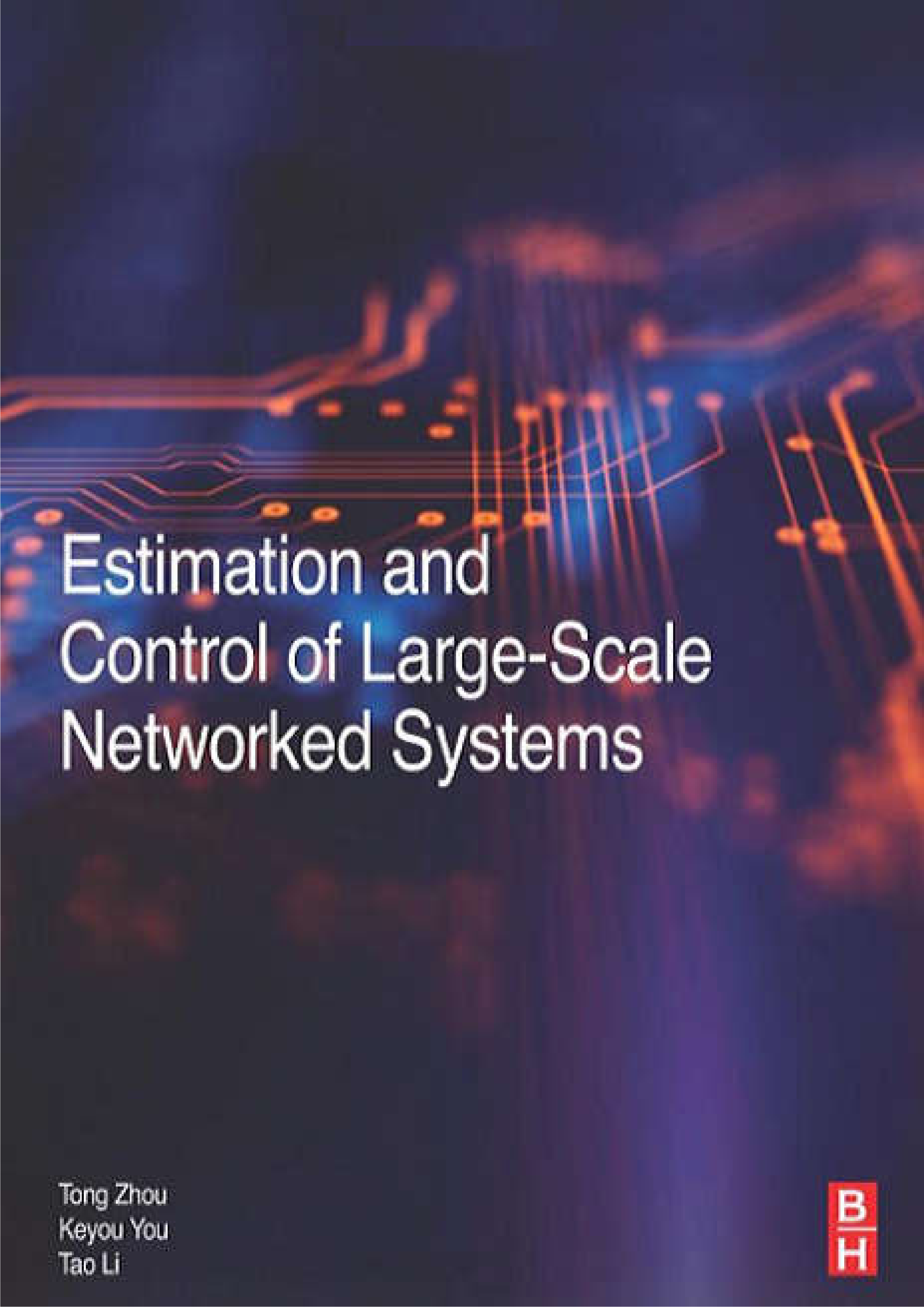

Estimation and Control of Large-Scale Networked Systems

Tong Zhou
Keyou You
Tao Li

B
H

Estimation and Control of Large-Scale Networked Systems

Estimation and Control of Large-Scale Networked Systems

Tong Zhou
Keyou You
Tao Li

ELSEVIER
Butterworth-Heinemann
An imprint of Elsevier

Butterworth-Heinemann is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-805311-9

Publisher: Mara Conner
Acquisition Editor: Sonnini R. Yura
Editorial Project Manager: Ana Claudia Garcia
Production Project Manager: R. Vijay Bharath
Designer: Greg Harris

Typeset by VTeX

*We dedicate this book to the "networked system"
whose members are both far from and near to us.
Without their stimulus, help, understanding, and
encouragements, it is impossible for us to observe,
concentrate, and finish this book.*

Preface

> *The ideal situation occurs when the things that we regard as beautiful are also regarded by other people as useful.*
>
> ***Donald Knuth***

Recent developments in sensing, communication, and computation provide great potentials of constructing a complicated system through digital communication techniques using strong local sensing and information processing capabilities, with the ambition of achieving more challenging tasks. Application areas include industrial automation, electrical power systems, transportation systems, defense systems, and so on. These systems are extensively called networked systems and usually have several different decision units, which may be spatially far from each other and connected through a real-time communication network. Information exchanges among these units may be neither instantaneous nor simultaneous. These characteristics ask an integrated thinking about control and communication in system analysis and synthesis, in which all the limitations of a digital communication channel must be explicitly taken into account. In addition, computation costs and numerical sensitivity must also be considered as metrics in evaluating a method developed for validating or designing these systems.

This book presents some recent results on identification, estimation, and control of a large-scale networked system, in which several fundamental issues have been attacked. These issues include controllability, observability, stability, robust stability, state estimation, robust state estimation, structure estimation, attack prevention, and so on. A general purpose here is to reach a global objective using local information and neighbor interactions/information exchanges. By introducing a novel model for a networked system, we are able to establish explicit relations between a system property and the system structure, which further enables utilization of structure information in system analysis and design. A Homographic transformation is adopted in expressing the recursive formulas of state estimations. This transformation leads to a very concise relation between the current and initial values of the recursions and greatly simplifies analysis of their convergence properties. A Riemannian metric is utilized in measuring the distance between two positive definite matrices. Under this metric, a Homographic transformation defined through a Hamiltonian matrix is always contractive. This property is quite attractive in analyzing estimation algorithms with random data droppings, which leads to some good approximations on its limit distributions. Statistical properties like the power law and so on have been explicitly taken into account in structure identification for a large-scale networked system.

This book is written for students and researchers with interests in studies and investigations on distributed estimation/control, estimation/control under communication restrictions, and estimation/control over a network. The material presented in this book includes some recent results of the authors reported as isolated conclusions in journals and conference proceedings. This book presents them in an integrated and coherent way and includes some related important results from other research groups. In each chapter, we present the material in a manner that strikes an appropriate balance between concepts on the one hand and technical depth and rigor on the other hand, with the expectation that an interested reader may refer to the associated primary literature for mathematical details that we have not included.

Chapters 8 and 9, as well as most of Chapter 5, are written by K.Y. You. Chapter 12 is written by T. Li. The remaining chapters and Section 5.6 of Chapter 5 are written by T. Zhou.

Tong Zhou
Beijing, China

Keyou You
Beijing, China

Tao Li
Shanghai, China

Acknowledgments

T. Zhou would like to thank the National Natural Science Foundation of China and the Education Ministry of China for their continuous financial supports, which enabled him to concentrate on his academic interests. Thanks also go to people who helped him in various different aspects. Specifically, T. Zhou would like to mention Prof. H. Kimura, his attitudes toward research have brought great influences. Encouragements from Prof. H.F. Chen and Prof. L. Guo are highly appreciated. Discussions with Prof. K.M. Zhou, Prof. J. Chen, and many other people were of significant help.

T. Zhou also wants to thank Ms. S.R. Yura and Ms. A.C. Garcia of Elsevier for their various efforts and help in the preparation of this book, as well as their kind permissions that allowed him to postpone the submission of the final manuscript more than half a year.

K.Y. You would like to acknowledge the support from the Excellent Young Scholar Fund from National Natural Science Foundation of China under grant no. 61722308.

T. Li wishes to express his indebtedness to the research collaborators Prof. J.F. Zhang, Prof. M.R. Fei, and Mr. N. Xiong. Many thanks to his former postdoctoral fellow and students, especially Dr. H.L. Liang, Mr. J.X. Wang, Mr. A.Q. Liu, Mr. J. Li, Mr. J.X. Chen, Mr. G.Q. Ma and Mr. W.S. Chen for their assistance in preparing the book.

T. Li also would like to acknowledge the Shanghai Key Laboratory of Pure Mathematics and Mathematical Practice and the following research grants for their support: the Excellent Young Scholar Fund from National Natural Science Foundation of China under grant 61522310, the Shu Guang project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under grant 17SG26, and the Shanghai Rising-Star Program under grant 15QA1402000.

Notation and Symbols

The Fourier transform of a time domain signal $u(t)$ is usually denoted by $u(j\omega)$, whereas its Laplace transform by $u(s)$.

The frequency response of a linear and time-invariant operator $G(\cdot)$ is usually denoted by $G(j\omega)$, whereas its transfer function matrix by $G(s)$.

$\exp(\cdot)$ exponential of a variable/number
$\log(\cdot)$ logarithm of a variable/number
$\Re(\cdot)$ real part of a variable/number/matrix/function
$\bar{\cdot}$ conjugate of a variable/number/matrix/function

x a lower case letter usually denotes a scalar variable or a vector
X a capital letter usually denotes a matrix
$\mathcal{X}$ a calligraphy capital letter usually denotes a set
$\mathcal{A}\bigcap\mathcal{B}$ the intersection of a set $\mathcal{A}$ and a set $\mathcal{B}$
$\mathcal{A}\bigcup\mathcal{B}$ the union of a set $\mathcal{A}$ and a set $\mathcal{B}$
$\mathcal{A}\backslash\mathcal{B}$ the relative complement of a set $\mathcal{B}$ in a set $\mathcal{A}$, which is also widely called the set-theoretic difference of set $\mathcal{A}$ from a set $\mathcal{B}$.

$(\cdot)^T$ transpose of a real/complex vector
$(\cdot)^H$ conjugate transpose of a complex vector
$(\star)^T WX$ or $XW(\star)^T$ abbreviation for $X^T WX$ or XWX^T
$\mathbf{col}\{X_i|_{i=1}^L\}$ the vector/matrix stacked by $X_i|_{i=1}^L$
$\mathbf{diag}\{X_i|_{i=1}^L\}$ a block diagonal matrix with its ith diagonal block being X_i. Sometimes, it is also written as $\bigoplus_{i=1}^L X_i$ or $X_1 \oplus X_2 \oplus \cdots \oplus X_L$.
$\left[X_{ij}|_{i=1,j=1}^{i=M,j=N}\right]$ a matrix with $M \times N$ blocks and its ith row jth column block being X_{ij}
0_m the m-dimensional zero column vector; the subscript is often omitted when it is clear or not important
$0_{m\times n}$ the $m \times n$-dimensional matrix with all zero-elements; the subscript is often omitted when it is clear or not important
I_n the $n \times n$-dimensional identity matrix; the subscript is often omitted when it is clear or not important
$||\cdot||_2$ Euclidean norm of a real/complex vector or the matrix norm induced from this Euclidean vector norm, which is the maximum singular value of that matrix
$||\cdot||_Q$ Euclidean norm of a real/complex vector weighted by a positive semidefinite matrix Q
$\mathrm{rank}(\cdot)$ the rank of a matrix
$\det(\cdot)$ determinant of a square matrix
$\mathrm{tr}(\cdot)$ the trace of a square matrix
$\rho(\cdot)$ the spectral radius of a matrix, that is, the largest magnitude of its eigenvalues
$\bar{\sigma}(\cdot)$ the maximum singular value of a matrix, sometimes also written as $\sigma_{\max}(\cdot)$

$A \otimes B$ the Kronecker product of matrices A and B
$X^{1/2}$ the square root of a positive definite matrix

$\mathbf{F}_l(*, \#)$ lower linear fractional transformation of the variable/vector/matrix/function # with a prescribed matrix/matrix-valued function $*$
$\mathbf{F}_u(*, \#)$ upper linear fractional transformation of the variable/vector/matrix/function # with a prescribed matrix/matrix-valued function $*$
$\mathbf{H}_m(\Phi, X)$ Homographic transformation of a matrix X defined as $(\Phi_{11}X + \Phi_{12})(\Phi_{21}X + \Phi_{22})^{-1}$ with $\Phi = \left[\Phi_{ij}|_{i,j=1}^{2}\right]$, in which Φ is a prescribed matrix/matrix valued function
$\mu_{\mathbf{\Delta}}(M)$ the structural singular value of a matrix M with respect to the uncertainty structure $\mathbf{\Delta}$

$\mathcal{B}\mathcal{X}$ the subset of a set $\mathcal{X}$, with each element having a norm not greater than 1
$\mathcal{RH}_{\infty}^{m\times n}$ the set of $m \times n$-dimensional rational and stable transfer function matrices
$\mathcal{R}^m$ the set of m-dimensional column real vectors
$\mathcal{R}^{m\times n}$ the set of $m \times n$-dimensional real matrices
$\mathcal{C}^m$ the set of m-dimensional column complex vectors
$\mathcal{C}^{m\times n}$ the set of $m \times n$-dimensional complex matrices
$\mathcal{L}_2^d$ the set of d-dimensional real-valued vectors with the square of their Euclidean norms integrable over the time interval $[0, \ \infty)$, that is,

$$\mathcal{L}_2^d = \left\{ v(t) \;\middle|\; v(t) \in \mathcal{R}^d, \ \int_0^{\infty} ||v(t)||_2^2 dt < \infty \right\}$$

$\mathbf{Pr}(\cdot)$ probability of a random event
$\mathbf{E}(\cdot)$ mathematical expectation of a random variable/vector/matrix
$\mathbf{E}(*|\#)$ conditional mathematical expectation of a random variable/vector/matrix
$\mathbf{Var}(\cdot)$ variance of a random variable
$\mathbf{Cov}(\cdot)$ covariance matrix of a random vector
$\mathcal{N}(*, \#)$ normal distribution of a random variable/vector with mathematical expectation $*$ and covariance matrix #

CHAPTER 1

Introduction

1.1 A General View on Control System Design

Control loops exist extensively in biological systems, engineering systems, financial systems, social systems, and so on [1]. Whereas feedback is an essential characteristic in control systems, factors like developments of technologies, performance requirement strengthening, and so on make analysis and synthesis of a control system very different in distinct stages. In a classic control system, both its input and output are scalars. A system with this property is usually called single-input single-output system, which is often abbreviated as SISO. For these systems, transfer functions and differential equations are widely utilized in describing their input–output properties. Methods like Bode diagram, root locus, and so on, play central roles in the analysis and synthesis of a linear and time-invariant SISO system. The most extensively adopted controller in classic control systems is the so called PID controller, with its abbreviations P, I, and D standing respectively for proportion, integration, and differentiation. This controller is a special kind of the so-called lead-lag controller, its realization has been well supported by standard industrial products, and various methods have been developed for the adjustments of this controller to satisfy design specifications of a control system. Methods like description functions, phase plane, and so on have also been developed to analyze systems with some special types of nonlinearities.

Around the 1960s, with the requirements from astronautics and aeronautics, plants began to emerge in which there are multiple variables that are to be controlled simultaneously. These plants are usually called multi-input multi-output systems, which are often abbreviated as MIMO, and state space models are invented to describe their dynamics. For these systems, the Luenberger observer and the Kalman filter have been developed to estimate their states, among many other methods. Moreover, a method called the linear quadratic Gaussian control, usually abbreviated as LQG, has been developed to construct a controller. In addition to this, many other methods, such as decoupling control, model predictive control, and so on, have been developed to control an industrial MIMO process.

When modeling errors are explicitly and directly taken into account in system analysis and synthesis, robust control theory has been developed to handle this problem. One of the well-known robust controller design methods is the H_∞ control, in which the H_∞-norm of a transfer function matrix is minimized guaranteeing internal stability of the feedback control system. This norm is originally introduced by Zames [2] and found to be appropriate in

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00001-4

specifying both the size of modeling errors and the induced gain from a disturbance input vector to an error output vector in control systems. It is extensively believed that one of the motivations for introducing this induced norm into system analysis and synthesis is to bring modeling errors in a plant dynamics description, which is described in the frequency domain, back to the central stage in system analysis and synthesis. Note that in classic control theories, robustness of a feedback control system is reflected by its gain margin and phase margin, which is measured by the frequency response of its open-loop transfer function, whereas in the control theories developed around 1960s, whose representative results include LQG control, pole placements, and so on, an explicit and accurate model is required, which is usually given in a state space form. Two of the most important results associated with this model error description seem to be the so-called small gain theorem and two Riccati-equation-based formulas for the H_∞ control problem. Although the H_∞-norm is widely regarded to be suitable in describing unmodeled dynamics, it usually introduces conservativeness in dealing with parametric modeling errors. To overcome this drawback, a block diagonal structure is suggested for modeling error descriptions, and structured singular values (SSV) are defined to measure the robustness of a feedback system with both parametric modeling errors and unmodeled dynamics [3,4]. SSV computations, however, are proven to be generally NP-hard [5]. This computational difficulty greatly hampers applications of the SSV to system analysis and synthesis.

In all these theories, communications have not been explicitly taken into account, and it is implicitly assumed that all data transmissions in a feedback control system, which include those among plant subsystems, those between a plant and its controller, and so on, are performed with an infinite precision in value, an infinite communication bandwidth, and an infinite speed. This implicit assumption makes state estimation algorithms, control algorithms, and so on completely independent of communications and greatly simplifies system analysis and synthesis. The associated results work well for traditional engineering systems. Due to technology developments in sensors, communications, and so on, as well as more complicated and demanding tasks expected for a system, the number of subsystems increases significantly, and they are expected to cooperate to achieve a much higher performance level. In addition, digital communication networks are expected to be used in transmitting information among different components of a control system in order to reduce hardware costs of the system and to improve maintenance capabilities. Systems with these characteristics are becoming ubiquitous, with applications ranging from electricity power systems to rescue robot teams, remote surgery, and so on.

However, with the increment of the number of subsystems and the introduction of public communication channels into a control system, some essentially new and challenging issues arise in system analysis and synthesis. For example, the structure of a plant becomes an important factor, as well as communication qualities. More precisely, in a networked system, data transmission and information processing cannot be performed instantaneously in general. Data may be delayed, be out of order, and even be dropped out, noting that data transmissions through communication networks are usually corrupted due to noises in the communication medium, congestion of a communication network, and protocol malfunctions and that communication channels may change from time to time, and so on. To make things worse, there may even exist some attackers who inject malicious disturbances into the system with an objective of destroying its functions. All these bring new challenging issues in system analysis and synthesis.

On the other hand, for most of large-scale networked systems, rather than utilizing a single processing unit, it is preferable to distribute the control tasks and/or estimation tasks among several processing units. In this task division, sparseness of the system is an important factor, which needs to be taken into account, as well as the topology of subsystem connections. In addition, these processing units may not be triggered by a common clock pulse, which makes the sampling, holding, and computation activities of these processing units not synchronized. In other words, different processing units may have different sampling periods. When communication channels are shared by several networked systems, the sampling period of each processing unit may have some irregularities, as time sharing of a communication channel often hampers a precise scheduling for data transmissions in a particular plant.

Generally speaking, analysis and synthesis of a networked system is a relatively new and challenging field in system theories, which requires knowledge on feedback control systems, communication theory, information theory, and so on, which previously belong to different engineering disciplines. The purposes of this book are to investigate some important issues and summarize some important recent works in this area.

1.2 Communication and Control

In a control system, both information and energy are transmitted. Information transmission is necessary as a deviation of the plant output from a desirable trajectory and must be recorded and processed by a signal processing unit, which is usually called a controller or a regulator, and the processed deviations or plant outputs must be sent back to the plant as its input to make the plant work properly. In other words, information transmission is essential in the improvement of plant performances. Different from signal processing, energy transmission is also necessary in a control system, noting that power is required in accomplishing any task in a mechanical system, a biology system, an electrical system, and so on. Particularly, energy transmissions usually cannot be performed exactly in a control system due to technology difficulties in accurately generating the required amount of energies for accomplishing the tasks, which will introduce model errors into the system. A general situation is that the bigger the transmitted energy, the larger the model error. As feedback is usually adopted in a control system, the influence of modeling errors may be amplified if the associated controller has not been well designed. This makes robustness a much more essential issue in control system designs comparing to those of other fields like signal processing, mechanical system design, and so on.

Traditionally, information transmissions in a control system are assumed to be instantaneous and to have an infinite precision. As mentioned in the previous section, with the introduction of public communication channels into a control system, these conditions are no longer satisfied. In addition to this, communication costs must also be taken into account in the design of a networked system. Naturally, it will be appreciated if less data is sent with a lower precision in a networked system, provided that the required tasks can be accomplished satisfactorily by the associated plant. In other words, in the analysis and synthesis of a networked system, one of the major concerns is on the description of the minimum amount of information transfer required for satisfactory system performances, in which stability is usually the most important factor. Information transfer may happen among various parts of the system, for example, between different subsystems, between a local controller and a subsystem, and so on. Another important issue is about the influences on system performances from the data transmission rate and the data expression precision. It is also interesting to investigate effective coordinations among different subsystems with minimum requirements on information exchanges.

When a system consists of a great amount of subsystems and these subsystems are spatially far from each other, and when a system has a great amount of states, it is usually not appreciative, and even prohibitive/impossible, in actual applications either to estimate all the plant states by a single centralized estimator or to control all the subsystems with a single centralized controller. In addition to computational costs, the reasons also include considerations from maintenances, robustness against failures of a subsystem, and so on. On the other hand, if each subsystem is independently estimated or controlled, then performances of estimations or control may be significantly deteriorated. In the worst case, even stability of the closed-loop system may not be reached although the whole system is controllable. This means that some coordinations, which is sometimes also called cooperations, are necessary among the estimators/controllers designed for each individual subsystem. To realize these coordinations/cooperations, communications among these estimators/controllers are also necessary. In other words, the estimator/controller itself also consists of several subestimators/subcontrollers, and these subestimators/subcontrollers may be connected through public communication channels, which may cause data missing, data disordering, and so on.

In communication networks, source signals are usually sampled and encoded into a sequence of channel input symbols, which is then transmitted through some communication media, for example, antenna, satellites, optical fibers, and so on and received by an equipment that gives a sequence of channel output symbols. A perfectly designed communication network intends to completely recover the original source signals from the channel output symbols. Since external disturbances are unavoidable during transmissions and quantification is widely adopted in communications, a completely perfect source signal recovery is usually impossible, noting that under these situations, two different resource signals may lead to a completely equal channel output sequence. To describe the capability of a communication channel in reliably transmitting signals, a concept called channel capacity is suggested by Shannon during World War II, which provides a mathematical model that can be accurately computed. Intuitively, the capacity of a communication channel is the tightest upper bound about the rate at which a signal can be reliably transmitted over it. Alternatively, a channel capacity can also be explained using the noisy-channel coding theorem as the highest information rate (in units of information per unit time) that can be achieved by a communication channel with an arbitrarily small error probability. Mathematically, the capacity of a communication channel is given by the maximal value of the mutual information between the input signal and output signal of the channel, in which the maximization is taken over all possible probability density functions of the input signal.

To deal with analysis and synthesis of a networked system, it appears necessary to introduce an appropriate model of communication channels into the descriptions of its dynamics. With the concept of channel capacity, various models are expected to be developed to meet this requirement, which characterize the communication constraints in a networked system depending on the underlying channel characteristics and information pattern.

More precisely, when data missing is concerned, a Bernoulli process model and a Markov chain model are extensively adopted. When the data missing is described by a Bernoulli process, the communication channel is usually called an erasure channel in communications, whereas in the case of a Markov chain, it is called a Gilbert–Elliott channel. In an erasure channel, data loss is assumed to be independent of each other, and its influences on state estimation and system control are relatively easy to be analyzed. However, the Gilbert–Elliott channel appears to be a more realistic model in the description of data losses due to imperfect communications, since it takes influences of the previous states of a communication network on its current states, which is closer to actual situations. A cost is that this model may make system analysis more complicated. With these data missing models, the capacity of a communication channel can be simply characterized. For example, when the Bernoulli process is used, if information is contained in the input signal to indicate that it is a signal, then the channel capacity can be proved to be equal to the probability that a data packet is not lost.

When external disturbances are taken into account, they are usually treated as an additive noise, which is simple and yet representative in various communication channels. Under such a situation, the channel capacity constraint arises as a bound that is usually put on the power of channel input signals with the purpose to reduce interferences among different communication users and to meet hardware requirements. When the external disturbance has a Gaussian distribution, the channel capacity can be simply computed using the signal-to-noise ratio of the channel.

However, when a channel capacity is derived using the mutual information, this capacity is achievable under several assumptions that in general cannot be easily satisfied in practical applications. These assumptions include that the capacity-achieving code can be arbitrarily long, there does not exist any restriction on the coding complexity, and so on. In addition, causality has not been explicitly taken into account in this derivation. Note that a control system usually requires feedbacks, and a long code often leads to significant time delays that are usually not very appreciative in control system designs. This means that in an actual engineering problem, the channel capacity usually cannot be achieved. On the other hand, when a wireless communication channel is used, due to the effects of multipaths and shadowing in a wireless channel, a signal may experience fluctuations in its transmissions. This phenomenon is still difficult to be modeled satisfactorily in a general setting, and only some simple statistical models have been proposed, such as the Rayleigh model, the Rician model, and so on, which depend heavily on the particular signal propagation environments and transmission scenarios. Moreover, for a large-scale networked system, a multiple-input multiple-output communication network appears to be necessary. However, there is still no mature theory that successfully deals with information transmissions in such a communication network facing channel inferences and external noises.

In summary, when public communication channels are adopted in a networked control system, various efforts are still required for establishing an appropriate model for a communication channel that satisfies requirements raised by controller analysis and synthesis.

1.3 Book Contents

Being aware of the importance of communications in a networked system, as well as that centralized estimation and/or centralized control is not very appropriate for a large-scale system and systems that are constituted from subsystems that are geometrically far away from each other, this book investigates five important issues in the analysis and synthesis of a large-scale networked system, which are listed as follows.

- Controllability, observability, stability, and robust stability. These are fundamental properties in system analysis and synthesis. Relations are revealed between these properties and subsystem connections, which are also called the structure or topology of a system. Conditions are clarified for each subsystem such that a controllable/observable system can be constructed.

- State estimation. Both centralized and distributed designs are investigated, as well as situations in which plant output measurements may be lost due to communications. Influences of data missing on estimation accuracies are investigated, as well as conditions under which a distributed estimator has the same steady estimation accuracy as a centralized one.
- Distributed control. Taking into account factors like robustness against subsystem failures, scalability of estimation/control procedures, and so on, a control using only local information is much more appreciative in actual engineering applications. Controller designs with relative state feedback and relative output feedback are attacked respectively for a networked system with time-invariant subsystem interactions, which is also called a static topology, and with time-varying interactions, which is sometimes called a dynamic topology.
- Attack estimation and identification. This is a relatively new issue in control system analysis and synthesis although some simple situations have been dealt with for power systems for a long time, in which system dynamics has not been taken into account. Relations among system observability, observer design, and estimation and identification of attacks are discussed.
- Structure identification. In many actual systems, such as an economy system, gene regulation networks, and so on, interactions among different subsystems are not clear from the underlying principles, and it is important to understand these interactions from observed data and available knowledge on system structure. Some methods are developed, which reveal subsystem interactions using respectively steady-state system output measurements and dynamic system output measurements.

These topics are dealt with respectively in the following Chapter 3 to Chapter 12.

1.3.1 Controllability and Observability of a Control System

In Chapter 3, controllability and observability of a networked system are discussed. A new description is introduced to model the dynamics of a system consisting of several subsystems, in which each subsystem is represented by a state space model. In this subsystem model, its input vector is divided into two ingredients, which are respectively called the internal input vector and external input vector. Completely the same division is also performed on the output vector of each subsystem model. Interactions among subsystems are described by a subsystem connection matrix, which reflects that an internal input of one subsystem is in fact an internal output of another subsystem, possibly transmitted by a public communication channel. This model is adopted throughout this book in the description of the dynamics of a networked system.

Necessary and sufficient conditions are given respectively for the controllability and observability of a networked system, which depend only on the subsystem connection matrix, transmission zeros of each subsystem, and their related vectors. With this condition, it has been proven that for a networked system to be controllable/observable, each subsystem must be controllable/observable. On the contrary, if each subsystem is controllable/observable, then a controllable/observable networked system can always be constructed. In addition, an explicit formula is given for the minimum number of inputs/outputs for each subsystem in a networked system, such that a controllable/observable networked system can be constructed. When the number of inputs/outputs is fixed, a parameterization is given for all the input/output matrices such that the associated subsystem is controllable/observable.

1.3.2 Centralized and Distributed State Estimations

Chapter 4 discusses observer designs for a lumped system. Both Kalman filtering and robust estimation are investigated. The Kalman filter is rederived using a maximum likelihood approach, which plays an essential role in obtaining the robust state estimator described in the same chapter. Through penalizing sensitivities of the cost function associated with the Kalman filter to parametric modeling errors, a robust state estimator is obtained, which can also be recursively realized and has a computational complexity similar to that of the Kalman filter. When the plant nominal model is time invariant, conditions are given for the convergence of the robust state estimator to a time-invariant system.

In Chapter 6, distributed state estimations are discussed for a system consisting of several subsystems. The estimator is assumed to have the same structure as that of the plant, and each of its subsystems is required to only use local plant output measurement; that is, the subsystem connection matrix of the estimator is the same as that of the plant, and state estimates of each subsystem in the estimator are updated using only the output measurements of one associated subsystem in the plant. Under the requirements that the state estimate for each subsystem is unbaised and the covariance matrix of estimation errors is minimal for each subsystem, a recursive formula is derived respectively for the update gain matrix of each subsystem in the state estimator and for the covariance matrix of estimation errors of the whole system. Conditions are also given for the distributed state estimator having the same steady estimation accuracy as the lumped Kalman filter.

1.3.3 State Estimations and Control With Imperfect Communications

When a public channel is adopted in a networked system, some measurements may be lost due to imperfect transmissions. In Chapter 5, state estimations are investigated under such a situation. It is shown that when an indicator is included in a transmitted package, which indicates that the package contains plant output measurements, the optimal state estimator is equal to a predictor when a measurement is lost, whereas it is simply the Kalman filter when a measurement is satisfactorily transmitted. These results have been extended to situations in which parametric errors exist in a plant state space model. Conditions are also derived for the boundedness of the covariance matrix of estimation errors, as well as its stationary distribution. In these investigations, rather than a Euclidean induced norm, a Riemannian metric is adopted in measuring the difference between two positive definite matrices.

The purpose of Chapter 8 is to understand interactions between the control components and the communication components of a networked system with communication channels. A problem of stabilizing a linear time-invariant plant is discussed, in which a sensor for the plant outputs is connected to the controller through a communication channel. Influences of quantization on controller designs have been investigated, and some limitations of the traditional approach to system stabilization have been established. The minimum data rate for stabilizing a linear time-invariant system has also been derived, which is further extended to a stochastically time-varying communication channel.

1.3.4 Verification of Stability and Robust Stability

Chapter 7 deals with how to verify stability and robust stability of a large-scale networked system when there exist both parametric modeling errors and unmodeled dynamics in a state space model of each of its subsystems. Some necessary and sufficient conditions have been derived, which explicitly depend on the subsystem connection matrix of the networked system under investigation. When the system under investigation has a great number of subsystems, these conditions are helpful in reducing computational complexity of the associated system analysis and synthesis, noting that a large-scale system usually has a sparse structure. Another characteristic of these conditions is that most of the involved matrices are block diagonal, which is also appreciative in improving numerical stability of the associated matrix computations and in reducing their computational complexities.

1.3.5 Distributed Controller Design for an LSS

Chapter 9 discusses distributed control design for the consensus and formation of a discrete-time multiagent system, in which all agents are required to reach an agreement using some shared data through local communications. A problem formulation is given for the consensus of a multiagent system with general linear agent dynamics. Distributed controller designs have been investigated under the situation that relative state feedbacks are available and under the situation that relative output feedbacks are available. The associated techniques and results are further extended to address a distributed formation problem.

1.3.6 Structure Identification for an LSS

Chapter 10 deals with causal relation inference for a networked system from experiment data. This problem is frequently encountered in many fields, including biologies, economy, finance, and so on. In these fields, measurements of direct influences between two different subsystems are generally time consuming or economically expansive,and sometimes may even be prohibitive. This chapter investigates possibilities of estimating these relations from experiment data and statistical properties of the structure of a large-scale networked system. Most of the discussions are concentrated on gene regulation networks, but the results may be helpful in solving similar problems in other fields. Both steady state data and dynamic data have been investigated. The so-called power law is incorporated into the structure inference using steady-state experiment data, in which the total least squares method and the maximum likelihood estimation method are respectively used to estimate interactions among different subsystems of a plant. A concept called relative variation is developed for these interaction inferences. When dynamic experiment data are available, a robust state estimation-based method is developed, which is proven to be more efficient than the methods based on the extended Kalman filter (EKF) or the unscented Kalman filter (UKF). However, application of this method is still restricted to systems with only a few subsystems, and statistical structure properties of a large-scale system has not been utilized.

1.3.7 Attack Estimation/Identification and Other Issues

In Chapter 11, estimation and identification are investigated for attacks in a networked system, which becomes more and more important in networked system analysis and design, as malicious and organized inputs to a networked system may even destroy its stability. Both static data-based methods and dynamic data-based methods are introduced. It is observed that these problems are closely related to system observability and system transmission zeros, and accurate system model is quite important in both attack preventions and attack constructions.

Time synchronization and state consensus problems are discussed in Chapter 12, which are also important topics in the control of a networked system. The former deals with how to calibrate clocks in each subsystem, which is essential to realize a distributed estimation algorithm or a distributed control algorithm. The latter investigates protocol designs such that with the increment of the temporal variable, the state vector of each subsystem reaches an equal value. Clearly, time synchronization can be regarded as a particular state consensus problem. These problems are discussed respectively under the situation that the subsystem connections are fixed and under the situation that the subsystem connections are time varying. Conditions have been established for the associated problems.

1.4 Bibliographic Notes

It appears that systematic investigations on a large-scale dynamic system emerged around the beginning of the 1960s. Various research literature and research monographs have been published after that time, in which a state space model is extensively adopted with an emphasis on subsystem interconnections through their states. Other models include hierarchical model, graph-based model, and so on. Relations among system stability, regulation performances, and system structure have been extensively studied from various different aspects. Major results can be found in [6] and the references therein.

Revealing causal relations between different time series record sets is also a long and attractive topic. A historic review on major achievements and recent advances in this field can be found in [7], [8], and the references included in these monographs. The former is more focused on relations between the recorded data of two time series, whereas the latter has put many efforts on data generated from a dynamic system that consists of a large amount of subsystems.

Recent interests in networked systems were mainly triggered by the introduction of a public communication channel into the system used to transfer signals between a plant and a controller or from a plant to a state estimator and to coordinate a group of subsystems with neighbor information to construct a desired formation. Imperfect signal transfers make it necessary to introduce a model of communication channels in system analysis and synthesis, whereas variations of neighbors ask to take system structure switchings into account in the design of a control law. The monograph [9] appears to be the first research that summarizes major research results in a unified way on estimation and control with communication channels, whereas [10] places special emphasis on the influences of model errors in the analysis and synthesis of a networked system. In [11], a game theory-based approach is adopted in the investigation of the interactions between information and control in a networked system, which consists of several subsystems, has heterogeneous communication media, uses decentralized and distributed measurements, and acquires information with possible delays. Unified studies are also given in [12] on the effects of imperfect communications on the stability and performances of a networked system.

References

[1] S.Y. Nof (Ed.), Springer Handbook of Automation, Springer-Verlag, Berlin, Heidelberg, Germany, 2009.

[2] G. Zames, Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses, IEEE Transactions on Automatic Control 26 (1981) 301–320.

[3] J.C. Doyle, Analysis of feedback systems with structured uncertainties, IEE Proceedings, Part D 133 (1982) 45–56.

[4] M.G. Safonov, Stability margins of diagonally perturbed multivariable feedback systems, IEE Proceedings, Part D 129 (1982) 251–256.

[5] A. Packard, J.C. Doyle, The complex structured singular value, Automatica 29 (1993) 71–109.
[6] D.D. Siljak, Large-Scale Dynamic Systems: Stability and Structure, North-Holland Books, New York, USA, 1978.
[7] J. Pearl, Causality: Models, Reasoning, and Inference, second edition, Cambridge University Press, UK, 2009.
[8] E.D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, 2009.
[9] A.S. Matveev, A.V. Savkin, Estimation and Control Over Communication Networks, Birkhäuser, Boston, USA, 2009.
[10] E. Garcia, P.J. Antsaklis, L.A. Montestruque, Model Based Control of Networked Systems, Birkhäuser, Springer International Publishing, Switzerland, 2000.
[11] S. Yuksel, T. Basar, Stochastic Networked Control Systems, Birkhäuser, Springer Science+Business Media, New York, USA, 2000.
[12] K.Y. You, N. Xiao, L.H. Xie, Analysis and Design of Networked Control Systems, Springer-Verlag, London, UK, 2015.

CHAPTER 2

Background Mathematical Results

2.1 Linear Space and Linear Algebra

In signal processing and control system analysis/synthesis, a usually encountered situation is that several variables must be handled simultaneously. To deal with the associated problems, knowledge on linear algebra is necessary. This section reviews some of these topics and results that are helpful to understand the following chapters of this book.

A linear space is a collection of vectors, which can be added together and multiplied by scalars. Usually, the scalars are taken to be real numbers. The operations of vector addition and scalar multiplication must satisfy the following eight requirements, which are respectively called associativity of addition, commutativity of addition, identity element of addition, inverse elements of addition, compatibility of scalar multiplication with multiplication between scalars, existence of an identity element of scalar multiplication, distributivity of scalar multiplication with respect to vector addition, and distributivity of scalar multiplication with respect to scalar addition. More precisely, we have the following definition.

Definition 2.1 (Linear Space). *A set $\mathcal{V}$ is a linear space if the following properties are simultaneously satisfied for all elements u, v, and w in this set and for all real numbers a and b.*

- *associativity of addition: $u + (v + w) = (u + v) + w$.*
- *commutativity of addition: $u + v = v + u$.*
- *identity element of addition: there exists an element $0 \in \mathcal{V}$, which is called the zero vector, that satisfies $v + 0 = v$ for each $v \in \mathcal{V}$.*
- *inverse elements of addition: for each $v \in \mathcal{V}$, there exists an element $z \in \mathcal{V}$ satisfying $v + z = 0$. This element is usually called the additive inverse of the vector v and denoted by $-v$.*
- *compatibility of scalar multiplication with multiplications among scalars: $a(bv) = (ab)v$.*
- *identity element of scalar multiplication: $1v = v$.*
- *distributivity of scalar multiplication with respect to vector addition: $a(u + v) = au + av$.*
- *distributivity of scalar multiplication with respect to scalar addition: $(a + b)v = av + bv$.*

Well-known examples of linear spaces include a finite-dimensional real Euclidean space, a space consisting of random variables with zero mathematical expectation and finite variance, a space of time series with finite energies, and so on.

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00002-6

Associated with linear spaces, an essential concept is an inner product, which is a mapping from a linear space to real numbers that satisfies simultaneously three particular properties, which are respectively called symmetry, linearity, and positive-definiteness. This mapping is usually denoted by $< \cdot, \cdot >$.

Definition 2.2 (Inner Product). *A mapping $< \cdot, \cdot >$ from $\mathcal{V} \times \mathcal{V}$ to $\mathcal{R}$ is an inner product over the linear space $\mathcal{V}$ if the following requirements are simultaneously met by it.*

- *symmetry: for each vector $u \in \mathcal{V}$ and each vector $v \in \mathcal{V}$, $< x, y >=< y, x >$.*
- *linearity: for arbitrary vectors $u, v, z \in \mathcal{V}$ and arbitrary real numbers a and b, $< ax + by, z > = a < x, z > +b < y, z >$.*
- *positive-definiteness: for each vector $u \in \mathcal{V}$, $< u, u >\geq 0$. Moreover, $< u, u >= 0$ if and only if $u = 0$.*

A space endowed with an inner product is called an inner product space.

A concept closely related to inner product is the norm of a linear space.

Definition 2.3 (Norm). *A real-valued function defined on a linear space $\mathcal{V}$, which is usually denoted by $|| \cdot ||$, is called a norm of this space if it simultaneously satisfies the following three conditions.*

- *homogeneity: for each vector $u \in \mathcal{V}$ and each real number α, $||\alpha u|| = |\alpha| \times ||u||$.*
- *triangle inequality: for arbitrary vectors $u, v \in \mathcal{V}$, $||x + y|| \leq ||x|| + ||y||$.*
- *positive-definiteness: for each vector $u \in \mathcal{V}$, $||u|| \geq 0$. Moreover, $||u|| = 0$ if and only if $u = 0$.*

A space endowed with a norm is called a normed space. An inner product space is obviously also a normed space, noting that if a space is endowed with an inner product $< \cdot, \cdot >$, then a norm can be simply defined as $||u|| = \sqrt{< u, u >}$. However, the converse is in general not true. In fact, some conditions should be satisfied for the existence of an inner product that leads to a prescribed norm, which is usually called the parallelogram condition, that is, for arbitrary $x, y \in \mathcal{V}$,

$$||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2).$$

For a finite-dimensional real Euclidean space $\mathcal{R}^n$, its element can be denoted as $x = \mathbf{col}\{x_i|_{i=1}^n\}$, in which each of the x_is is a real number. With this representation, an inner product can be defined as

$$< x, y >= \sum_{i=1}^{n} x_i y_i.$$

When a linear space is constituted from n-dimensional random vectors with zero mathematical expectation and finite covariance matrix, the following inner product is extensively adopted:

$$< x,\ y >= \mathbf{E}(x^T y).$$

This inner product has been proven to be very helpful in developing signal processing algorithms, including the well-known Wiener and Kalman filters. When a linear space consists of time series with finite energies,

$$< x,\ y >= \int_{-\infty}^{\infty} x(t)y(t)dt$$

is one of the well-known inner products. This inner product plays an essential role in robustness analysis of a control system and design of a robust control system.

For a finite-dimensional real Euclidean space $\mathcal{R}^n$, an extensively adopted norm is

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

in which p is a real number not smaller than 1. Especially, this norm is called the Euclidean norm if the parameter $p = 2$. Usually, this vector norm is called the p-norm. When a linear space consists of time series with finite energies, a widely utilized norm is

$$||x(t)||_p = \left(\int_{-\infty}^{\infty} |x(t)|^p dt \right)^{1/p}.$$

Once again, the parameter p is required to be real and not smaller than 1.

To discuss relations between two finite-dimensional real Euclidean spaces, a mapping is required. When this mapping is linear, it is usually represented by a matrix. Specifically, a linear mapping from $\mathcal{R}^m$ to $\mathcal{R}^n$, which is usually denoted $\mathbf{T}:\ \mathcal{R}^m \rightarrow \mathcal{R}^n$, can be represented by an $n \times m$-dimensional matrix T such that, for each $x \in \mathcal{R}^m$,

$$y = Tx. \tag{2.1}$$

For an $n \times n$-dimensional square matrix A, its eigenvalue is a number that satisfies $|\lambda I_n - A| = 0$. Associated with each eigenvalue, there is a nonzero n-dimensional vector x satisfying $Ax = \lambda x$, which is called a right eigenvector of this matrix. Correspondingly, there is also an n-dimensional nonzero vector y satisfying $y^H A = \lambda y^H$, which is called a left eigenvector of this matrix. It is worth noting that even if each element of a square matrix is real valued, its eigenvalues and left and right eigenvectors may be complex valued. On the other hand, when the matrix A is real valued and $\lambda \in \mathcal{C}$ is one of its eigenvalues, then the conjugate of λ, that is, $\bar{\lambda}$, is also an eigenvalue of this matrix.

In linear algebra, it is a basic result that any square matrix admits a Jordan canonical representation, which is given by the following theorem.

Theorem 2.1. *Let A be an $n \times n$-dimensional complex-valued matrix. Then there always exists a nonsingular matrix T such that*

$$A = TJT^{-1}$$

in which

$$J = \mathbf{diag}\{J_i|_{i=1}^{p}\}, \quad J_i = \mathbf{diag}\{J_{i,j}|_{j=1}^{q_i}\},$$

$$J_{i,j} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix} \in \mathcal{C}^{n_{i,j} \times n_{i,j}},$$

where λ_i, $i = 1, 2, \cdots, p$, are the distinct eigenvalues of the matrix A. Moreover, $n_{i,j}$ satisfy $\sum_{i=1}^{p} \sum_{j=1}^{q_i} n_{i,j} = n$.

In this representation, q_i is called the geometric multiplicity of the matrix A associated with its eigenvalue λ_i, whereas $\sum_{j=1}^{q_i} n_{i,j}$ is its algebraic multiplicity associated with that eigenvalue. Moreover, the matrix T can be constructed from the right eigenvectors of the matrix A. Obviously, when the geometric and algebraic multiplicities are equal to each other for each eigenvalue of a matrix, its associated Jordan canonical form reduces to a diagonal matrix. In this case, the matrix is said to have a diagonal representation.

A matrix is said to be symmetric if its transpose equals to the matrix itself. If a matrix is complex valued and its conjugate transpose equals to the matrix itself, then the matrix is said to be Hermitian. It is well known that both symmetric and Hermitian matrices only have real eigenvalues and diagonal representations.

A symmetric or Hermitian matrix A is said to be positive definite if all its eigenvalues are positive, which is usually denoted as $A > 0$. It is said to be positive semidefinite and denoted as $A \geq 0$ if all its eigenvalues are not negative. Negative definiteness and negative semidefiniteness of a matrix can be defined similarly.

The following lemma gives some interesting properties of the eigenvalues of the sum and product of two symmetric matrices and some relations between two positive definite matrices

[1,2]. These properties and relations play important roles in the remaining parts of this book. One example is the analysis of the convergence characteristics of an identification algorithm given in Chapter 10.

Lemma 2.1. *Let $A, B \in \mathcal{R}^{n\times n}$ be two symmetric matrices. Denote their minimum eigenvalues respectively by $\lambda_{\min}(A)$ and $\lambda_{\min}(B)$. Then*

$$\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B).$$

If both A and B are positive semidefinite matrices, then

$$\lambda_{\min}(AB) \geq \lambda_{\min}(A)\,\lambda_{\min}(B).$$

Moreover, if $A \geq B$, then

$$\mathrm{tr}(A) \geq \mathrm{tr}(B) \quad \textit{and} \quad \lambda_{\min}(A) \geq \lambda_{\min}(B).$$

Furthermore, if $A \geq B > 0$, then

$$A^{-1} \leq B^{-1}$$

and

$$TAT^T \geq TBT^T \geq 0$$

for every $m \times n$-dimensional real matrix T.

Another extensively utilized matrix representation is called singular value decomposition (SVD), which is valid even when the matrix is not square. It is now well known that the maximum singular value of a matrix is an appropriate measure of modeling errors in system analysis and synthesis. Moreover, if we consider the vector x in Eq. (2.1) as an input, and the vector y in this equation as an output, then a singular vector of the matrix T is an appropriate indicator for the direction of inputs or outputs. In particular, a singular vector associated with a small singular value provides the directions of inputs that have weak influence on outputs, whereas a large singular value provides the directions of inputs that have strong influence on outputs. In addition, the ratio between the maximum and the minimum singular values of a matrix is also a good indicator for the robustness of an algorithm for matrix computations, for example, computing the inverse of a matrix.

A $p \times p$-dimensional matrix T is called unitary if $T^H T = T T^H = I_p$.

Theorem 2.2. *Let A be an $m \times n$-dimensional complex-valued matrix. Then there exist an $m \times m$-dimensional unitary matrix U and an $n \times n$-dimensional unitary matrix V such that*

$$A = U\Sigma V^H, \quad \Sigma = \begin{bmatrix} \Sigma_* & 0 \\ 0 & 0 \end{bmatrix},$$

in which $\Sigma_* = \mathbf{diag}\{\sigma_i|_{i=1}^{\min\{m,\,n\}}\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,\,n\}} \geq 0$.

Differently from the eigenvalue decomposition of a matrix, the matrix Σ_* is always guaranteed to be diagonal in its singular value decomposition.

The concept of a structured singular value, usually abbreviated SSV, is widely used in system analysis and synthesis. This concept is originally developed by Doyle [3] and has been extensively investigated by many researchers in robust control system designs. Summaries of its properties, computations, applications, and connections with other concepts developed in robust control theories are given in [4,5]. A similar concept has also been suggested almost at the same time by Safonov [6].

Definition 2.4 (SSV). *Given an $m \times n$-dimensional complex matrix A and an uncertainty description $\mathcal{U}$ with each of its elements having a dimension of $n \times m$, the structured singular value (SSV) of the matrix A with respect to the uncertainty description $\mathcal{U}$, denoted $\mu_{\mathcal{U}}(A)$, is defined as*

$$\mu_{\mathcal{U}}(A) = \begin{cases} \frac{1}{\min\{\ \bar{\sigma}(U)\ |\ |I-AU|=0,\ \ U\in\mathcal{U}\,\}} & \textit{if there is at least one } U \in \mathcal{U}, \\ & \ \textit{such that } |I - AU| = 0; \\ 0 & \textit{otherwise}. \end{cases}$$

Usually, the uncertainty description $\mathcal{U}$ has a block diagonal form, which can include all parametric errors of unmodeled dynamics of a state space model. This makes it possible to investigate influences of a modeling error on system performances, not only using its size, but also using its directions into the plant model. When there are no constraints on the structure of the uncertainties, that is, every element of the matrix U belonging to the uncertainty description $\mathcal{U}$ is allowed to vary independently, it can be proven that the structured singular value $\mu_{\mathcal{U}}(A)$ is simply equal to the maximum singular value of the matrix A.

The following theorem is well known as the main loop theorem in robust system analysis and synthesis, which relates the structured singular value of a matrix with that of its submatrix and has been proven very helpful in robustness analysis for a system with several model uncertainties [4,5].

Theorem 2.3. *For two prescribed uncertainty structures* $\mathbf{\Delta}_1$ *and* $\mathbf{\Delta}_2$*, define another uncertainty structure* $\mathbf{\Delta}$ *as* $\mathbf{\Delta} = \{\Delta \,|\, \Delta = \mathbf{diag}\{\Delta_1,\ \Delta_2\},\ \Delta_1 \in \mathbf{\Delta}_1,\ \Delta_2 \in \mathbf{\Delta}_2\}$*. Assume that a matrix* A *has a dimension compatible with the uncertainty structure* $\mathbf{\Delta}$ *and partition it as* $A = \left[A_{ij}|_{i,j=1}^{2}\right]$ *with each of its submatrices* A_{ij}*,* $i, j = 1, 2$*, having compatible dimensions. Then, a necessary and sufficient condition for* $\mu_{\mathbf{\Delta}}(A) < 1$ *is that*

$$\mu_{\mathbf{\Delta}_2}(A_{22}) < 1 \quad \textit{and} \quad \max_{\Delta_2 \in \mathcal{B}\mathbf{\Delta}_2} \mu_{\mathbf{\Delta}_1}(\mathcal{F}_l(A,\ \Delta_2)) < 1.$$

Although the structured singular value of a matrix is in general very hard to compute, it has been proven to be a powerful tool in robust system analysis and synthesis, combined with properties of the linear fractional transformation given in Section 2.3. In addition, various upper and lower bounds have been derived for it [4,5].

The next results are also well known in linear algebra [7].

Lemma 2.2. *For arbitrary matrices* $A,\ B,\ C,\ D$ *with compatible dimensions, assume that all the involved matrix inverses exist. Then*

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix}\begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}\begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix}\begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix},\end{aligned} \tag{2.2}$$

$$[A + CBD]^{-1} = A^{-1} - A^{-1}C[B^{-1} + DA^{-1}C]^{-1}DA^{-1}, \tag{2.3}$$

$$A(I + BA)^{-1} = (I + AB)^{-1}A. \tag{2.4}$$

The following results are related to matrix optimizations, which are well known in matrix analysis and linear estimations and can be straightforwardly proved through square completions [7,8]. These results are extensively utilized in the derivations of an optimal estimator, an optimal controller, and so on [9,10].

Lemma 2.3. *For real-valued matrices* $S_{ij}|_{i,j=1}^{2}$ *with compatible dimensions, assume that the matrix* S_{11} *is symmetric and positive definite and the matrix* S_{22} *is symmetric. Define the matrix-valued function* $J(K)$ *as*

$$J(K) = [-K \ \ I]\begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix}\begin{bmatrix} -K^T \\ I \end{bmatrix}.$$

Then

$$K = S_{12}^T S_{11}^{-1}$$

is the unique optimizer that makes the matrix-valued function $J(K)$ achieve its minimal value

$$S_{22} - S_{12}^T S_{11}^{-1} S_{12}.$$

The following result, which is well known as the Schur complement formula and widely utilized in system analysis and synthesis, reveals relations about negative definiteness between a Hermitian matrix and its submatrices.

Lemma 2.4. *For a Hermitian matrix S with partition $S = \left[S_{ij}\big|_{i,j=1}^{2} \right]$, in which both S_{11} and S_{22} are square, the following three statements are equivalent.*

- *the matrix S is negative definite;*
- *both S_{11} and $S_{22} - S_{12}^H S_{11}^{-1} S_{12}$ are negative definite matrices;*
- *both S_{22} and $S_{11} - S_{21}^H S_{22}^{-1} S_{21}$ are negative definite matrices.*

Related to a matrix, two linear spaces are extensively utilized in systems, control theory, and signal processing. One is its null space, whereas the other is the space spanned by this matrix.

Definition 2.5 (Null Space). *Given a matrix A, its null space is constituted of all vectors α satisfying $A\alpha = 0$.*

The null space of a matrix is often denoted by $\mathcal{N}ull(\cdot)$. This subspace is sometimes also called the kernel of a matrix.

Definition 2.6 (Span). *For an arbitrary $m \times n$ matrix A, denote its column vectors by α_i, $i = 1, 2, \ldots, n$. The linear space spanned by the matrix A, usually denoted by $\mathcal{S}pan(A)$, is defined as*

$$\mathcal{S}pan(A) = \left\{ x \;\middle|\; x = \sum_{i=1}^{n} c_i \alpha_i \right\}.$$

This space is also widely called the image space or simply the image of the matrix A. In the definition, the scalars c_i, $i = 1, 2, \ldots, n$, belong to $\mathcal{R}$ or $\mathcal{C}$, according to the associated problem under investigation. It can be easily proven that both the null space and the span of a matrix are subspaces.

2.1.1 Vector and Matrix Norms

In system analysis and synthesis, the extensively adopted vector norm is the p-norm with $p = 1, 2, \infty$. Obviously, when x is an n-dimensional real vector with its elements x_i,

$i = 1, 2, \ldots, n$, we have that

$$||x||_1 = \sum_{i=1}^{n} |x_i|, \quad ||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Moreover, it can be easily shown that

$$||x||_\infty = \max_{1 \le i \le n} |x_i|.$$

On the other hand, it can be straightforwardly proven that the set consisting of $n \times m$-dimensional real matrices is also a linear space. Hence, the norm of a matrix can be defined in two different ways. One is to consider it as an element of a linear space. The other one is to consider it as a mapping. When the latter is adopted, the associated norm is usually called an induced norm. Particularly, when a vector is measured by its p-norm or q-norm, the induced matrix norm associated with Eq. (2.1) is

$$||T||_{q,p} = \sup_{x \neq 0} \frac{||Tx||_q}{||x||_p}.$$

When $p = q$, $||T||_{q,p}$ is usually abbreviated as $||T||_p$. Let t_{ij} denote the ith row jth column element of the matrix T. Then, when $p = 1, 2, \infty$, the associated induced matrix norm can be proven to be respectively

$$\begin{aligned} ||T||_1 &= \max_{1 \le i \le m} \sum_{j=1}^{n} |t_{ji}| \quad \text{(column summation)}, \\ ||T||_2 &= \sqrt{\lambda_{\max}(T^T T)} = \sigma_{\max}(T), \\ ||T||_\infty &= \max_{1 \le j \le n} \sum_{i=1}^{m} |t_{ji}| \quad \text{(row summation)}, \end{aligned}$$

in which $\lambda_{\max}(\cdot)$ is the maximum eigenvalue of a matrix, whereas $\sigma_{\max}(\cdot)$ is its greatest singular value. When the associated vector and matrix are complex valued, the corresponding vector norm and induced matrix norm can be defined.

When a matrix is considered to be an element of a linear space, an extensively used norm is the so-called Frobenius norm, which is denoted by $||T||_F$ and defined as

$$||T||_F = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} |t_{ji}|^2}.$$

However, this norm is not an induced one, which means that it may not be very convenient to use it in property analysis for signal transfers.

When a set under investigation is constituted of positive definite matrices (PDM), there is a quite specific distance function, which has been proven useful in property analysis for recursive estimations.

Definition 2.7 (Riemannian Distance). *Let P and Q be two $n \times n$-dimensional positive definite matrices. Let $\lambda_i(PQ^{-1})$ denote the ith eigenvalue of the matrix PQ^{-1}. The Riemannian distance between these two matrices, denoted $\delta(P, Q)$, is defined as*

$$\delta(P, Q)= \sqrt{\sum_{i=1}^{n} \log^2 \lambda_i(PQ^{-1})}.$$

Note that although the product of two positive definite matrices is usually even not symmetric, each of its eigenvalues has been proven to be positive [1,7]. Since the inverse of a positive definite matrix is still positive definite, it can be declared that this distance is well defined.

An attractive property of this distance is its invariance under conjugacy transformations and inversions. It is now also known that when equipped with this distance, the space of $n \times n$-dimensional PDMs is complete. This metric, although not widely known, has been recognized very useful for many years in studying asymptotic properties of the Kalman filtering with random system matrices [11]. Its effectiveness in studying asymptotic properties of the Kalman filter with intermittent observations (KFIO) has been discovered in [12]. This distance has also played an important role in analyzing properties of a robust recursive state estimator [13,14].

2.1.2 Hamiltonian Matrices and Distance Among Positive Definite Matrices

Hamiltonian matrices play important roles in systems and control theory. In this subsection, a definition is given for this matrix, and some of their important properties are discussed, which are relevant to topics in this book. These properties are helpful in investigating convergence characteristics of recursive operations defined by a homographic transformation. The latter is given in Section 2.3 and denoted by $\mathbf{H}_m(*, \#)$.

Definition 2.8 (Hamiltonian Matrix). *A matrix $\Phi = [\Phi_{ij}|_{i,j=1}^{2}]$ with $\Phi_{ij} \in \mathcal{R}^{n \times n}$, $i, j = 1, 2$, is said to be Hamiltonian if it satisfies*

$$\Phi^T J \Phi = J \qquad \textit{with} \qquad J = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & 0_{n\times n} \end{bmatrix}.$$

Hamiltonian matrices have found various applications in settling important issues in systems and control. One example is solving a Riccati equation, which is extensively used in controller design and state estimations.

According to properties of the submatrices Φ_{ij}, $i, j = 1, 2$, of the Hamiltonian matrix Φ, Hamiltonian matrices can be further divided into several subclasses. In particular, the following four subsets of Hamiltonian matrices $\mathcal{H}$, $\mathcal{H}_l$, $\mathcal{H}_r$, and $\mathcal{H}_{lr}$ are widely adopted:

$$\begin{aligned}
\mathcal{H} = & \Big\{\Phi \Big| \Phi = \left[\Phi_{ij}\right]_{i,j=1}^{2}, \ \Phi_{ij} \in \mathcal{R}^{n\times n}, \ \Phi^T J \Phi = J, \ \Phi_{11} \text{ invertible}, \ \Phi_{12}\Phi_{11}^T \geq 0, \\
& \Phi_{11}^T \Phi_{21} \geq 0 \Big\}, \\
\mathcal{H}_{lr} = & \Big\{\Phi \Big| \Phi \in \mathcal{H}, \ \Phi_{12}\Phi_{11}^T > 0, \ \Phi_{11}^T\Phi_{21} > 0 \Big\}, \\
\mathcal{H}_l = & \Big\{ \Phi \ \Big| \ \Phi \in \mathcal{H}, \ \Phi_{11}^T \Phi_{21} > 0 \Big\}, \\
\mathcal{H}_r = & \Big\{ \Phi \ \Big| \ \Phi \in \mathcal{H}, \ \Phi_{12}\Phi_{11}^T > 0 \Big\}.
\end{aligned}$$

From their definitions the following relations are obvious:

$$\mathcal{H}_l \subset \mathcal{H}, \ \ \mathcal{H}_r \subset \mathcal{H}, \ \ \mathcal{H}_{lr} \subset \mathcal{H}, \ \ \mathcal{H}_{lr} = \mathcal{H}_r \cap \mathcal{H}_l.$$

The following properties of Hamiltonian matrices are given in [11,14], which reveal further relations among these four subsets of Hamiltonian matrices and some relations between the matrix and its homographic transformation defined by a Hamiltonian matrix. In the establishment of these relations, the Riemannian metric is used, which is defined in the previous subsection for positive definite matrices.

Lemma 2.5. *Assume that all the involved matrices have compatible dimensions. Then, among elements of the sets $\mathcal{H}$, $\mathcal{H}_l$, $\mathcal{H}_r$, and $\mathcal{H}_{lr}$ and (semi)positive definite matrices (PDM), we have the following relations:*

- *if $\Phi(1) \in \mathcal{H}$ and $\Phi(2) \in \mathcal{H}$ (or $\mathcal{H}_l$, or $\mathcal{H}_r$, or $\mathcal{H}_{lr}$), then both $\Phi(2)\Phi(1)$ and $\Phi(1)\Phi(2)$ belong to $\mathcal{H}$ (or $\mathcal{H}_l$, or $\mathcal{H}_r$, or $\mathcal{H}_{lr}$);*
- *Assume that $\Phi(i) = \left[\Phi_{pq}(i)\big|_{p,q=1}^{2}\right] \in \mathcal{H}$, $i = 1, 2, \cdots, m$. Then*
  - *$\prod_{i=m}^{1} \Phi(i) \in \mathcal{H}_l$ if and only if*

$$\det\left\{\Phi_{11}^T(1)\Phi_{21}(1) + \sum_{i=2}^{m}\left[\left(\prod_{k=1}^{i}\Phi_{11}^T(k)\right)\Phi_{21}(i)\left(\prod_{k=i-1}^{1}\Phi_{11}(k)\right)\right]\right\} \neq 0; \qquad (2.5)$$

- $\prod_{i=m}^{1}\Phi(i)\in\mathcal{H}_r$ *if and only if*

$$\det\left\{\sum_{i=1}^{m-1}\left[\left(\prod_{k=m}^{i+1}\Phi_{11}(k)\right)\Phi_{12}(i)\left(\prod_{k=i}^{m}\Phi_{11}^T(k)\right)\right]+\Phi_{12}(m)\Phi_{11}^T(m)\right\}\neq 0; \qquad (2.6)$$

- *Assume that* $\Phi\in\mathcal{H}$. *Then, for arbitrary* $X\geq 0$, $\mathbf{H}_m(\Phi,\ X)$ *is well defined and is at least semipositive definite. If, in addition,* $\det(X)\neq 0$, *then* $\det\{\mathbf{H}_m(\Phi,\ P)\}$ *is also positive;*
- *Assume that* $\Phi\in\mathcal{H}_{lr}$. *Then, for every* $X\geq 0$, $\mathbf{H}_m(\Phi,\ X)$ *is a PDM;*
- *Assume that* $\Phi\in\mathcal{H}$. *Then,* $\delta\{\mathbf{H}_m(\Phi,\ X),\ \mathbf{H}_m(\Phi,\ Y)\}\leq\delta(X,\ Y)$ *whenever* $X,\ Y>0$;
- *Assume that* $\Phi\in\mathcal{H}_l$ *or* $\Phi\in\mathcal{H}_r$. *Then, for any* $X,\ Y>0$, $\delta\{\mathbf{H}_m(\Phi,\ X),\ \mathbf{H}_m(\Phi,\ Y)\}<\delta(X,\ Y)$;
- *Assume that* $\Phi\in\mathcal{H}_{lr}$. *Then, there exists* $\rho(\Phi)$ *belonging to* $(0,1)$ *such that, for all* $X,\ Y>0$, $\delta\{\mathbf{H}_m(\Phi,\ X),\ \mathbf{H}_m(\Phi,\ Y)\}\leq\rho(\Phi)\delta(X,\ Y)$.

The lemma reveals that when a homographic transformation is defined by a Hamiltonian matrix, under the Riemannian metric of Definition 2.7, the distance between two positive definite matrices does not increase after the transformation. This is quite attractive in solving many theoretical problems, such as the H_∞ optimal controller design [15], the asymptotic properties of the Kalman filter with random coefficients [11,13,14], and so on.

2.2 Generalized Inverse of a Matrix

The inverse of a matrix is extensively utilized in solving various engineering problems. When a square matrix is of full rank, its inverse exists and is unique. Under several important situations, a square matrix is not of full rank, but its inverse is still required. In this case, a generalized inverse is widely adopted, which is usually denoted by $A^\dagger$.

Definition 2.9 (**Generalized Inverse**). *A matrix $A^\dagger$ is said to be the generalized inverse of the matrix A if it satisfies the following four equalities:*

$$AA^\dagger A=A,\quad A^\dagger AA^\dagger=A^\dagger,$$
$$\left(AA^\dagger\right)^H=AA^\dagger,\quad \left(A^\dagger A\right)^H=A^\dagger A.$$

If a matrix is invertible, then obviously its inverse satisfies these four equalities, which means that this definition is consistent with that of the matrix inverse. Differently from the inverse of a matrix, which not always exists, the generalized inverse of a matrix always exists. In addition, in the computations of the generalized inverse of a matrix, the matrix is even not required to be square. As a matter of fact, the generalized inverse of a matrix can be uniquely determined by its singular value decomposition, which is possible for every matrix [7,16].

Theorem 2.4. *For a prescribed $m \times n$-dimensional complex-valued matrix A, denote its singular value decomposition by*

$$A = U\Sigma V^H, \quad \Sigma = \begin{bmatrix} \Sigma_* & 0 \\ 0 & 0 \end{bmatrix},$$

in which U is an $m \times m$-dimensional unitary matrix, V is an $n \times n$-dimensional unitary matrix, and $\Sigma_* = \mathbf{diag}\{\sigma_i|_{i=1}^r\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Here, $r \leq \min\{m, n\}$ is the rank of the matrix A. Then, its generalized inverse is uniquely determined by

$$A = V \begin{bmatrix} \Sigma_*^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^H.$$

Using the generalized inverse of a matrix, conditions for the existence of a solution to a set of linear algebraic equations can be concisely expressed, as well as all the solutions if there exists one.

Theorem 2.5. *Let A, B, and Y be some prescribed matrices of dimensions $n_1 \times n_2$, $n_3 \times n_4$, and $n_1 \times n_4$, respectively. Then, there exists an $n_2 \times n_3$-dimensional matrix X such that*

$$AXB = Y$$

if and only if the matrices A, B, and Y satisfy

$$(I_{n_1} - AA^\dagger)Y = 0 \quad \text{and} \quad Y(I_{n_3} - B^\dagger B) = 0.$$

When these conditions are satisfied, all the solutions to the above equation can be parameterized as

$$X = A^\dagger Y B^\dagger + Z - A^\dagger AZBB^\dagger, \tag{2.7}$$

where Z is an arbitrary $n_2 \times n_3$-dimensional matrix.

When both A and B are invertible matrices, it can be directly verified that the conditions in the theorem are always satisfied. In this case, all the solutions to the equation $AXY = B$ given by Eq. (2.7) reduce to $X = A^{-1}YB^{-1}$, which is consistent with the results obtained through directly solving this equation.

The results of Theorem 2.5 include those on the equation $AX = Y$ and the equation $XB = Y$ as a particular case. Particularly, the equation $AX = Y$ is a special situation of the equation in Theorem 2.5 with $B = I_{n_4}$. In this case, the condition $Y(I_{n_3} - B^\dagger B) = 0$ is always satisfied, whereas the condition $(I_{n_1} - AA^\dagger)Y = 0$ implies that each column of the matrix Y belongs to the space spanned by the columns of the matrix A.

2.3 Some Useful Transformations

In system analysis and synthesis, as well as in signal processing, there are two extensively adopted transformations. One is called linear fractional transformation, which is usually abbreviated as LFT. The other is called a homographic transformation and is often abbreviated as HM. Under some mild conditions, these two transformations can be converted to each other. However, both of them have their own properties, which makes them convenient to be applied to dealing with different problems.

Definition 2.10 (Homographic Transformation). *Given a matrix* $\Phi = \left[\Phi_{ij}\big|_{i,j=1}^{2}\right]$ *with its submatrices having compatible dimensions, the homographic transformation* $\mathbf{H}_m(\Phi,\ X)$ *of a matrix* X *with the matrix* Φ *is defined as*

$$\mathbf{H}_m(\Phi,\ X) = [\Phi_{11}X + \Phi_{12}][\Phi_{21}X + \Phi_{22}]^{-1},$$

where the matrix $\Phi_{21}X + \Phi_{22}$ *is assumed to be square and of full rank.*

Attractive properties of this transformation include its invariance and simplicity under cascade connections, which is given in the following lemma. Another property of this transformation is that contractiveness of the matrix X can be kept, provided that some conditions are satisfied by the matrix Φ. These properties have been proven useful in analyzing asymptotic behaviors of recursive estimators and iterative computations, and in H_∞ controller design using the so-called chain scattering approach.

Lemma 2.6. *Assume that the dimensions of the matrices* $\Phi(1)$, $\Phi(2)$*, and* X *are compatible with each other. Moreover, assume that all the involved matrix inverses exist. Then*

$$\mathbf{H}_m(\Phi(2),\ \mathbf{H}_m(\Phi(1),\ X)) = \mathbf{H}_m(\Phi(2)\Phi(1),\ X).$$

This property follows immediately from the concatenation property of a so-called chain scattering representation of the homographic transformation and can be established through straightforward algebraic manipulations. The details can be found, for example, in [9,15].

From the definition of the homographic transformation it is clear that $X = \mathbf{H}_m(I,\ X)$. On the basis of this property and Lemma 2.6, it can be further proven that if the matrix Φ is invertible, then from $Y = \mathbf{H}_m(\Phi,\ X)$ we can get $X = \mathbf{H}_m(\Phi^{-1},\ Y)$.

However, some difficulties may arise if the homographic transformation is adopted in investigating system connections like addition, cascade interconnection, feedback interconnection, and so on. To illustrate possible difficulties, we discuss the cascade interconnection of

$\mathbf{H}_m(\Phi,\ X)$ and $\mathbf{H}_m(\Psi,\ Y)$. Assume that all the involved matrix inverses exist. Then direct algebraic manipulations show that

$$\begin{aligned}
& \mathbf{H}_m(\Phi,\ X)\mathbf{H}_m(\Psi,\ Y) \\
= \ & [\Phi_{11}X+\Phi_{12}][\Phi_{21}X+\Phi_{22}]^{-1}\times[\Psi_{11}Y+\Psi_{12}][\Psi_{21}Y+\Psi_{22}]^{-1} \\
= \ & \Phi_{12}\Phi_{22}^{-1}\Psi_{12}\Psi_{22}^{-1}+\begin{bmatrix}\Phi_{11}-\Phi_{12}\Phi_{22}^{-1}\Phi_{21} & \Phi_{12}\Phi_{22}^{-1}(\Psi_{11}-\Psi_{12}\Psi_{22}^{-1}\Psi_{21})\end{bmatrix}\begin{bmatrix} X & \\ & Y\end{bmatrix} \\
& \times\left\{I-\begin{bmatrix}-\Phi_{22}^{-1}\Phi_{21} & \Phi_{22}^{-1}(\Psi_{11}-\Psi_{12}\Psi_{22}^{-1}\Psi_{21}) \\ 0 & -\Psi_{22}^{-1}\Psi_{21}\end{bmatrix}\begin{bmatrix} X & \\ & Y\end{bmatrix}\right\}^{-1}\begin{bmatrix}\Phi_{22}^{-1}\Psi_{12}\Psi_{22}^{-1} \\ \Psi_{22}^{-1}\end{bmatrix}.
\end{aligned}$$

The rightmost expression in this equation is in fact a linear fractional transformation, which will be defined immediately, of the matrix $\mathbf{diag}\{X,\ Y\}$. From this expression it may not be difficult to understand that a cascade connection of two homographic transformations is generally hard to be expressed by a compact homographic transformation.

In some literature, a homographic transformation is defined as

$$\mathbf{H}_m(\Phi,\ X)=[\Phi_{11}X+\Phi_{12}]^{-1}[\Phi_{21}X+\Phi_{22}]$$

using matrices X and $\Phi=\left[\Phi_{ij}|_{i,j=1}^{2}\right]$ with compatible dimensions. To clarify their differences, the former is called a right homographic transformation, whereas the latter is called a left homographic transformation. It can be proven that these two homographic transformations can be converted to each other under some mild conditions. In this book, only the former is adopted. Therefore, there will be no confusion if it is simply called as a homographic transformation.

Another well-adopted transformation in system analysis and synthesis is called a linear fractional transformation (LFT). In particular, there are two LFTs. One is called a lower LFT, and the other is called an upper LFT.

Definition 2.11 (Linear Fractional Transformation). *Given a matrix* $\Phi=\left[\Phi_{ij}|_{i,j=1}^{2}\right]$ *with its submatrices having compatible dimensions, the upper LFT of a matrix X, denoted* $\mathbf{F}_u(\Phi, X)$*, is defined as*

$$\mathbf{F}_u(\Phi,\ X)=\Phi_{22}+\Phi_{21}X[I-\Phi_{11}X]^{-1}\Phi_{12}.$$

Moreover, the lower LFT of this matrix, denoted $\mathbf{F}_l(\Phi, X)$*, is defined as*

$$\mathbf{F}_l(\Phi,\ X)=\Phi_{11}+\Phi_{12}X[I-\Phi_{22}X]^{-1}\Phi_{21}.$$

Here, all the involved matrix inverses are assumed to exist.

Obviously, an upper LFT can also be expressed as a lower LFT, and vice versa. More precisely, for a prescribed matrix $\Phi = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}$, define the matrix $\bar{\Phi} = \begin{bmatrix} \Phi_{22} & \Phi_{21} \\ \Phi_{12} & \Phi_{11} \end{bmatrix}$. Then, it is obvious from the definitions of the lower and upper LFTs that

$$\mathbf{F}_u(\Phi,\ X) = \mathbf{F}_l(\bar{\Phi},\ X).$$

A linear fractional transformation is a natural way to describe relations among a plant, a controller, and a closed-loop transfer function matrix in a feedback control system. It is also well known that addition, multiplication, and feedback connection of LFTs can still be expressed by an LFT. It has also been proven that under some weak conditions, the inverse of an LFT is still an LFT. These properties make LFT very convenient in control system analysis and synthesis.

For example, assume that the matrices $\Phi = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}$ and $\Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$ have compatible dimensions, as well as their submatrices. Then, when the matrices X and Y are also compatible in their dimensions and all the involved matrix inverses exist, it can be straightforwardly proven that

$$\mathbf{F}_u(\Phi,\ X)\mathbf{F}_u(\Psi,\ Y) = \mathbf{F}_u(\Theta,\ Z),$$

in which

$$\Theta = \left[\begin{array}{cc|c} \Phi_{11} & \Phi_{12}\Psi_{21} & \Phi_{12}\Psi_{22} \\ 0 & \Psi_{11} & \Psi_{12} \\ \hline \Phi_{21} & \Phi_{22}\Psi_{21} & \Phi_{22}\Psi_{22} \end{array}\right], \quad Z = \begin{bmatrix} X & \\ & Y \end{bmatrix}.$$

However, although the composite of two LFTs can still be expressed by an LFT, it generally has no concise form as that of the homographic transformation, which is given in Lemma 2.6. More precisely, assume that dimensions are compatible for each other in all the involved matrix operations and that the existence is also guaranteed for all the involved matrix inverses. Then, some tedious but direct algebraic manipulations show that

$$\mathbf{F}_l\{\Phi,\ \mathbf{F}_l(\Psi,\ X)\} = \mathbf{F}_l(\Theta,\ X),$$

where

$$\begin{aligned} \Theta &= \begin{bmatrix} \Phi_{11} + \Phi_{12}\Psi_{11}(I - \Phi_{22}\Psi_{11})^{-1}\Phi_{21} & \Phi_{12}\Psi_{12} + \Phi_{12}\Psi_{11}(I - \Phi_{22}\Psi_{11})^{-1}\Phi_{22}\Psi_{12} \\ \Psi_{21}(I - \Phi_{22}\Psi_{11})^{-1}\Phi_{21} & \Psi_{22} + \Psi_{21}(I - \Phi_{22}\Psi_{11})^{-1}\Phi_{22}\Psi_{12} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{F}_l(\Phi,\ \Psi_{11}) & \Phi_{12}(I - \Psi_{11}\Phi_{22})^{-1}\Psi_{12} \\ \Psi_{21}(I - \Phi_{22}\Psi_{11})^{-1}\Phi_{21} & \mathbf{F}_u(\Psi,\ \Phi_{22}) \end{bmatrix}, \end{aligned}$$

which is quite complicated, compared with the product of the matrices Φ and Ψ. In addition, even when both matrices Φ and Ψ are Hamiltonian, there is no guarantee that the matrix Θ is also Hamiltonian.

As a matter of fact, the expression for the matrix Θ is called the Redheffer star product between the matrices Φ and Ψ, which is able to include most of system interconnections, such as cascade connection, feedback connection, additive connection, and so on as particular cases [5,15,17].

In addition, under some conditions, a Homographic transformation can also be expressed as a lower/upper LFT, and vice versa. In particular, assume that Φ_{22} is invertible. Define the

$$\Psi = \begin{bmatrix} \Phi_{12}\Phi_{22}^{-1} & \Phi_{11} - \Phi_{12}\Phi_{22}^{-1}\Phi_{21} \\ \Phi_{22}^{-1} & -\Phi_{22}^{-1}\Phi_{21} \end{bmatrix}.$$

Then direct algebraic manipulations show that

$$\mathbf{H}_m(\Phi,\ X) = \mathbf{F}_l(\Psi,\ X).$$

On the contrary, assume that Φ_{21} is invertible. Define the matrix

$$\Psi = \begin{bmatrix} \Phi_{11}\Phi_{21}^{-1} & \Phi_{12} - \Phi_{11}\Phi_{21}^{-1}\Phi_{22} \\ \Phi_{21}^{-1} & -\Phi_{21}^{-1}\Phi_{22} \end{bmatrix}.$$

Then it can also be shown through direct algebraic manipulations that

$$\mathbf{F}_l(\Phi,\ X) = \mathbf{H}_m(\Psi,\ X).$$

However, in these three different transformations, it is now extensively understood that one transformation is more convenient than the other two transformations in dealing with a particular problem. As an example, the homographic transformation is used in this book for studying convergence properties of some recursive state estimation algorithms, whereas the linear fractional transformation is adopted in the investigations on controllability/observability of a large-scale networked system.

In the above three transformations, all the matrices can be replaced by transfer function matrices, which leads to the corresponding transformations of a transfer function matrix.

2.4 Set Function and Submodularity

For a given set $\mathcal{V}$ that only has M finite elements, denoted $v_i|_{i=1}^{M}$, a set function $f(\mathcal{S})$ is a mapping that assigns a real number to each subset of the set $\mathcal{V}$, that is,

$$f:\ 2^{\mathcal{V}} \longrightarrow \mathcal{R}. \tag{2.8}$$

For a set $\mathcal{S}$, let $|\mathcal{S}|$ denote the number of its elements. Many design problems encountered in engineering, including the sensor placements that will be discussed in the following Chapter 11 of this book, can be mathematically described as

$$\max_{\mathcal{S}\subseteq\mathcal{V},\ |\mathcal{S}|\leq k} f(\mathcal{S}), \tag{2.9}$$

where the function $f(\cdot)$ stands for a composite description of some performances that must be maximized in the designs, whereas the set $\mathcal{V}$ consists of possible selections, and k is the permissible number of the maximal selections.

Submodularity of a set function is extensively utilized in combinatorial optimizations. More precisely, this property plays an essential role in combinatorial optimizations similar to that of convexity in the optimization of a function with continuous variables and shares some attractive characteristics with a concave function [18]. Interesting properties of submodularity include its preservation under various operations, supported by mathematical theories that are application oriented and mathematically rigorous, numerically feasible optimization algorithms, and so on.

Definition 2.12. *A set function $f: \ 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ is said to be submodular if for each subset pair $(\mathcal{A},\ \mathcal{B})$ satisfying $\mathcal{A}\subseteq\mathcal{B}\subseteq\mathcal{V}$ and each element α that does not belong to the set $\mathcal{B}$, we have the inequality*

$$f\left(\mathcal{A}\bigcup\{\alpha\}\right) - f(\mathcal{A}) \geq f\left(\mathcal{B}\bigcup\{\alpha\}\right) - f(\mathcal{B}) \tag{2.10}$$

If this inequality is valid when the relation "$\geq$" is replaced by the relation "$\leq$", then the associated set function is called supermodular. If a set function is both submodular and supermodular, then it is called modular.

It has been proven that inequality (2.10) is equivalent to that the inequality

$$f\left(\mathcal{A}\bigcap\mathcal{B}\right) + f\left(\mathcal{A}\bigcup\mathcal{B}\right) \leq f(\mathcal{A}) + f(\mathcal{B})$$

for all subsets $\mathcal{A}$ and $\mathcal{B}$ satisfying the restriction $\mathcal{A}\subseteq\mathcal{B}\subseteq\mathcal{V}$. Submodular functions, supermodular functions, and modular functions are analogous to convex functions, concave functions, and linear functions with variables taking continuous values. Particularly, it has been proven that in a modular function, each element of a subset in the set $\mathcal{V}$ contributes independently to the value of the function, which is very attractive in optimizations. More precisely, we have the following results [18,19].

Lemma 2.7. *A set function $f: \ 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ is modular if and only if there exists a function $w(\cdot)$, usually called a weight function, such that for each subset $\mathcal{S}$ of the set $\mathcal{V}$, we have the equality*

$$f(\mathcal{S}) = w(\emptyset) + \sum_{s\in\mathcal{S}} w(s), \tag{2.11}$$

where $\emptyset$ is the empty set.

When a set function is modular, an algorithm can always be developed that optimizes its value and whose computational complexity increases only linearly with the increment of the number of the elements in the set $\mathcal{V}$. In fact, the associated optimization can be simply settled through computing values of the set function at each element of the set $\mathcal{V}$, sorting the computed values, and selecting the elements that lead to the first k largest or smallest function values.

A concept closely related to submodularity is monotone increasing, which is mathematically clearer to be understood.

Definition 2.13. *If for each subset pair $(\mathcal{A},\ \mathcal{B})$ satisfying $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$, the set function $f:\ 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ satisfies*

$$f(\mathcal{A}) \leq f(\mathcal{B}), \tag{2.12}$$

then this set function is called monotone increasing. If inequality (2.12) is valid with "$\leq$" replaced by "$\geq$" for all subsets $\mathcal{A}$ and $\mathcal{B}$ of the set $\mathcal{V}$, then this set function is called monotone decreasing.

The following lemma establishes a relation between submodular set functions and monotone decreasing set functions.

Lemma 2.8. *For an element v of the set $\mathcal{V}$ and a given set function $f:\ 2^{\mathcal{V}} \longrightarrow \mathcal{R}$, define the derived set function $f_v:\ 2^{\mathcal{V}\backslash\{v\}} \longrightarrow \mathcal{R}$ as*

$$f_v(\mathcal{S}) = f\left(\mathcal{S}\bigcup\{v\}\right) - f(\mathcal{S}), \tag{2.13}$$

where $\mathcal{S}$ is any subset of the set $\mathcal{V}\backslash\{v\}$. Then, the set function $f(\cdot)$ is submodular if and only if for each $v \in \mathcal{V}$, the derived set function $f_v(\cdot)$ is monotone decreasing.

It can be simply shown from the definition that a nonnegative weighted sum of several supermodular functions is still a supermodular function. Moreover, assume that a supermodular set function $f:\ 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ takes only nonnegative values and satisfies $f(\mathcal{S}) \geq f(\mathcal{T})$ whenever $\mathcal{S} \subseteq \mathcal{T}$. If $g(*)$ is a nondecreasing convex function that is differentiable and defined on the set $\mathcal{R}$, then the composite function $g(f(\star))$ is also a supermodular function that maps $2^{\mathcal{V}}$ to $\mathcal{R}$.

In general, maximization of a set function that is both monotone increasing and submodular is NP-hard [20]. On the other hand, it has also been proven that the so-called greedy heuristic method, which selects the element from all the candidates that maximize the increment of the set function in each step, leads to a solution that is close to the optimizer. In particular, the greedy heuristic method selects elements from the set $\mathcal{V}$ as follows.

Algorithm 2.4.1. The greedy heuristic method for maximizing a set function

1. *Set $\mathcal{S}_0 = \emptyset$.*
2. *Compute the conditional increment of the set function $f(\mathcal{S})$ as*

$$\delta(v|\mathcal{S}_i) = f\left(\{v\}\bigcup\mathcal{S}_i\right) - f(\mathcal{S}_i)$$

   for each $v \in \mathcal{S}\backslash\mathcal{S}_i$. Take v_i as

$$v_i = \arg\max_{v\in\mathcal{S}\backslash\mathcal{S}_i} \delta(v|\mathcal{S}_i).$$

3. *Let $\mathcal{S}_{i+1} = \mathcal{S}_i \bigcup\{v_i\}$ and $i \Longrightarrow i+1$.*
4. *If $i \geq k$, then end the computation and output the set $\mathcal{S}_i$. If $i < k$, then go to Step 2 and start another round.*

Moreover, the following approximation results have been proven, which give a universal upper bound on the relative error of the algorithm [20,21].

Theorem 2.6. *Assume that the set function $f(\mathcal{S})$ in the maximization problem of Eq. (2.9) is submodular and monotone increasing. Let $f^\star$ and $f^{[greedy]}$ denote respectively its optimal value and the value obtained through the greedy heuristic method given by Algorithm 2.4.1. Then*

$$\frac{f^\star - f^{[greedy]}}{f^\star - f(\emptyset)} \leq \left(1-\frac{1}{k}\right)^k \leq \frac{1}{e}. \tag{2.14}$$

These results are quite attractive, as they reveal that the greedy heuristic method, which is computationally simple and *only* seeks a local optimum in each element selection in the sense of maximizing *only* the increment of the cost function, usually works approximately well for a complicated combinatorial problem.

2.5 Probability and Random Process

In the description of uncertainties, an extensively adopted and efficient approach is probability, in which a random variable is used to represent uncertainties. When a series of random variables is to be investigated, a concept of a random process is usually introduced to deal with their properties. This section gives some basic concepts and results on probability and random process.

If the result of an experiment cannot be determined before performing the experiment, this experiment is called a random experiment. The set consisting of all possible results of a random experiment is called a sample space, which is usually denoted by Ω. A set constituted from some subsets of the sample space Ω is called a class, which is usually denoted by $\mathcal{F}$. A nonempty class $\mathcal{F}$ is called a σ-algebra if it satisfies the following three conditions:

- Ω belongs to $\mathcal{F}$;
- if A belongs to $\mathcal{F}$, then, $X\backslash A$ also belongs to $\mathcal{F}$;
- if $A_1, A_2, \ldots$ belong to $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i$ also belongs to $\mathcal{F}$.

Moreover, $(\Omega,\ \mathcal{F})$ is called a measurable space.

A measure μ defined on a measurable space $(\Omega,\ \mathcal{F})$ is called a probability measure if the following three conditions are satisfied:

- for each A belonging to $\mathcal{F}$, $\mu(A) \geq 0$;
- $\mu(\Omega) = 1$;
- for all A_1, A_2 belonging to $\mathcal{F}$ and satisfying $A_1 \bigcap A_2 = \emptyset$, $\mu\left(A_1 \bigcup A_2\right) = \mu(A_1) + \mu(A_2)$.

When $(\Omega,\ \mathcal{F})$ is a measurable space and a measure μ is a probability measure over this measurable space, the triple $(\Omega,\ \mathcal{F},\ \mu)$ is called a probability space.

Assume that $X(\omega)$ is a real-valued function defined on the set Ω. If for an arbitrary real number α, the set $\{\,\omega \mid X(\omega) \leq \alpha\,\}$ belongs to the class $\mathcal{F}$, then the function $X(\omega)$ is called a random variable. Accordingly, the function $F(x)$ defined as

$$F(x) = \mu(X(\omega) \leq x)$$

is called the distribution function of this random variable, whereas a function $f(x)$ taking real nonnegative values and satisfying

$$F(x) = \int_{-\infty}^{x} f(u)du$$

for every real number x is called the probability density function of this random variable, usually abbreviated as PDF.

In the studies of random events, in addition to the probability of the occurrence of a random event, some numerical characteristics have also been extensively adopted. Among them, the most widely adopted ones appear to be the mathematical expectation $\mathbf{E}(X)$ and the variance $\mathbf{Var}(X)$ of a random variable $X(\omega)$, which are defined respectively as

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x)dx, \quad \mathbf{Var}(X) = \int_{-\infty}^{\infty} [x - \mathbf{E}(X)]^2 f(x)dx,$$

where $f(x)$ is the probability density function of the random variable $X(\omega)$.

When two or more random variables are under investigation, similar concepts have also been developed, such as the joint probability density function, conditional mathematical expectation, covariance, independence of two random variables, and so on. The details can be found in various standard textbooks, such as [22].

These concepts can be straightforwardly extended to situations where a random variable takes vectorial values.

A random process is a mathematical object that is usually defined as a collection of random variables/vectors indexed by a mathematical set, meaning that each random variable/vector of this random process is uniquely associated with an element in that set. In particular, a random process $X = \{ x(t),\ t \in \mathcal{T} \}$ can be interpreted as a parameterized family of random variables defined on the same probability space $(\Omega,\ \mathcal{F},\ \mu)$, in which the parameter t may be either a scalar or a vector. In fact, this parameter may be a temporal variable, or a spatial variable, or both of them. Moreover, the index set $\mathcal{T}$ may consist of only finitely or countably many elements and may even be uncountable. The set from which each random variable in the aforementioned collection takes values is called the state space of the random process.

Distribution functions, PDFs, joint PDFs, and so on can be defined for random variables belonging to the same random process or different random processes. Characteristics of a random process are completely determined by the class of its finite-dimensional probability distributions $F\left(t_i|_{i=1}^n,\ x_i|_{i=1}^n\right)$ defined as

$$F\left(t_i|_{i=1}^n,\ x_i|_{i=1}^n\right) = \mu\left(X(t_1) \le x_1,\ X(t_2) \le x_2,\ \cdots,\ X(t_n) \le x_n\right),$$

where, for each $i = 1, 2, \ldots, n$, $t_i \in \mathcal{T}$, and x_i is a real scalar or a real vector with a compatible dimension. Moreover, n is an arbitrary positive integer. In case that x_i is a vector, the inequality in this definition must be understood elementwise.

In addition to the distribution function and PDF, numerical characteristics like mathematical expectation, correlation, and so on are also extensively adopted in the description of properties of a random process and relations between different random processes. We list some widely adopted ones:

- mathematical expectation function: $m_X(t) = \mathbf{E}(X(t))$;
- variance function: $\mathbf{Var}_X(t) = \mathbf{E}([X(t) - m(t)]^2)$;
- autocorrelation function: $r_X(s,\ t) = \mathbf{E}(X(s)X(t))$;
- cross-correlation function: $r_{XY}(s,\ t) = \mathbf{E}(X(s)Y(t))$.

Differently from a random variable, all these numerical characteristics depend on the index t in general, which means that they are deterministic functions with the index t as their variable. These characteristics can be extended to situations in which a random process takes vector values. For instance, when the random process $X = \{ x(t),\ t \in \mathcal{T} \}$ takes values from an n-dimensional Euclidean real space, the definition of its variance function must be modified as $\mathbf{Var}_X(t) = \mathbf{E}([X(t) - m(t)][X(t) - m(t)]^T)$, which is in fact an $n \times n$-dimensional positive semidefinite matrix.

There are various special kinds of random processes that are widely adopted in engineering, economics, biology, and so on in the behavior description of actual signal and/or disturbances. The following are some of them that are used in the rest of this book.

- **Second-Order Process.** A random process is said to be a second-order process if the mathematical expectation of the squares of its absolute value exists at each $t \in \mathcal{T}$. This requirement is quite natural, since most of random processes are of finite energy, which leads to a finite value of the aforementioned quantity.
- **Stationary Process.** When a random process has the property that each its finite-dimensional probability distribution $F\left(t_i|_{i=1}^n,\ x_i|_{i=1}^n\right)$ does not depend on the particular values of $t_i|_{i=1}^n$, it is called stationary and sometimes strictly stationary. This property is usually adopted in the description of a random process in its steady state.
  When both the mathematical expectation function of a random process and its variance function do not depend on a particular value of the index, this random process is also called stationary. To differentiate it from the previous one, the terminology "weak stationary" is much more widely adopted. Compared with the distribution function-based requirements, these two numerical characteristic-based requirements are much easier to be verified using sampled data of a random process.
- **Ergodic Process.** A random process $X = \{ x(t),\ t \in \mathcal{T} \}$ is said to be ergodic with respect to a function $f(\cdot)$ if its ensemble average $\mathbf{E}(f(X(t)))$ equals to its index average, that is,

$$\mathbf{E}(f(X(t))) = \frac{1}{\mu(\mathcal{T})} \sum_{t \in \mathcal{T}} f(X(t)),$$

  where $\mu(\mathcal{T})$ is the measure of the index set $\mathcal{T}$.
  According to different forms of the function $f(\cdot)$, a random process is called respectively as mean-ergodic, mean-square ergodic in the first moment, autocovariance ergodic, mean-square ergodic in the second moment, and so on. Intuitively, a random process is said to be ergodic if its associated statistical properties can be estimated from one of its single but sufficiently long realization.
  Ergodicity is a quite attractive property in applications, such as those in biology, econometrics, signal processing, and so on since it means that any sufficiently large collection of random samples from a process can represent the ensemble average statistical properties of the process. The former can be obtained through one experiment, whereas the latter generally asks for a great amount of experiments, which is usually time consuming and sometimes economically expensive, and even prohibitive.

- **Martingale process.** If a discrete-time or continuous-time random process has the property that the expectation of its value at the next index is equal to its value at the current index conditional on all its previous values, then, this random process is called a martingale process. This process is often encountered in iterative estimations, adaptive estimations, adaptive controls, and so on.
- **Bernoulli process.** A simple, yet a widely adopted random process, is the so-called Bernoulli process. It is a sequence of independent and identically distributed (extensively abbreviated as iid) random variables, in which each random variable takes a value from the set $\{0,\ 1\}$. A Bernoulli process may be considered as a process of repeated coin flipping in which the coin is permitted to be unfair with its unfairness being restricted to be time invariant. This process is extensively adopted in the description of data transmissions in a communication network and so on.
- **Markov process.** This is another random process that is well encountered in applications. The most prominent characteristic of this random process is that its future is uniquely determined by its current state. Some important details are discussed in the next section.

Definition 2.14. *Let $\left\{x(k)\,|_{k=1}^{\infty}\right\}$ be a sequence of random vectors. This sequence is said to converge to a random vector x in the mean-square if the second absolute moments of both $\left\{x(k)\,|_{k=1}^{\infty}\right\}$ and x exist and*

$$\lim_{k\to\infty} E\left\{(x(k)-x)^T\,(x(k)-x)\right\}=0. \tag{2.15}$$

When there is a random vector x satisfying Eq. (2.15), this random vector is called the mean-square limit of the sequence $\left\{x(k)\,|_{k=1}^{\infty}\right\}$. In addition, the corresponding convergence is usually indicated as $x(k)\xrightarrow{m.s.}x$.

2.6 Markov Process and Semi-Markov Process

In signal processing, system analysis/synthesis, economy, biology, and so on, a well-encountered random process is called a Markov process. Differently from a general random process, a Markov process has the so-called memorylessness property, which is also extensively called the Markov property. Briefly, if we call each possible value of a random process $X=\{\,x(t),\ t\in\mathcal{T}\,\}$ at a fixed t, that is, the random variable $x(t)$, a state of this random process at that index, then the memorylessness property means that the conditional probability of any state of this random process, given its past observed values, depends only on the most recent past observed value. In other words, conditional on the present state of the random process, its future and past states are independent. This property is quite useful, as it implies that information contained in its current state is as rich as that contained in the process full history.

Hence, predictions of the future of a Markov process can be performed using only its current state without sacrificing any prediction performance.

When both t and $x(t)$ take discrete values, a Markov process is called a Markov chain. In other words, a Markov chain is a sequence of random variables $X_0, X_1, \cdots$ such that, for each $t \geq 1$,

$$\mathbf{Pr}\,(\,X_t = x_t \mid X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2},\ \cdots,\ X_0 = x_0\,) = \mathbf{Pr}\,(\,X_t = x_t \mid X_{t-1} = x_{t-1}\,)\,.$$

An attractive property of Markov chains is that its joint probability has a significantly simple form. More precisely, according to the Markov property and properties of conditional probability, it is straightforward to derive the following relations:

$$\begin{aligned}
& \mathbf{Pr}\,(\,X_t = x_t,\ \ X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2},\ \cdots,\ X_0 = x_0) \\
= \ & \mathbf{Pr}\,(\,X_t = x_t \mid X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2},\ \cdots,\ X_0 = x_0\,) \\
& \qquad \times \mathbf{Pr}\,(\,X_{t-1} = x_{t-1},\ \ X_{t-2} = x_{t-2},\ X_{t-3} = x_{t-3},\ \cdots,\ X_0 = x_0) \\
= \ & \mathbf{Pr}\,(\,X_t = x_t \mid X_{t-1} = x_{t-1}\,) \\
& \qquad \times \mathbf{Pr}\,(\,X_{t-1} = x_{t-1},\ \ X_{t-2} = x_{t-2},\ X_{t-3} = x_{t-3},\ \cdots,\ X_0 = x_0) \\
= \ & \cdots \\
= \ & \prod_{i=t}^{1} \mathbf{Pr}\,(\,X_i = x_i \mid X_{i-1} = x_{i-1}\,) \times \mathbf{Pr}\,(\,X_0 = x_0)\,.
\end{aligned} \tag{2.16}$$

If the possible values of $x(t)$ are countable, denote them by $\alpha_i|_{i=1}^{\infty}$, then from properties of conditional probability we have that

$$\begin{bmatrix} \mathbf{Pr}(X_t = \alpha_1) \\ \mathbf{Pr}(X_t = \alpha_2) \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{Pr}(X_t = \alpha_1 | X_{t-1} = \alpha_1) & \mathbf{Pr}(X_t = \alpha_1 | X_{t-1} = \alpha_2) & \cdots \\ \mathbf{Pr}(X_t = \alpha_2 | X_{t-1} = \alpha_1) & \mathbf{Pr}(X_t = \alpha_2 | X_{t-1} = \alpha_2) & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix} \times$$
$$\begin{bmatrix} \mathbf{Pr}(X_{t-1} = \alpha_1) \\ \mathbf{Pr}(X_{t-1} = \alpha_2) \\ \vdots \end{bmatrix}, \tag{2.17}$$

in which the matrix $\left[\mathbf{Pr}(X_t = \alpha_i | X_{t-1} = \alpha_j)|_{i,j=1}^{\infty}\right]$ is called the probability transition matrix. The relations revealed in Eqs. (2.16) and (2.17), together with results on matrix analysis, are useful in establishing properties of a Markov chain, such as reducibility, periodicity, transience and recurrence, mean recurrence time, and so on, especially when $x(t)$ only takes finitely many values.

When the conditional probability $\mathbf{Pr}(X_t = \alpha_i | X_{t-1} = \alpha_j)$ is independent of the index t for all $i, j = 1, 2, \cdots$, the probability transition matrix is usually written as $\left[p_{ij} |_{i,j=1}^{\infty} \right]$.

The next theorem gives a simple, yet very useful, method to generate a Markov chain, which can also be applied to the verification on whether or not a random process is a Markov chain.

Theorem 2.7. *Let $\mathcal{S}$ denote the set consisting of all the values of a random process $X = \{ x(t),\ t \in \mathcal{T} \}$ with countable index set. Represent its index set as $\mathcal{T} = \{1,\ 2,\ \cdots\}$. Assume that $f(\cdot,\ \cdot)$ is a function mapping $\mathcal{S} \times \mathcal{S}$ into $\mathcal{S}$. Suppose that this random process satisfies the following two conditions:*

- *for each integer $t \geq 1$, $X(t) = f(X(t-1),\ \xi(t))$, where $\xi(t)$ also belongs to the set $\mathcal{S}$;*
- *the random process $\{\xi(t),\ t \geq 1\}$ is a sequence of independent and identically distributed random variables or vectors. Moreover, $X(0)$ is also independent of this random sequence.*

Then, the random process $\{X(t),\ t \geq 0\}$ is a Markov chain. Moreover, for each $i, j = 1, 2, \cdots$, its one-step transition probability is

$$p_{ij} = \mu\{f(x(i),\ \xi) = x(j)\}.$$

Note that the equation $X(t) = f(X(t-1),\ \xi(t))$ takes completely the same form as that of the state transition equation in a state space model of a dynamic system. Applicability of Markov chains can be highly expected and in fact has been extensively investigated in dealing with signal processing problems and problems in dynamic system analysis and synthesis.

More generally, there is also a Markov chain of order m, in which m is a finite positive integer. In this random process, we have

$$\begin{aligned} & \mathbf{Pr}\,(X_t = x_t \mid X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2},\ \cdots,\ X_0 = x_0) \\ = \ & \mathbf{Pr}\,(X_t = x_t \mid X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2},\ \cdots,\ X_{t-m} = x_{t-m}) \end{aligned}$$

for each $t \geq m$, that is, the probability of a random event at the tth index associated with a Markov chain depends only on the past m random event. Clearly, when $m = 1$, this process reduces to the normal Markov chain.

Definition 2.15 (Semi-Markov Process). *For a random process with a finite or countable number of states, assume that it has a stepwise trajectory with jumps at indices. If the values of this random process at its jump indices form a Markov chain, then it is called a semi-Markov process.*

Semi-Markov processes provide a model for many processes widely used in fields like queueing theory, reliability theory, and so on.

2.7 Bibliographic Notes

Mathematics have been extensively adopted in system analysis and synthesis, in which almost all branches of mathematics have found their impacts. Examples include differential geometry in nonlinear system analysis and synthesis, functional analysis in robust control system designs, and so on. This chapter provides some mathematical concepts and results that are closely related to analysis and synthesis of a networked system, which are mostly from linear algebra, probability theory, and random processes. Detailed investigations on these concepts and results can be easily found in many standard textbooks and monographes in mathematics, for example, [1,2,7,22,23].

Basic concepts and results for combinatorial optimizations can be found, for example, in [18, 19]. Concerning with convex optimization, [24,25] provides an excellent introduction on its essential motivations and basic results.

References

[1] F.Z. Zhang, Matrix Theory: Basic Results and Techniques, Springer, New York, 1999.

[2] P. Lancaster, M.T. Tismenetsky, The Theory of Matrices: With Applications, Academic, New York, 1985.

[3] J.C. Doyle, Analysis of feedback systems with structured uncertainties, IEE Proceedings, Part D 133 (1982) 45–56.

[4] A. Packard, J.C. Doyle, The complex structured singular value, Automatica 29 (1993) 71–109.

[5] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[6] M.G. Safonov, Stability margins of diagonally perturbed multivariable feedback systems, IEEE Proceedings, Part D 129 (1982) 251–256.

[7] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[8] H. Abou-Kandil, G. Freiling, V. Ionescu, G. Jank, Matrix Riccati Equations in Control and System Theory, Birkhäuser Verlag, Basel, 2003.

[9] T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[10] T. Zhou, Coordinated one-step optimal distributed state prediction for a networked dynamical system, IEEE Transactions on Automatic Control 58 (2013) 2756–2771.

[11] P. Bougerol, Kalman filtering with random coefficients and contractions, SIAM Journal on Control and Optimization 31 (1993) 942–959.

[12] A. Censi, Kalman filtering with intermittent observations: convergence for semi-Markov chains and an intrinsic performance measure, IEEE Transactions on Automatic Control 56 (2011) 376–381.

[13] T. Zhou, Robust recursive state estimation with random measurement droppings, IEEE Transactions on Automatic Control 61 (2016) 156–171.

[14] T. Zhou, Asymptotic behavior of recursive state estimations with intermittent measurements, IEEE Transactions on Automatic Control 61 (2016) 400–415.

[15] H. Kimura, Chain-Scattering Approach to H^{∞}-Control, Birkhäuser, Boston, USA, 1997.

[16] R.E. Skelton, T. Iwasaki, K. Grigoriadis, A Unified Algebraic Approach to Linear Control Design, Taylor & Francis, London, UK, 1998.

[17] R.M. Redheffer, On a certain linear fractional transformation, Journal of Mathematics and Physics 39 (1960) 269–286.

[18] L. Lovasz, Submodular functions and convexity, in: A. Bachem, M. Grotschel, B. Korte (Eds.), Mathematical Programming: the State of the Art, Springer, Bonn, Germany, 1983.

[19] L.A. Wolsey, Integer Programming, John Wiley & Sons, USA, 1988.

[20] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, An analysis of approximation for maximizing submodular set functions I, Mathematical Programming 14 (1978) 265–294.

[21] T.H. Summers, F.L. Cortesi, J. Lygeros, On submodularity and controllability in complex dynamical networks, IEEE Transactions on Control of Network Systems 3 (2016) 91–101.

[22] Y.S. Chow, H. Teicher, Probability Theory: Independence, Interchangeability, Martingales, 3rd edition, Springer-Verlag, New York, USA, 1997.

[23] G.R. Grimmett, D.R. Stirzaker, Probability and Random Processes, Oxford University Press, USA, 2001.

[24] D.P. Bertsekas, Convex Optimization Theory, Athena Scientific, Boston, USA, 2009.

[25] J.B. Hiriart-Urruty, C. Lemarechal, Fundamentals of Convex Analysis, Springer, Berlin, Germany, 2001.

CHAPTER 3

Controllability and Observability of an LSS

3.1 Introduction

A prominent characteristic of modern control theory is the adoption of a state space model, which uses a set of first-order differential equations in the description of system input–output dynamics. The states of a plant might be either its actual physical variables or some of their linear combinations. Usually, not all of these states can be directly manipulated and/or measured. An essential issue is therefore whether it is possible to maneuver a plant state to a desirable value using feasible inputs and to estimate a plant state through measurements of accessible variables. The former is usually called controllability of the plant, whereas the latter its observability [1,2].

It is now extensively known that controllability and observability are closely related to other important characteristics of a plant. For example, controllability is required to locate eigenvalues of a linear system into an arbitrary desirable area through state feedbacks, whereas to guarantee the existence of a linear control law that makes the H_2 or H_∞ norm smaller than a prescribed value, the plant must be both controllable and observable. In addition to these, convergence of a state estimator is also closely related to the observability of a plant.

Controllability and observability are primarily formulated and investigated by Kalman [3] in his pursuit of system analysis and synthesis using a state space model. Through extensive pursuits of many researchers, various results have been obtained for the verification of the controllability and observability of a system. Originally, this problem is investigated for a linear dynamic system without any restrictions on plant inputs, states, and outputs. Afterward, these results are extended to more practical situations in which plant inputs and/or states and/or outputs are restricted to some prescribed sets and to nonlinear dynamic systems with the help of Lie brackets and Lie algebras. There are also studies on relations between controllability/observability of a system and its structure, which are usually called structural controllability and structural observability, respectively. Rather than system parameters, the corresponding results depend only on the positions of plant inputs and outputs, directed connections among plant states, way that the plant states are connected to its inputs, and the way that the plant outputs are connected to its states.

Although issues related to system controllability/observability have been investigated for more than half a century, it is still an active research topic. Especially, stimulated by the development of sensor technologies, communication technologies, computer technologies, and

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00003-8

so on, network technologies are widely recognized to be very helpful in the sense of providing more structural flexibilities, reducing greatly hardware investments, and so on in the construction of a control system. With the dream of making these advantages realistic, recently, there emerged extensive interests in the verification of controllability and observability of a networked system.

In this chapter, we aim at developing a numerically stable and computationally feasible method for checking these properties of a system constituted from a great number of subsystems. Verification of controllability and observability of a linear time-invariant (LTI) system is investigated without any constraints on its inputs, outputs, and states. At first, we summarize major associated results for a lumped LTI system. A model for spatially connected systems is introduced afterward, which is more convenient than the available descriptions and can represent the dynamics of a larger plant class. Necessary and sufficient conditions are given in Section 4.3 for the controllability and observability of a spatially connected system, which independently depends on parameters of each of its subsystems and its subsystem connection matrix. As a model is no longer required for the associated lumped system, this property is quite attractive in large-scale system analysis and synthesis.

3.2 Controllability and Observability of an LTI System

To investigate controllability and observability of a large-scale system, we first discuss these properties for a general finite-dimensional discrete LTI system. For these systems, its input–output relation can be described by the state space model of the following equations:

$$x(k+1) = Ax(k) + Bu(k), \tag{3.1a}$$

$$y(k) = Cx(k) + Du(k), \tag{3.1b}$$

where $x(k)$, $u(k)$, and $y(k)$ are respectively the plant state vector, input vector, and output vector. As in the remaining chapters of this book, their dimensions are assumed to be respectively n, q, and p.

Generally speaking, controllability is concerned with capabilities of a system in maneuvering its states through external inputs, whereas observability is concerned with capabilities of estimating its states using measurements of external outputs. Formally, they are defined as follows.

Definition 3.1. *The matrix pair* $(A,\ B)$ *or, equivalently, the discrete dynamical system described by Eqs. (3.1a) and (3.1b) is said to be controllable if for an arbitrary state vector pairs* $(x_0,\ x_1)$*, there exist a positive integer k and an input vector sequence* $u(0), u(1), \ldots, u(k-1)$ *such that under the drive of this sequence, the state vector of the system satisfies* $x(k) = x_1$ *when starting from* $x(0) = x_0$*. Otherwise, the system or the matrix pair is said to be uncontrollable.*

In some literature, controllability of a discrete system is also called reachability. To be consistent with continuous systems and avoid possible confusions, in this book, we use the term controllability.

Concerning a discrete LTI system, several criteria are available for the verification of its controllability using its system parameters. The following theorem summarizes some of the most widely used ones.

Theorem 3.1. *The discrete dynamical system of equations (3.1a) and (3.1b) is controllable if and only if one of the following conditions is satisfied:*

- *The controllability matrix $\mathcal{C} = [\,B \;\; AB \;\; A^2B \;\; \cdots \;\; A^{n-1}B\,]$ is of full row rank.*
- *The matrix $[\lambda I - A \;\; B]$ is of full row rank for every complex number λ.*

The first condition of the theorem is relatively obvious noting that

$$\begin{aligned} x(k) &= Ax(k-1) + Bu(k-1) \\ &= A[Ax(k-2) + Bu(k-2)] + Bu(k-1) \\ &= \cdots \\ &= A^k x(0) + [\,B \;\; AB \;\; A^2B \;\; \cdots \;\; A^{k-1}B\,] \begin{bmatrix} u(k-1) \\ u(k-2) \\ \vdots \\ u(0) \end{bmatrix}. \end{aligned} \tag{3.2}$$

On the other hand, note that according to matrix theories [4], for an arbitrary square matrix A of dimension $n \times n$, A^k can always be expressed as a linear combination of the matrices I_n, $A, \ldots,$ and A^{n-1} whenever k is a nonnegative integer. From these results it is not difficult to imagine that to guarantee the existence of a control consequence $u(0), u(1), \ldots, u(k-1)$ for an arbitrary positive integer $k \geq n$ and arbitrary n-dimensional column vectors x_0 and x_1, such that Eq. (3.2) is satisfied with $x(0) = x_0$ and $x(k) = x_1$, it is necessary that the controllability matrix $\mathcal{C}$ is of full row rank.

To understand the second criterion, we assume that the system is controllable, but the matrix $[\lambda I - A \;\; B]$ is not of full row rank at every complex number λ. Then, there exist at least one complex number λ^* and one nonzero n-dimensional row vector x^* such that

$$x^*[\lambda^* I - A \;\; B] = 0, \tag{3.3}$$

which can be equivalently rewritten as

$$\lambda^* x^* = x^* A, \quad x^* B = 0, \tag{3.4}$$

Hence

$$\begin{aligned} x^*\mathcal{C} &= x^*[\,B \;\; AB \;\; A^2B \;\; \cdots \;\; A^{k-1}B\,] \\ &= \lambda^*[\,0 \;\; x^*B \;\; x^*AB \;\; \cdots \;\; x^*A^{k-2}B\,] \\ &= \cdots \\ &= 0, \end{aligned} \tag{3.5}$$

which is in contradiction with the first criterion.

A mathematically rigorous proof of the theorem can be found in many textbooks on linear systems and/or linear estimations, such as [1,2].

Note that except at an eigenvalue of the matrix A, the matrix $\lambda I - A$ is always of full row rank and of full column rank. This implies that verification of the second criterion is only required at the eigenvalues of the matrix A.

Another fundamental concept in system analysis and synthesis is observability of a system. While controllability and observability of a system are completely different from an engineering point of view, it is now well known that they are mathematically dual.

Definition 3.2. *The discrete dynamical system described by Eqs. (3.1a) and (3.1b) or, equivalently, the matrix pair $(A,\ C)$, is said to be observable if there exits a positive integer k such that each initial system state vector $x(0)$ can be revealed from the input–output vector pairs $(u(s),\ y(s))_{s=0}^{k}$. Otherwise, this system or matrix pair is said to be unobservable.*

Note that when the input sequence $u(k)|_{k=0}^{\infty}$ is assumed to be known, it is clear from Eq. (3.1a) that all the uncertainties in the plant state vectors are caused by its initial values. This means that when the initial state vector of a plant is revealed from its output measurements, all of its state vectors can be revealed also from these measurements.

Similar to Theorem 3.1, we also have some algebraic criteria for the verification of observability of a system.

Theorem 3.2. *The discrete dynamical system of equations (3.1a) and (3.1b) is observable if and only if one of the following conditions is satisfied:*

- *The observability matrix $\mathcal{O} = \left[C^T \;\; A^TC^T \;\; (A^2)^TC^T \;\; \cdots \;\; (A^{n-1})^TC^T\right]^T$ is of full column rank.*
- *The matrix $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ is of full column rank for every complex number λ.*

Note that through substituting Eq. (3.2) into Eq. (3.1b) we have that, for an arbitrary positive integer s,

$$\begin{aligned} y(s) &= C\left[A^s x(0) + [\,B\ \ AB\ \ A^2B\ \ \cdots\ \ A^{s-1}B\,]\begin{bmatrix} u(s-1) \\ u(s-2) \\ \vdots \\ u(0) \end{bmatrix}\right] + Du(s) \\ &= CA^s x(0) + [\,D\ \ CB\ \ CAB\ \ \cdots\ \ CA^{s-1}B\,]\begin{bmatrix} u(s) \\ u(s-1) \\ u(s-2) \\ \vdots \\ u(0) \end{bmatrix}. \end{aligned} \tag{3.6}$$

Hence, the following equality is valid whenever k is a nonnegative integer:

$$\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^k \end{bmatrix} x(0) = \begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(k) \end{bmatrix} - \begin{bmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{k-1}B & CA^{k-2}B & CA^{k-3}B & \cdots & D \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ \vdots \\ u(k) \end{bmatrix}. \tag{3.7}$$

Once again, recall that when k is a nonnegative integer, A^k can be expressed as a linear combination of the matrices I_n, A, $\ldots$, and A^{n-1}. It is clear that to recover the system initial state vector $x(0)$ from its input–output vector pairs $(u(s),\ y(s))_{s=0}^{k}$ with $k \geq n$, it is necessary that the observability matrix $\mathcal{O}$ is of full column rank. The sufficiency of this condition is also clear, noting that when the matrix $\mathcal{O}$ is of full column rank, there exists one and only one vector $x(0)$ that satisfies Eq. (3.4) whenever $k \geq n$; that is, the system initial state vector is uniquely determined by its input–output pairs under this condition.

To understand the necessity and sufficiency of the second condition, assume that the system is observable, but the matrix $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ is not of full column rank at each complex number λ. Then, there exists a complex number λ^* and a nonzero n dimensional complex column vector x^*, such that

$$\begin{bmatrix} \lambda^* I - A \\ C \end{bmatrix} x^* = 0, \tag{3.8}$$

which is equivalent to

$$Ax^* = \lambda^* x^*, \quad Cx^* = 0 \tag{3.9}$$

Hence

$$\mathcal{O}x^* = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} x^* = \lambda^* \begin{bmatrix} 0 \\ Cx^* \\ CAx^* \\ \vdots \\ CA^{n-2}x^* \end{bmatrix} = \cdots = 0. \tag{3.10}$$

That is, the observability matrix $\mathcal{O}$ is not of full column rank. This is a contradiction with the first criterion.

Similarly to system controllability verifications, the verification of the second condition about system observability is only required to be performed at the eigenvalues of the matrix A.

Our discussions only provide some engineering insights about the criteria for system observability. For a more mathematically rigorous proof, a reader is recommended to refer to [1,2] among many excellent textbooks on optimal control and estimations.

From Theorems 3.1 and 3.2 it is clear that the verification of the controllability of the matrix pair $(A,\ B)$ is equivalent to the verification of the observability of the matrix pair $(A^T,\ B^T)$ and that the verification of the observability of the matrix pair $(A,\ C)$ is equivalent to the verification of the controllability of the matrix pair $(A^T,\ C^T)$. This relation is extensively known as the duality between controllability and observability of a system, which is sometimes very helpful in the analysis and synthesis of a complex system.

The second conditions in the aforementioned two theorems are usually called the PBH test, abbreviated from the names of three researchers V.M. Popov, V. Belevitch, and M. Hautus.

From these results the minimal number of inputs can be obtained for a plant with a prescribed state transition matrix such that an input matrix can be constructed that makes the plant controllable, as well as a parameterization for all input matrices that has a fixed number of inputs and can construct a controllable plant with this prescribed state transition matrix. Using dualities between controllability and observability of a linear time-invariant system, similar results can be obtained for the minimal number of outputs that leads to an observable plant and a parameterization for all plant output matrices [5]. These results can be extended to situations in which there are some constraints on the plant input/output matrices, which are frequently encountered in actual engineering systems, biological systems, and so on [6].

3.2.1 Minimal Number of Inputs/Outputs Guaranteeing Controllability/Observability

In the design of a large-scale system, a fundamental issue is to appropriately, if not optimally, locate actuators and sensors [7]. Toward a solution to this problem, we discuss in this subsection that how many actuators/sensors are required such that a controllable/observable system can be constructed.

In the following investigations, subscripts r and c are adopted to indicate whether a quantity, variable, and so on is associated with a real or complex eigenvalue of the STM A. This indication is important to clarify differences of real and complex eigenvalues in constructing either a real input matrix B or a real output matrix C such that the LTI system $\mathbf{\Sigma}$ is controllable or observable. Moreover, it is assumed that among the n eigenvalues of the STM A, there are k_r distinct real values and k_c distinct complex values, which are respectively denoted by $\lambda_{r,i}|_{i=1}^{k_r}$ and $\lambda_{c,i}|_{i=1}^{k_c}$. This assumption is introduced only for a concise presentation, and does not sacrifice any generality of the obtained results. Furthermore, for each $i = 1, 2, \ldots, k$ and all $* = r, c$, let $\left\{ x_{*,i}(j)|_{j=1}^{p_*(i)} \right\}$ denote a set of linearly independent vectors that span the null space of the matrix $\bar{\lambda}_{*,i} I_n - A^T$. In addition, define $p_{\max}$ as

$$p_{\max} = \max \left\{ \max_{1 \le i \le k_r} p_r(i), \; \max_{1 \le i \le k_c} p_c(i) \right\}. \tag{3.11}$$

Obviously, the quantity $p_{\max}$ is actually the maximum geometric multiplicity of the plant state transition matrix. In this subsection, it is proven that both the minimal number of inputs for controllability assurance and the minimal number of outputs for observability assurance are equal to p_{max}.

Note that the state transition matrix A is in general real. It can be proven that any of its left eigenvectors associated with an eigenvalue $\lambda_{*,i}$ is also a right eigenvector of the matrix A^T associated with its eigenvalue $\bar{\lambda}_{*,i}$. This means that the vectors $x_{*,i}(j)\Big|_{j=1}^{p_*(i)}$ are well defined, and $p_*(i) \ge 1$ equals the dimension of the $\lambda_{*,i}$ related eigenspace of the STM A, that is, for all $* = r, c$ and $i = 1, \cdots, k_*$, $p_*(i)$ equals the geometric multiplicity of its eigenvalue $\lambda_{*,i}$ [8]. Hence, $p_{\max}$ is the maximum of the geometric multiplicities of this matrix. (The geometric multiplicity of an eigenvalue is different from its algebraic multiplicity. The latter refers to its multiplicity as a root of the characteristic polynomial of the matrix.)

From Theorem 3.1 the following conclusion is established, giving a necessary and sufficient condition on an input matrix B for the corresponding system $\mathbf{\Sigma}$ to be controllable.

Lemma 3.1. *For all $* = r, c$ and $i = 1, 2, \cdots, k_*$, define the matrix*

$$X_{*,i} = \left[\, x_{*,i}(1) \;\; x_{*,i}(2) \;\; \cdots \;\; x_{*,i}(p_*(i)) \, \right]. \tag{3.12}$$

Then, a system with its state transitions described by Eq. (3.1a) is controllable if and only if the matrix $B^T X_{*,i}$ is always of full column rank (FCR).

Proof. Let the set $\mathcal{X}$ represent $\mathcal{C}$ when $* = c$ and $\mathcal{R}$ when $* = r$. From the definition of the matrix $X_{*,i}$ and properties of the eigenvectors of a matrix it is clear that, for all $* = r, c$ and $i = 1, 2, \cdots, k_*$, x is a left eigenvector of the matrix A associated with its eigenvalue $\lambda_{*,i}$ if and only if there exists a nonzero vector $\alpha \in \mathcal{X}^{p_*(i)}$ such that $x = X_{*,i}\alpha$. The proof can now be completed through a direct application of the PBH test in Theorem 3.1. This completes the proof. □

It is worth mentioning that a left eigenvector of a real-valued square matrix may still be complex valued [4]. On the other hand, for a practically realizable system, its input matrix is usually required to be real valued. This real–complex mixture asks careful investigations about the minimal inputs required to guarantee system controllability. Note also that, for a real valued matrix A, if a vector $x \in \mathcal{C}^n$ is one of its left eigenvectors associated with a complex eigenvalue λ, then the complex number $\bar{\lambda}$ is also its eigenvalue. Moreover, the vector $\bar{x}$ is a left eigenvector associated with this eigenvalue $\bar{\lambda}$ [4]. It can therefore be declared that if $k_c \neq 0$, then it is certainly an even number. Hence, it can be assumed, without any loss of generality, that $\lambda_{i+k_c/2} = \bar{\lambda}_i$ and $X_{i+k_c/2} = \bar{X}_i$, $i = 1, 2, \cdots, \frac{k_c}{2}$, provided that $k_c \geq 1$. This assumption is adopted throughout the rest of section.

To give a clear description on the minimal inputs required for controllability assurance, the state transition matrix A is first expressed through its Jordan canonical form. From the assumption that the eigenvalues $\lambda_{*,i}$ are different and each of them has $p_*(i)$ independent left eigenvectors, $i = 1, 2, \ldots, k_*$, $* = r$ or c, it can be declared from results on matrix analysis [4] that, associated with each $\lambda_{*,i}$, there are $p_*(i)$ Jordan blocks. Denote these Jordan blocks by $J_{*,i,j}$, and assume their dimensions being $m_{*,i,j}$ respectively, $j = 1, 2, \cdots, p_*(i)$. Moreover, for each Jordan block, there exists an $(n \times m_{*,i,j})$-dimensional FCR matrix $T_{*,i,j}$ satisfying

$$T_{*,i,j} J_{*,i,j} = A T_{*,i,j}, \tag{3.13}$$

and the vectors of the first columns of the matrices $T_{*,i,j}|_{j=1}^{p_r(i)}$ are linearly independent. Furthermore, the matrix $T_{*,i,j}$ is real when the associated eigenvalue is real and is generally complex when the associated eigenvalue is complex.

For a given scalar $\alpha_{*,i,j}$, define the matrix

$$\hat{B}_{*,i} = \left[\mathbf{diag}\left\{ \begin{bmatrix} 0_{m_{*,i,j}-1} \\ \alpha_{*,i,j} \end{bmatrix}_{j=1}^{p_*(i)} \right\},\ 0_{\sum_{j=1}^{p_*(i)} m_{*,i,j} \times (p_{max} - p_*(i))} \right],$$

in which $\alpha_{*,i,j}$ belongs to the set $\mathcal{R}$ when $* = r$ and to the set $\mathcal{C}$ when $* = c$. On the basis of these matrices, construct the input matrix

$$B = \sum_{i=1}^{k_r} \left[T_{r,i,j}\ T_{r,i,2}\ \cdots\ T_{r,i,p_r(i)}\right] \hat{B}_{r,i} + 2\sum_{i=1}^{k_c/2} \Re\left\{\left[T_{c,i,j}\ T_{c,i,2}\ \cdots\ T_{c,i,p_r(i)}\right] \hat{B}_{c,i}\right\}. \tag{3.14}$$

Using this particularly constructed input matrix, the following results are obtained, which reveal the minimal input number for System $\mathbf{\Sigma}$ being controllable.

Theorem 3.3. *There exists a matrix $B \in \mathcal{R}^{n\times q}$ such that a system with its state transitions described by Eq. (3.1a) is controllable if and only if q is not smaller than p_{max}.*

Proof. Assume that there exists a matrix $B \in \mathcal{R}^{n\times q}$ with $q < p_{\max}$ such that the matrix pair (A, B) is controllable. Let $\mathcal{I}$ denote the set consisting of the indices of the eigenvalues of the matrix A such that the maximum number of the associated linearly independent left eigenvectors achieves $p_{\max}$, that is,

$$\mathcal{I} = \mathcal{I}_r \bigcup \mathcal{I}_c, \tag{3.15}$$

where $\mathcal{I}_* = \{\, i \mid p_*(i) = p_{\max},\ 1 \le i \le k_* \}$ with $* = r, c$.

From the definitions it is clear that the sets $\mathcal{I}_r$ and $\mathcal{I}_c$ may be empty, but it is certain that they cannot be simultaneously empty.

Assume that the set $\mathcal{I}_c$ is not empty. For an arbitrary positive integer $i \in \mathcal{I}_c$, from the definition of the matrix $X_{c,i}$ given by Eq. (3.12) we have that the dimension of the matrix $B^T X_{c,i}$ is $q \times p_{\max}$, which cannot be FCR when $q < p_{\max}$. This contradicts with Lemma 3.1. Similar arguments apply when the set $\mathcal{I}_r$ is not empty. Hence, to guarantee the controllability of the matrix pair $(A,\ B)$, the matrix B must have at least $p_{\max}$ columns.

On the other hand, from Eq. (3.13) and the fact that the state transition matrix A is real, by the arrangements of its eigenvalues it is obvious that $\bar{T}_{*,i,j}\bar{J}_{*,i,j} = A\bar{T}_{*,i,j}$, which further implies that $\bar{T}_{c,i,j} J_{c,i+k_c/2,j} = A\bar{T}_{c,i,j}$ for all $i = 1, 2, \ldots, \frac{k_c}{2}$ and $j = 1, 2, \ldots, p_c(i)$. Define the matrix

$$T = [T_r\ \ T_c] \tag{3.16}$$

with $T_c = \left[\left[T_{c,i,j}\right]_{j=1,i=1}^{j=p_c(i),i=k_c/2}\ \left[\bar{T}_{c,i,j}\right]_{j=1,i=1}^{j=p_c(i),i=k_c/2}\right]$ and $T_r = \left[T_{r,i,j}\right]_{j=1,i=1}^{j=p_r(i),i=k_r}$. Since eigenvectors of a matrix associated with different eigenvalues are linearly independent [4] and the first columns of the matrices $T_{*,i,j}|_{j=1}^{p_r(i)}$ are linearly independent, it can be straightforwardly proven that the matrix T is invertible. Moreover,

$$T^{-1}AT = \mathbf{diag}\left\{ J_{*,i,j}\Big|_{j=1,i=1,*=r}^{j=p_*(i),i=k_*,*=c}\right\}, \tag{3.17}$$

$$T^{-1}B = \mathbf{col}\left\{ \hat{B}_{r,i}\Big|_{i=1}^{k_r},\ \hat{B}_{c,i}\Big|_{i=1}^{k_c/2},\ \bar{\hat{B}}_{c,i}\Big|_{i=1}^{k_c/2} \right\}. \tag{3.18}$$

Note that, for an arbitrary complex number $\lambda \in \mathcal{C}$,

$$[\lambda I_n - A,\ B] = T\left[\lambda I_n - T^{-1}AT,\ T^{-1}B\right]\mathbf{diag}\{T^{-1},\ I_{p_{max}}\}. \tag{3.19}$$

It is clear that the matrix $[\lambda I_n - A,\ B]$ is of FRR for each $\lambda \in \mathcal{C}$ if and only if every $\alpha_{*,i,j}$ in the definition of the matrix B is not equal to zero. Hence, through selecting an appropriate value for $\alpha_{*,i,j}$, we can construct a real-valued input matrix B with exactly $p_{\max}$ columns such that the associated matrix pair $(A,\ B)$ is controllable.

This completes the proof. □

Theorem 3.3 makes it clear that to construct a controllable system, the minimal number of inputs is exactly equal to the maximum geometric multiplicity of the state transition matrix. From its proof, especially from Eqs. (3.17)–(3.19), it can be understood that if the input number is smaller than $p_{\max}$, then, no matter how the input matrix B is selected, there certainly exist some states whose transition processes cannot be independently maneuvered by external inputs. This means that in the plant state space $\mathcal{R}^n$, there are some places that cannot be reached by the plant states. Hence, controllability of the system cannot be guaranteed.

Now, consider the problem of finding the minimal number of outputs such that the system is observable.

Note that all the system matrices A, B, C, and D are real valued. From Theorem 3.2 it is clear that the observability of the matrix pair $(A,\ C)$ is equivalent to the controllability of the matrix pair $(A^T,\ B^T)$, which is well known in systems and control theory as the duality between system observability and system controllability [1,2,9]. These mean that the results of Theorem 3.3 can be directly applied to finding the minimal number of outputs such that an output matrix C can be constructed that makes the matrix pair $(A,\ C)$ observable. The results are given in the following corollary. Their proof is omitted due to the straightforwardness.

Corollary 3.1. *There exists a matrix C such that the system $\mathbf{\Sigma}$ is observable if and only if the dimension of the output vector $y(k)$ is not smaller than the maximum geometric multiplicity of the STM A, that is, $p_{\max}$.*

3.2.2 A Parameterization of Desirable Input/Output Matrices

In the previous subsection, a necessary and sufficient condition is given for the existence of an input/output matrix B/C such that the associated system is controllable/observable. In many engineering problems, there usually exist some other requirements on a system input/output matrix. For example, some system states can hardly be straightforwardly affected by an external input or can hardly be directly measured by a sensor, constraints exist on input energy, restrictions are put on the number of the states that can be directly affected/measured, and so on [9–11]. To satisfy these requirements, it appears desirable to have a parameterization for all system input/output matrices.

In this subsection, we give a complete parameterization for all the input matrices B that have the minimal column number and construct a controllable system with the STM A. Using the duality between controllability and observability, these results can be directly applied to the parameterization of the system output matrix C that has a minimal row number and construct an observable system with the same state transition matrix A.

To get this parameterization, for all $* = r, c$ and $i = 1, 2, \ldots, k_*$, define the integer

$$m_{*,i} = \sum_{j=1}^{p_*(i)} m_{*,i,j}. \tag{3.20}$$

Moreover, for an arbitrary function of an integer variable j, define $\sum_{j=a}^{b} f(j) = 0$ whenever $b < a$. Then, we have the following results; their proof is deferred to the appendix of this chapter.

Theorem 3.4. *The matrix pair* $(A,\ B)$ *is controllable with a matrix* B *having the minimal number of columns if and only if there exist* $\hat{B}_{r,i} \in \mathcal{R}^{m_{r,i} \times p_{max}}$, $i = 1, 2, \ldots, k_r$, *and* $\hat{B}_{c,i} \in \mathcal{C}^{m_{c,i} \times p_{max}}$, $i = 1, 2, \ldots, \frac{k_c}{2}$, *such that* $\hat{B}_{c,i+k_c/2} = \bar{\hat{B}}_{c,i}$, $i = 1, 2, \ldots, \frac{k_c}{2}$, *and*

$$B = T\mathbf{col}\left\{\mathbf{col}\left\{\hat{B}_{r,i}|_{i=1}^{k_r}\right\},\ \mathbf{col}\left\{\hat{B}_{c,i}|_{i=1}^{k_c}\right\}\right\}, \tag{3.21}$$

and the matrix

$$\tilde{B}_{*,i} = \left[\hat{b}_{*,i}\left(\sum_{j=1}^{k-1} m_{*,i,j} + 1, s\right)\right]_{s=1,k=1}^{s=p_{max},k=p_*(i)} \tag{3.22}$$

is of FCR for each $i = 1, 2, \ldots, k_r$ *with* $* = r$ *and for each* $i = 1, 2, \ldots, \frac{k_c}{2}$ *with* $* = c$. *Here,* $\hat{b}_{*,i}(k, l)$ *is the kth row lth column element of the matrix* $\hat{B}_{*,i}$.

From the definition of the matrix B in Eq. (3.21) it is clear that it has just $p_{\max}$ columns. Hence, the results of Theorem 3.4 in fact give a complete parameterization for all input matrices B that have the minimal number of inputs and construct a controllable system with the STM A. On the other hand, from the proof of Theorem 3.4 it is obvious that its conclusions are in fact also valid when the number of system inputs is greater than $p_{\max}$. It can therefore be declared that when $p_{\max}$ is replaced by an arbitrary integer q satisfying $q \geq p_{\max}$, Eq. (3.21) also gives a complete parameterization for all input matrices B that have q columns and the associated matrix pair $(A,\ B)$ is controllable.

When the STM A is prescribed, similar results can be obtained for parameterizing all output matrices of an observable system with its output number not smaller than $p_{\max}$. These results can simply be obtained through the duality between system controllability and observability.

On the other hand, based on Theorem 3.4, the minimal numbers of inputs and outputs and a parameterization of the plant input/output matrix can be derived for a controllable/observable system, in which some rows/columns of its input/output matrix are prescribed to be zero. This is a well-encountered situation in actual applications, in which some plant states cannot be directly affected by an external signal or cannot be directly measured by a sensor. A detailed discussion is given in [6].

3.2.3 Some Nitpicking

Closely related to controllability/observability of a system, there is a concept called structural controllability/observability [10,12,13], which appears to be originally introduced in [14]. Loosely speaking, assume that in the state space model of a system, its state transition matrix, input matrix, output matrix, and a direct feed matrix depend on a parameter vector. Then, the system is said to be strongly structurally controllable if for all feasible values of this parameter vector, the system is controllable. On the other hand, if there exists at least one particular value for this parameter vector in its feasible sets such that the system is controllable when this parameter vector is fixed at that fixed value, then the system is said to be weakly structurally controllable, which is usually abbreviated as structurally controllable. Similar concepts have also been developed for system observability.

On the other hand, it has been proven that in various types of extensively adopted system models, either controllability or observability is a generic property of the system [15], which means that if the system is controllable/observable for one particular value of its parameter vector, then it is controllable/observable for almost all other values of its parameter vector in its definition set. The essential reason for this result is that if system matrices of a plant depend on a parameter vector, denote it by p, then the controllability matrix $\mathcal{C}$ of this system, defined in Theorem 3.1, is also a matrix-valued function of this parameter vector. To clarify this dependence, denote it by $\mathcal{C}(p)$. Then, according to Theorem 3.1, the values of this parameter vector that lead to a uncontrollable system must satisfy

$$\det\left(\mathcal{C}(p)\mathcal{C}^{T}(p)\right) = 0. \tag{3.23}$$

Assume that there is a particular value of the parameter vector p, say $p(0)$, that makes the associated system controllable. Then it is necessary that

$$\mathcal{C}(p(0))\mathcal{C}^T(p(0)) > 0. \tag{3.24}$$

This inequality implies that the determinant of the matrix $\mathcal{C}(p)\mathcal{C}^T(p)$ is not constantly equal to zero. Under such a situation, it can be proven that the parameter vectors that satisfy Eq. (3.23) usually construct a proper algebraic variety in the parameter space. The most obvious one is the case in which $\det(\mathcal{C}(p)\mathcal{C}^T(p))$ is a rational function of each element of the parameter vector, which is well encountered in actual problems. Note that a proper algebraic variety has a zero Lebesgue measure. It is not very hard to understand that if the system is controllable at a particular value of the parameter vector p, then at almost each value of this parameter vector, the system is also controllable.

This generic property makes it possible to connect system controllability/observability with its structure and to significantly reduce gaps between strong and weak structural controllabilities/observabilities. In addition, it also makes graph theory applicable to verifications of system controllability/observability. As a matter of fact, rather than numerical computations, almost all results about structural controllability/observability are expressed with terminologies of graph theory, such as path, cactus, and so on. Attractive characteristics of the associated results include their clear graphical illustrations, which is helpful in understanding information flows in a system and quite important in system analysis and synthesis.

Another issue in system analysis and synthesis is that even if it is controllable, the system may still not be very easy to be actually controlled. This requires investigations on appropriate measures on system controllability, which is usually related to energy needed to maneuver plant states [16,17].

3.3 A General Model for an LSS

Large-scale systems (LSS) are encountered frequently in engineering practice, biology systems, cyber-physical systems, and so on. Researches on LSS are extensively regarded to be started around 1970s. Various research articles and monographs have been published on the analysis and synthesis of an LSS [18,19]. In these studies, influences among subsystems are usually described through their state vectors. This description is general enough to represent the dynamics of a large class of interconnected systems. Sometimes, however, it may lose important structure information of these systems, which is helpful in reducing computational complexity in LSS analysis and design.

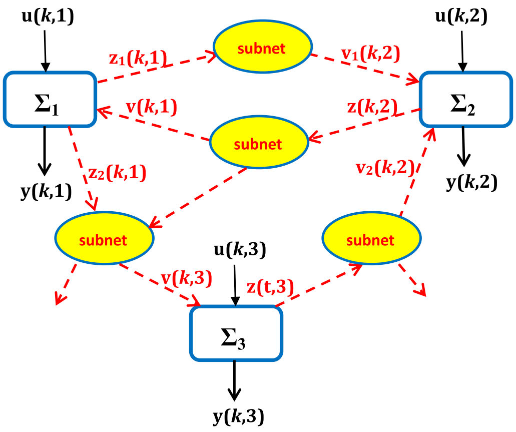

Figure 3.1: A schematic diagram for a networked system.

In this book, we adopt an LSS model in which subsystems affect each other through their outputs. More precisely, we consider the following networked system $\mathbf{\Sigma}$ constituted from N linear dynamic subsystems with the dynamics of its ith subsystem $\mathbf{\Sigma}_i$ described by

$$\begin{bmatrix} x(k+1,i) \\ z(k,i) \\ y(k,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k,i) & A_{\mathbf{xv}}(k,i) & B_{\mathbf{x}}(k,i) \\ A_{\mathbf{zx}}(k,i) & A_{\mathbf{zv}}(k,i) & B_{\mathbf{z}}(k,i) \\ C_{\mathbf{x}}(k,i) & C_{\mathbf{v}}(k,i) & D_{\mathbf{u}}(k,i) \end{bmatrix} \begin{bmatrix} x(k,i) \\ v(k,i) \\ u(k,i) \end{bmatrix} \tag{3.25}$$

and interactions among its subsystems described by

$$v(k) = \Phi(k) z(k), \tag{3.26}$$

where $z(k) = \mathbf{col}\left\{z(k,i)|_{i=1}^{N}\right\}$ and $v(k) = \mathbf{col}\left\{v(k,i)|_{i=1}^{N}\right\}$. Moreover, k and i stand respectively for the temporal variable and the index number of a subsystem, $x(k,i)$ represents the state vector of the ith subsystem $\mathbf{\Sigma}_i$ at the time instant k, $z(k,i)$ and $v(k,i)$ are its output vector to other subsystems and input vector from other subsystems, and $y(k,i)$ and $u(k,i)$ are its output and input vectors. To distinguish the output vector $z(k,i)$ and the input vector $v(k,i)$ from the output vector $y(k,i)$ and the input vector $u(k,i)$, the vectors $z(k,i)$ and $v(k,i)$ are called internal output/input vectors, whereas $y(k,i)$ and $u(k,i)$ are called external output/input vectors.

Note that in this model, except the linearity requirement, there are no other restrictions on the subsystems and their connection matrix. In fact, the dynamics of each subsystem may be completely different, and connections among different subsystems are arbitrary.

An illustrative diagram is given in Fig. 3.1 for the model, in which $z_{\star}(k,\sharp)$ is the $\star$-th subvector of the internal output vector $z(k,\sharp)$ of the plant's $\sharp$-th subsystem $\Sigma_{\sharp}$, whereas $v_{\star}(k,\sharp)$ is the $\star$-th subvector of its internal input vector $v(k,\sharp)$.

To obtain a model for the whole system, we define the matrices $A_{*\#}(k) = \mathbf{diag}\left\{A_{*\#}(k,i)|_{i=1}^{N}\right\}$, $B_{*}(k) = \mathbf{diag}\left\{B_{*}(k,i)|_{i=1}^{N}\right\}$, $C_{*}(k) = \mathbf{diag}\left\{C_{*}(k,i)|_{i=1}^{N}\right\}$, and $D_{\mathbf{u}}(k) = \mathbf{diag}\left\{D_{\mathbf{u}}(k,i)|_{i=1}^{N}\right\}$, in which $*, \# = \mathbf{x}$, $\mathbf{v}$, or $\mathbf{z}$. Moreover, we denote $\mathbf{col}\left\{u(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{x(k,i)|_{i=1}^{N}\right\}$, and $\mathbf{col}\left\{y(k,i)|_{i=1}^{N}\right\}$ by $u(k)$, $x(k)$, and $y(k)$, respectively. Then, straightforward algebraic manipulations show that when the dynamic system $\mathbf{\Sigma}$ is well-posed, which is equivalent to the regularity of the matrix $I - A_{\mathbf{zv}}(k)\Phi(k)$, its dynamics can be equivalently described by the following state space model:

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \left\{\begin{bmatrix} A_{\mathbf{xx}}(k) & B_{\mathbf{x}}(k) \\ C_{\mathbf{x}}(k) & D_{\mathbf{u}}(k) \end{bmatrix} \right.$$
$$\left. + \begin{bmatrix} A_{\mathbf{xv}}(k) \\ C_{\mathbf{v}}(k) \end{bmatrix} \Phi(k) \left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} \left[A_{\mathbf{zx}}(k) \;\; B_{\mathbf{z}}(k)\right] \right\} \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}. \tag{3.27}$$

Note that well-posedness is essential in system designs. In fact, a non-well-posed plant is usually hard to be controlled and/or unable to estimate [1,2]. Therefore, in this book, we assume that all involved systems are well-posed. This means that the inverse of the matrix $I - A_{\mathbf{zv}}(k)\Phi(k)$ always exists.

Define the matrices $A(k)$, $B(k)$, $C(k)$, and $D(k)$ as

$$\begin{aligned}
A(k) &= A_{\mathbf{xx}}(k) + A_{\mathbf{xv}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} A_{\mathbf{zx}}(k), \\
B(k) &= B_{\mathbf{x}}(k) + A_{\mathbf{xv}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} B_{\mathbf{z}}(k), \\
C(k) &= C_{\mathbf{x}}(k) + C_{\mathbf{v}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} A_{\mathbf{zx}}(k), \\
D(k) &= D_{\mathbf{u}}(k) + C_{\mathbf{v}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} B_{\mathbf{z}}(k).
\end{aligned}$$

Clearly, all these matrices are in general time varying. Moreover, on the basis of these matrices, the input–output relation of system $\mathbf{\Sigma}$ can be further concisely expressed as

$$x(k+1) = A(k)x(k) + B(k)u(k), \tag{3.28a}$$
$$y(k) = C(k)x(k) + D(k)u(k). \tag{3.28b}$$

This is the lumped state space model of the large-scale system described by Eqs. (3.25) and (3.26) and is very similar to that of Eq. (3.1), except that the associated system matrices are time dependent. These similarities make Theorems 3.1 and 3.2 applicable to the analysis of the controllability and observability for system Σ.

In this book, the matrix $\Phi(k)$ is usually called a subsystem connection matrix. Note that in addition to this subsystem connection matrix, the system matrices $A_{\mathbf{zx}}(k,i)$, $A_{\mathbf{zv}}(k,i)$, and $A_{\mathbf{xv}}(k,i)$ of Eq. (3.25) are also capable of describing influence strengths among plant subsystems. It can be assumed, without any loss of generality, that each row of this matrix has only one nonzero element equal to 1. When this condition is not satisfied by an original system model, the model can be modified to meet this requirement through augmenting the vector $z(k)$ and/or $v(k)$ by repeating some of its elements and adjusting associated matrices. On the other hand, most results in this book remain valid for a general subsystem connection matrix. However, we adopt this assumption for computational considerations and presentation simplicities.

From Eq. (3.28) it is clear that influences among the subsystems of the system described by Eqs. (3.25) and (3.26) can also be represented by their subsystem state vectors. However, it is worth mentioning that a large-scale system usually has a sparse structure [18,20–22], which means that most of the elements of the subsystem connection matrix $\Phi(k)$ are often equal to zero. On the other hand, although the matrices $A_{\mathbf{zv}}(k)$, $A_{\mathbf{xx}}(k)$, and so on are block diagonal from their definitions, the inverse of the matrix $I - A_{\mathbf{zv}}(k)\Phi(k)$ is usually dense even when the matrix $\Phi(k)$ is sparse. In addition, for an LSS, it is sometimes even not very easy to compute the inverse of this matrix. This means that from a computational point of view, the description of Eqs. (3.25) and (3.26) is more attractive than that of Eqs. (3.28a) and (3.28b).

3.4 Controllability and Observability for an LSS

In the following discussions, we assume that both the subsystem dynamics of Eq. (3.25) and the subsystem connection matrix of Eq. (3.26) are time invariant. To simplify mathematical expressions, in this section, the time index symbol k is omitted from all the system matrices. Moreover, $m_{\star j}$ is used to denote the dimension of the vector $\star(k, j)$ in which $j = 1,\ 2,\ \ldots,\ N$ and $\star = \mathbf{x}, \mathbf{v}, \mathbf{z}, \mathbf{u}$, or $\mathbf{y}$. For example, the dimension of the state vector $x(k, j)$ of the jth subsystem $\mathbf{\Sigma}_j$ is represented by $m_{\mathbf{x}j}$.

From the definitions of the matrices $A(k)$, $B(k)$, $C(k)$, and $D(k)$ it is clear that under the aforementioned assumptions, all these four matrices are also time invariant. This means that the results of Theorems 3.1 and 3.2 can in principle be straightforwardly applied to the verification of both the controllability and the observability of the system described by Eqs. (3.25) and (3.26). However, when the number of the subsystems is large, this direct application may be computationally prohibitive.

In this section, we investigate possibilities of checking controllability and observability of this system directly utilizing parameters of its subsystem and the subsystem connection matrix. It summarizes major results of [9]. For this purpose, define the integers $M_{\star i}$ and $M_{\star}$ as $M_{\star} = \sum_{j=1}^{N} m_{\star j}$ and $M_{\star j} = 0$ when $j = 1$ and $M_{\star i} = \sum_{j=1}^{i-1} m_{\star j}$ when $2 \leq i \leq N$. Here, once again, $\star = \mathbf{x}, \mathbf{v}, \mathbf{z}, \mathbf{u}$, or $\mathbf{y}$.

Theorem 3.5. *Assume that the dynamic system $\mathbf{\Sigma}$ is well-posed. Define the matrix-valued polynomial*

$$M(\lambda) = \begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} & -A_{\mathbf{xv}} \\ -C_{\mathbf{x}} & -C_{\mathbf{v}} \\ -\Phi A_{\mathbf{zx}} & I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}} \end{bmatrix}. \tag{3.29}$$

Then, this dynamic system is observable if and only if the matrix-valued polynomial $M(\lambda)$ is of full column rank (FCR) for each complex scalar λ.

Proof. Assume that the dynamic system $\mathbf{\Sigma}$ is observable when both its subsystem dynamics and subsystem connection matrix are time invariant. Then, it can be declared from Theorem 3.2 that, for every scalar complex number λ and every nonzero $M_{\mathbf{x}}$-dimensional complex vector y satisfying $(\lambda I_{M_{\mathbf{x}}} - A)y = 0$, we certainly have that $Cy \neq 0$. On the basis of the definitions of the matrices A and C, this is clearly equivalent to that there does not exist any $(\lambda,\ y)$ pair with a complex scalar λ and a nonzero complex vector y such that

$$\begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - \left[A_{\mathbf{xx}} + A_{\mathbf{xv}}(I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{zx}}\right] \\ C_{\mathbf{x}} + C_{\mathbf{v}}(I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{zx}} \end{bmatrix} y = 0. \tag{3.30}$$

Now, assume that there exists a complex scalar λ, denote it by λ_0, such that the matrix-valued polynomial $M(\lambda)$ is not of FCR. Then, there must exist a nonzero $(M_{\mathbf{x}} + M_{\mathbf{v}})$-dimensional complex vector z such that $M(\lambda_0)z = 0$. Partition this vector as $z = \mathbf{col}\{z_1,\ z_2\}$ with z_1 and z_2 respectively having dimensions of $M_{\mathbf{x}}$ and $M_{\mathbf{v}}$. Then, according to the definition of the matrix-valued polynomial $M(\lambda)$, we have

$$(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}})z_1 - A_{\mathbf{xv}}z_2 = 0, \tag{3.31a}$$

$$C_{\mathbf{x}}z_1 + C_{\mathbf{v}}z_2 = 0, \tag{3.31b}$$

$$-\Phi A_{\mathbf{zx}}z_1 + (I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})z_2 = 0. \tag{3.31c}$$

From the subsystem connections of system $\mathbf{\Sigma}$ and its well-posedness we have that $I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}}$ is invertible [2,23]. It can therefore be claimed from Eq. (3.31c) that $z_2 = (I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{zx}}z_1$. Note that by the adopted assumptions we have that $\mathbf{col}\{z_1,\ z_2\} \neq 0$. The aforementioned relation between the vectors z_1 and z_2 further makes it clear that $z_1 \neq 0$. Substituting this expression for the vector z_2 into Eqs. (3.31a) and (3.31b), direct algebraic manipulations show that

$$\left\{\lambda_0 I_{M_{\mathbf{x}}} - \left[A_{\mathbf{xx}} + A_{\mathbf{xv}}(I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{ST}}\right]\right\} z_1 = 0, \tag{3.32a}$$

$$\left[C_{\mathbf{x}} + C_{\mathbf{v}}(I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{zx}}\right] z_1 = 0. \tag{3.32b}$$

These two equations and Eq. (3.30) clearly contradict each other, which means that the existence of the aforementioned λ_0 and z is not possible, and therefore the matrix-valued polynomial $M(\lambda)$ is always of FCR.

On the contrary, assume that the matrix-valued polynomial $M(\lambda)$ is always of FCR but System $\mathbf{\Sigma}$ is not observable. Then, according to Theorem 3.2 and the definitions of the matrices A and C, there exist at least one scalar complex number λ and one nonzero $M_{\mathbf{x}}$-dimensional complex vector y such that Eq. (3.30) is satisfied. Define the vector $\psi = (I_{M_{\mathbf{v}}} - \Phi A_{\mathbf{zv}})^{-1}\Phi A_{\mathbf{zx}}y$. Then, it can be straightforwardly proved from Eq. (3.29) and the definition of the vector ψ that for this complex scalar λ, $M(\lambda)\mathbf{col}\{y,\ \psi\} = 0$. As the vector y is not equal to zero, it is clear that $\mathbf{col}\{y,\ \psi\}$ is also not a zero vector, which means that this is a contradiction to the assumption on $M(\lambda)$. Hence, the dynamic system $\mathbf{\Sigma}$ must be observable. This completes the proof. □

From Lemma 11.3 given in the following Chapter 11, it is clear that the networked system described by Eqs. (3.25) and (3.26) is *observable*, only if the following dynamic system

$$x(k+1) = A_{\mathbf{xx}}x(k) + A_{\mathbf{xv}}u(k)$$
$$y(k) = \begin{bmatrix} C_{\mathbf{x}} \\ \Phi A_{\mathbf{zx}} \end{bmatrix} x(k) + \begin{bmatrix} C_{\mathbf{v}} \\ \Phi A_{\mathbf{zv}} - I_{M_{\mathbf{v}}} \end{bmatrix} u(k)$$

is *strongly observable*.

An attractive characteristic of the results of Theorem 3.5 is that it does not require the computation of the inverse of the matrix $I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}$, which may lead to a more numerically stable verification procedure. It is also worth noting that except the subsystem connection matrix Φ, all the other matrices are block diagonal. This property has been proven very helpful in developing a computationally efficient algorithm for verifying the observability of the dynamic system $\mathbf{\Sigma}$ [9].

Remark 3.1. *When each subsystem is completely isolated from other subsystems in the networked system described by Eqs. (3.25) and (3.26), the subsystem interconnection matrix Φ becomes a zero matrix. In this case, the matrix-valued polynomial $M(\lambda)$ defined in Eq. (3.29) reduces to*

$$M(\lambda) = \begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} & -A_{\mathbf{xv}} \\ -C_{\mathbf{x}} & -C_{\mathbf{v}} \\ 0 & I_{M_{\mathbf{v}}} \end{bmatrix}.$$

Obviously, at an arbitrary value of the complex variable λ, this matrix is of full column rank if and only if the matrix $\begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} \\ -C_{\mathbf{x}} \end{bmatrix}$ is of full column rank at this value. Note that from their definitions it is clear that both matrices $A_{\mathbf{xx}}$ and $A_{\mathbf{xv}}$ are block diagonal, and the number of the columns in their ith diagonal blocks are equal to each other for every $i = 1, 2, \cdots, N$. It can be straightforwardly proven that the matrix $\begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} \\ -C_{\mathbf{x}} \end{bmatrix}$ is of full column rank if *and only if for each $i \in \{1, 2, \cdots, N\}$, the matrix* $\begin{bmatrix} \lambda I_{m_{\mathbf{xi}}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{x}}(i) \end{bmatrix}$ *is of full column rank. From Theorem 3.2 it follows that the latter is also a necessary and sufficient condition for the observability of each subsystem in the networked system with its own external outputs.*

These observations imply that Theorem 3.5 also makes it clear that when all the subsystems are isolated from each other, then to guarantee the observability of the whole system, it is necessary and sufficient that each subsystem is observable with its own external outputs. This is clear from an engineering aspect, since under such a situation, information on the state vector of a subsystem cannot be transferred to the external outputs of any other subsystem, so that estimations on its state vector can only be performed on its own external subsystem.

On the other hand, when subsystem interactions exist in the networked system, the subsystem interconnection matrix Φ will no longer be a zero matrix. In this case, even if there exist a complex number λ and $i \in \{1, 2, \cdots, N\}$ such that the matrix-valued polynomial $\begin{bmatrix} \lambda I_{m_{\mathbf{xi}}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{x}}(i) \end{bmatrix}$ is not of full column rank, which further leads to that at this complex number λ, the matrix-valued polynomial $\begin{bmatrix} \lambda I_{m_{\mathbf{x}}} - A_{\mathbf{xx}} \\ C_{\mathbf{x}} \end{bmatrix}$ is not of full column rank, it is still possible that the matrix-valued polynomial

$$\begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} \\ -C_{\mathbf{x}} \\ -\Phi A_{\mathbf{zx}} \end{bmatrix}$$

and therefore the matrix-valued polynomial $M(\lambda)$ is of full column rank at this complex number λ. In other words, even if there is a subsystem in the networked system that is not observable with its own external outputs, appropriate subsystem connections are still able to lead to an observable networked system. This is due to that when a subsystem is connected to some other subsystems, not only its own external output vector, but also the external output vector of other subsystems, contain information about its state vector. Intuitively, observability of a networked system depends on appropriate information transfers among its subsystems. This is made mathematically clear in the next subsection, which claims that observability of a networked system is closely related to transmission zeros of its subsystems.

3.4.1 Subsystem Transmission Zeros and Observability of an LSS

Although Theorem 3.5 provides a necessary and sufficient condition on the observability of a networked system, it cannot be directly applied to actual verifications of system observability, noting that it requires to check every complex number λ that is computationally intensive and usually impossible. On the other hand, discussions in the previous subsection make it intuitively clear that to construct an observable networked system, appropriate information transfer among its subsystems is required. To develop a computationally feasible criterion for the observability of the dynamic system $\mathbf{\Sigma}$ and to mathematically clarify this engineering intuitions, the following results are required for the transmission zeros of a dynamic system [2].

Lemma 3.2. *Assume that the transfer function matrix (TFM) $G(\lambda)$ is proper and of full column normal rank (FCNR). Then, a complex number λ_0 is a transmission zero of this TFM if and only if there exists a nonzero complex vector z_0 such that $G(\lambda_0)z_0 = 0$.*

A transmission zero is an important concept in system analysis and synthesis, which reflects the existence of a signal whose influence on a system cannot be measured by its outputs [2]. Clearly, if a signal is completely blocked out in the outputs of a system, then variations of its states due to the stimulus of that signal will not cause any change in the outputs of the system, which possibly make the system unobservable. This means that there may exist some close relations between the observability of a system and its transmission zeros. For a networked system, this relation is first revealed in [9].

Using this result, we can derive a condition for the observability of the dynamic system $\mathbf{\Sigma}$, which is equivalent to that of Theorem 3.5.

Theorem 3.6. *Define a TFM $G(\lambda)$ as*

$$G(\lambda) = \begin{bmatrix} C_{\mathbf{v}} \\ \Phi A_{\mathbf{zv}} - I_{M_{\mathbf{z}}} \end{bmatrix} + \begin{bmatrix} C_{\mathbf{x}} \\ \Phi A_{\mathbf{zx}} \end{bmatrix} \left(\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^{-1} A_{\mathbf{xv}}. \tag{3.33}$$

Then, the matrix-valued polynomial $M(\lambda)$ is of FCR at every complex number λ only if $G(\lambda)$ is of FCNR. Moreover, when $G(\lambda)$ is of FCNR, there exists a complex number λ at which $M(\lambda)$ does not have an FCR if and only if $G(\lambda)$ has a transmission zero.

Proof. Assume that the TFM $G(\lambda)$ is not of FCNR. This is equivalent to that at every value of the variable λ, the matrix $G(\lambda)$ is column rank deficient. Hence, for any particular λ_0, there must exist a nonzero vector z_0 such that $G(\lambda_0)z_0 = 0$. Denote the vector $y_0 = \left(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^{-1} A_{\mathbf{xv}} z_0$ by y_0. Then, it can be directly proved using the definition of the TFM $G(\lambda)$ that $M(\lambda_0)\mathbf{col}\{z_0,\ y_0\} = 0$. As z_0 is not equal to zero, it is clear that $\mathbf{col}\{z_0,\ y_0\}$ also is not a zero vector. It can therefore be declared that under the assumption that the TFM $G(\lambda)$ is not of FCNR, we certainly have that there exists at least one complex λ_0 such that the matrix $M(\lambda_0)$ is not of FCR. Hence, to guarantee that $M(\lambda)$ is always of FCR, it is necessary that $G(\lambda)$ is of FCNR.

Now, assume that the TFM $G(\lambda)$ is of FCNR but has some transmission zeros. Let λ_0 denote one of them. According to Lemma 3.2, there exists a nonzero vector z_0 such that

$G(\lambda_0)z_0 = 0$. Using completely the same arguments as those for the transfer function matrix $G(\lambda)$ being not of FCNR, it can be shown that the matrix $M(\lambda_0)$ is not of FCR.

On the contrary, assume that the matrix-valued polynomial $M(\lambda)$ is column rank deficient at some particular values of the variable λ, and let λ_0 be one of them. Then, there exists at least one nonzero vector z such that $M(\lambda_0)z = 0$. Partition this vector as $z = \mathbf{col}\{z_1,\ z_2\}$ with the vector z_1 having a dimension of $M_{\mathbf{x}}$. Then we have that $\left(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right) z_1 - A_{\mathbf{xv}} z_2 = 0$. As $A_{\mathbf{xx}}$ is square according to its definition, it can be straightforwardly claimed that the matrix-valued polynomial $\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}$ is of full normal rank. Hence, z_1 can always be formally expressed as[1]

$$z_1 = \left(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^{-1} A_{\mathbf{xv}} z_2. \tag{3.34}$$

Substituting this expression for z_1 back into $M(\lambda_0)z = 0$, it can be shown through direct algebraic manipulations that

$$\left[C_{\mathbf{v}} + C_{\mathbf{x}} \left(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^{-1} A_{\mathbf{xv}}\right] z_2 = 0, \tag{3.35a}$$

$$\left[\Phi A_{\mathbf{zv}} - I_{M_{\mathbf{z}}} + \Phi A_{\mathbf{zx}} \left(\lambda_0 I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^{-1} A_{\mathbf{xv}}\right] z_2 = 0. \tag{3.35b}$$

By the definition of the TFM $G(\lambda)$ simultaneous satisfaction of these two equalities is clearly equivalent to $G(\lambda_0)z_2 = 0$. As the vector z is not equal to zero, Eq. (3.34) and $z = \mathbf{col}\{z_1,\ z_2\}$ imply that z_2 also is not a zero vector. It can therefore be declared from Lemma 3.2 that the TFM $G(\lambda)$ is of FCNR, and λ_0 is one of its transmission zeros. This completes the proof. □

From the definition of the TFM $G(\lambda)$ it is clear that, except the subsystem connection matrix Φ, all the other involved matrices have a consistent block diagonal structure. This is an attractive property. Especially, the block diagonal structure of the matrix $A_{\mathbf{xx}}$ makes the inverse of the matrix $\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}$ also block diagonal, which is significantly different from that of the matrix $I_{M_{\mathbf{z}}} - \Phi A_{\mathbf{zv}}$. In the latter case, although the matrix $A_{\mathbf{zv}}$ is block diagonal, the inverse of the matrix $I_{M_{\mathbf{z}}} - \Phi A_{\mathbf{zv}}$ is usually dense and does not keep that structure property. The block diagonal structure is violated by the subsystem connection matrix, which is in general not block diagonal.

This particular structure property of the TFM $G(\lambda)$ enables developments of computationally efficient procedures for verifying observability of the dynamic system $\mathbf{\Sigma}$. For this purpose, define TFMs $G^{[1]}(\lambda)$, $G^{[2]}(\lambda)$, $G_i^{[1]}(\lambda)$ and $G_i^{[2]}(\lambda)$, $i = 1, \cdots, N$, respectively as

$$G^{[1]}(\lambda) = \mathbf{diag}\{G_i^{[1]}(\lambda)|_{i=1}^{N}\}, \quad G^{[2]}(\lambda) = \mathbf{diag}\{G_i^{[2]}(\lambda)|_{i=1}^{N}\},$$

[1] If this matrix is not invertible at some particular λ_0, which is in fact equal to an eigenvalue of the matrix $A_{\mathbf{xx}}$, then $A_{\mathbf{xv}} z_2$ must belong to the space spanned by its vectors. This guarantees the validness of the expression for z_1 with the matrix inverse interpreted as a generalized inverse [4].

$$G_i^{[1]}(\lambda) = C_{\mathbf{v}}(i) + C_{\mathbf{x}}(i)[\lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i)]^{-1} A_{\mathbf{xv}}(i),$$
$$G_i^{[2]}(\lambda) = A_{\mathbf{zv}}(i) + A_{\mathbf{zx}}(i)[\lambda I_{m_{\mathbf{x}}i} - A_{\mathbf{xx}}(i)]^{-1} A_{\mathbf{xv}}(i).$$

From Lemma 3.2 and the block diagonal structure of the TFM $G^{[1]}(\lambda)$ it can be directly proved that every transmission zero of the TFM $G_i^{[1]}(\lambda)$ with $i \in \{1,\ 2,\ \cdots,\ N\}$ is also a transmission zero of $G^{[1]}(\lambda)$. Note that the existence of differences among different TFMs does not mean that all their transmission zeros are distinctive. Assume that the TFMs $G^{[1]}(\lambda)$ and $G_i^{[1]}(\lambda)$ have respectively m and m_i distinctive transmission zeros. These arguments imply that $\max_{1 \le i \le N} m_i \le m \le \sum_{i=1}^{N} m_i$. In addition, the block diagonal structure of the TFM $G^{[1]}(\lambda)$ also implies that, for each its transmission zero, there exists at least one TFM $G_i^{[1]}(\lambda)$ with i belonging to $\{1,\ 2,\ \cdots,\ N\}$ such that it is also a transmission zero of $G_i^{[1]}(\lambda)$.

3.4.2 Observability Verification

On the basis of the results obtained in the previous subsections, we can develop a procedure for the observability verification of the networked system defined by Eqs. (3.25) and (3.25). For this purpose, denote the transmission zeros of the TFM $G^{[1]}(\lambda)$ by $\lambda_0^{[k]}\Big|_{k=1}^{m}$. Moreover, assume that its kth transmission zero $\lambda_0^{[k]}$ is shared by the TFMs $G_{k(s)}^{[1]}(\lambda)$, $s = 1, \cdots, s^{[k]}$. Without any loss of generality, we can assume that $1 \le k(1) < k(2) < \cdots < k(s^{[k]}) \le N$. Let $\mathbf{Y}_s^{[k]}$ denote the null space of the matrix $G_{k(s)}^{[1]}(\lambda_0^{[k]})$. Construct the vector set

$$\mathbf{Y}^{[k]} = \left\{ y \left| \begin{array}{l} y = \mathbf{col}\left\{ \left(0_{m_{\mathbf{S}(k(i)+1)}},\ \cdots,\ 0_{m_{\mathbf{S}(k(i+1)-1)}},\ y_{i+1,0}^{[k]}\right)\Big|_{i=0}^{s^{[k]}-1},\ 0_{m_{\mathbf{S}(k(s^{[k]})+1)}},\ \cdots,\ 0_{m_{\mathbf{S}N}} \right\} \\ \qquad y_{i,0}^{[k]} \in \mathbf{Y}_i^{[k]},\ i = 1, 2, \cdots, s^{[k]};\ \ y \ne 0 \end{array} \right. \right\}, \tag{3.36}$$

in which $k(0) = 0$. From this construction it can be directly proved that this set belongs to the null space of the matrix $G^{[1]}(\lambda_0^{[k]})$, that is, for every $y^{[k]} \in \mathbf{Y}^{[k]}$, we have

$$G^{[1]}(\lambda_0^{[k]}) y^{[k]} = \mathbf{diag}\left\{ G_i^{[1]}(\lambda_0^{[k]})\Big|_{i=1}^{N} \right\} y^{[k]} = 0. \tag{3.37}$$

From these results a necessary and sufficient condition is derived for the observability of system $\mathbf{\Sigma}$ with time-invariant subsystem dynamics and subsystem connection matrix, which can lead to a computationally efficient verification procedure.

Theorem 3.7. *Assume that all TFMs $G_i^{[1]}(\lambda)|_{i=1}^{N}$ are of FCNR. Then, the dynamic system $\mathbf{\Sigma}$ is observable if and only if for every $1 \le k \le m$ and every $y^{[k]} \in \mathbf{Y}^{[k]}$, $\Phi G^{[2]}(\lambda_0^{[k]}) y^{[k]} \ne y^{[k]}$.*

Proof. Note that according to the definitions of the matrices $A_{\star\#}$ and $C_{\star}$ with $\star,\ \# = \mathbf{x},\ \mathbf{z},\ \mathbf{v}$, all they are block diagonal with consistent dimensions. Moreover, their block diagonal matrices are constituted from system matrices of the plant subsystems. On the basis of these observations and the definitions of the TFMs $G(\lambda)$, $G^{[1]}(\lambda)$, and $G^{[2]}(\lambda)$, straightforward algebraic manipulations show that

$$G(\lambda) = \begin{bmatrix} G^{[1]}(\lambda) \\ \Phi G^{[2]}(\lambda) - I_{M_{\mathbf{z}}} \end{bmatrix}. \tag{3.38}$$

On the other hand, from the block diagonal structure of the TFM $G^{[1]}(\lambda)$ it is obvious that the requirement that $G^{[1]}(\lambda)$ is of FCNR is equivalent to the requirement that each TFM $G_i^{[1]}(\lambda)|_{i=1}^N$ is of FCNR. This means that when the adopted assumption is satisfied, the TFM $G(\lambda)$ is certainly of FCNR.

Assume that the dynamic system $\mathbf{\Sigma}$ is observable. If there exist an integer k and a complex vector $y^{[k]}$, with k belonging to the set $\{1, 2, \cdots, m\}$ and $y^{[k]}$ belonging to the set $\mathbf{Y}^{[k]}$, such that $\Phi G^{[2]}(\lambda_0^{[k]})y^{[k]} = y^{[k]}$, then from the definition of the set $\mathbf{Y}^{[k]}$ it can be straightforwardly declared that $y^{[k]} \neq 0$ and $G^{[1]}(\lambda_0^{[k]})y^{[k]} = 0$. We therefore have from Eq. (3.36) that

$$G(\lambda_0^{[k]})y^{[k]} = \begin{bmatrix} G^{[1]}(\lambda_0^{[k]})y^{[k]} \\ \Phi G^{[2]}(\lambda_0^{[k]})y^{[k]} - y^{[k]} \end{bmatrix} = 0. \tag{3.39}$$

As $y^{[k]} \neq 0$, it can be claimed from Lemma 3.2 that $\lambda_0^{[k]}$ is one of the transmission zeros of the TFM $G(\lambda)$. Hence, on the basis of Theorems 3.5 and 3.6, we can declare that system $\mathbf{\Sigma}$ is not observable, which is a contradiction to the observability assumption. Therefore, the assumption about the existence of k and $y^{[k]}$ cannot be satisfied simultaneously.

On the contrary, assume that the system $\mathbf{\Sigma}$ is unobservable. Then, according to Theorems 3.1 and 3.5 and Lemma 3.2, there exist at least one complex number λ_0 and one nonzero complex vector y_0 such that

$$G(\lambda_0)y_0 = 0. \tag{3.40}$$

From Eq. (3.36) and the definition of the TFM $G^{[1]}(\lambda)$ this equality means that

$$\mathbf{diag}\{G_i^{[1]}(\lambda_0)|_{i=1}^N\}y_0 = 0.$$

Partition the vector y_0 as $y_0 = \mathbf{col}\{y_{0i}|_{i=1}^N\}$ with the column vector y_{0i} having a dimension of $m_{\mathbf{v}i}$. Then there exists at least one integer i simultaneously satisfying $1 \leq i \leq N$ and $y_{0i} \neq 0$, denote it by i_0, such that $G_{i_0}^{[1]}(\lambda_0)y_{0i_0} = 0$. Therefore, λ_0 is also a transmission zero of the TFM $G_{i_0}^{[1]}(\lambda)$ and $y_{0i_0} \in \mathbf{Y}_{i_0}^{[*]}$, in which $*$ is an integer belonging to the set

$\{1,\ 2,\ \cdots,\ m\}$. Hence, $y_0 \in \mathbf{Y}^{[*]}$. On the other hand, from Eqs. (3.36) and (3.37) we have that $\left[\Phi G^{[2]}(\lambda_0) - I_{M_{\mathbf{z}}}\right] y_0 = 0$. Therefore, $\Phi G^{[2]}(\lambda_0^{[*]}) y_0 = y_0$, that is, if the dynamic system $\mathbf{\Sigma}$ is not observable, then there certainly exist an integer k and a nonzero complex vector y simultaneously satisfying $1 \le k \le m$, $y \in \mathbf{Y}^{[k]}$ and $\Phi G^{[2]}(\lambda_0^{[k]}) y = y$. This completes the proof. □

Note that the TFM $G^{[2]}(\lambda)$ is block diagonal and all the elements of the subsystem connection matrix Φ belong to the set $\{0,\ 1\}$. Note also that in each row of the matrix Φ, there is only one nonzero element. This means that for a prescribed vector y and a prescribed number λ, the computation of $\Phi G^{[2]}(\lambda) y$ is in fact reduced to the exchange of the elements of a column vector and to multiplications between a matrix and a vector with small sizes that depend only on the dimensions of the subsystems $\mathbf{\Sigma}_i|_{i=1}^{N}$. This is a quite attractive property in the analysis and synthesis of a large-scale networked system, as it means that the computation complexity for verifying its observability increases only linearly with the increment of the number of its subsystems.

A numerical simulation comparison is provided in [9] between the computation time of the PBH test and that of a prototype algorithm based on Theorem 3.7. That comparison shows that when computational complexity is concerned, this theorem is significantly superior to the PBH test when the system $\mathbf{\Sigma}$ has a large number of subsystems.

3.4.3 A Condition for Controllability and Its Verification

Owing to the duality between the controllability of a LTI system and its observability, similar results can be obtained for controllability verification of a networked system.

More precisely, define TFMs $\bar{G}^{[1]}(\lambda)$ and $\bar{G}^{[2]}(\lambda)$ respectively as $\bar{G}^{[1]}(\lambda) = \mathbf{diag}\{\bar{G}_i^{[1]}(\lambda)|_{i=1}^{N}\}$ and $\bar{G}^{[2]}(\lambda) = \mathbf{diag}\{\bar{G}_i^{[2]}(\lambda)|_{i=1}^{N}\}$, where $\bar{G}_i^{[1]}(\lambda) = B_{\mathbf{z}}^T(i) + B_{\mathbf{x}}^T(i)\left[\lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}^T(i)\right]^{-1} A_{\mathbf{zx}}^T(i)$ and $\bar{G}_i^{[2]}(\lambda) = (G_i^{[2]}(\lambda))^T$. Let $\bar{\lambda}_0^{[1]}, \bar{\lambda}_0^{[2]}, \cdots, \bar{\lambda}_0^{[\bar{m}]}$ denote all the distinctive transmission zeros of the TFM $\bar{G}^{[1]}(\lambda)$, and let $\bar{G}_{\bar{k}(s)}^{[1]}(\lambda)|_{s=1}^{\bar{s}^{[k]}}$ be the TFMs that have $\bar{\lambda}_0^{[k]}$ as its transmission zero, $k = 1, 2, \cdots, \bar{m}$.

As in the observability verification, we can also assume without any loss of generality that $\bar{k}(1) < \bar{k}(2) < \cdots < \bar{k}(\bar{s}^{[k]})$. Define the set

$$\bar{\mathbf{Y}}_s^{[k]} = \left\{ y \;\middle|\; \bar{G}_{\bar{k}(s)}^{[1]}(\bar{\lambda}_0^{[k]}) y = 0 \right\}$$

and the integer $\bar{k}(0) = 0$. Using these vector sets, construct the other vector set

$$\bar{\mathbf{Y}}^{[k]} = \left\{ y \;\middle|\; \begin{array}{l} y = \mathbf{col}\left\{ \left(0_{m_{\mathbf{z}(\bar{k}(i)+1)}},\ \cdots,\ 0_{m_{\mathbf{z}(\bar{k}(i+1)-1)}},\ \bar{y}_{i+1,0}^{[k]}\right)\Big|_{i=0}^{\bar{s}^{[k]}-1},\ 0_{m_{\mathbf{z}(\bar{k}(\bar{s}^{[k]})+1)}},\ \cdots,\ 0_{m_{\mathbf{z}N}} \right\} \\ \bar{y}_{i,0}^{[k]} \in \bar{\mathbf{Y}}_i^{[k]},\ i = 1, 2, \cdots, \bar{s}^{[k]};\ \ y \neq 0 \end{array} \right\}. \tag{3.41}$$

Then we have the following results on the controllability of the dynamic system $\mathbf{\Sigma}$, which are very similar to those of Theorem 3.7.

Corollary 3.2. *Assume that all the TFMs $\bar{G}_i^{[1]}(\lambda)|_{i=1}^N$ are of FCNR. The dynamic system $\mathbf{\Sigma}$ is controllable if and only if for every $k \in \{1, 2, \cdots, \bar{m}\}$ and every $\bar{y}^{[k]} \in \bar{\mathbf{Y}}^{[k]}$, $\Phi^T \bar{G}^{[2]}(\bar{\lambda}_0^{[k]})\bar{y}^{[k]} \neq \bar{y}^{[k]}$.*

Proof. The matrices A and B are clearly real by their definitions. Therefore, the existence of a nonzero complex vector x simultaneously satisfying $x^H B \neq 0$ and $x^H A = \lambda x^H$ is equivalent to the existence of a nonzero real vector x simultaneously satisfying $A^T x = \lambda x$ and $B^T x = 0$.

From these observations and from Theorems 3.1 and 3.2 it can be claimed that the controllability of the matrix pair $(A,\ B)$ is equivalent to the observability of the matrix pair $(A^T,\ B^T)$. On the other hand, from the definitions of the matrices A and B it can be proved that

$$A^T = A_{\mathbf{xx}}^T + A_{\mathbf{zx}}^T \Phi^T \left[I - A_{\mathbf{zv}}^T \Phi^T \right]^{-1} A_{\mathbf{xv}}^T, \tag{3.42}$$

$$B^T = B_{\mathbf{x}}^T + B_{\mathbf{v}}^T \Phi^T \left[I - A_{\mathbf{xx}}^T \Phi^T \right]^{-1} A_{\mathbf{xv}}^T, \tag{3.43}$$

which have completely the same forms respectively as those of the matrices A and C. The proof can now be completed through directly utilizing Theorem 3.7. □

In the above derivations, for each $i = 1,\ 2,\ \cdots,\ N$, both the TFMs $G_i^{[1]}(\lambda)$ and the TFMs $\bar{G}_i^{[1]}(\lambda)$ are required to be of FCNR. These requirements cannot be satisfied in general by an actual networked system. To be more specific, to satisfy the FCNR condition on the TFM of the ith subsystem $\mathbf{\Sigma}_i$, which is denoted by $G_i^{[1]}(\lambda)$, it is necessary that $m_{\mathbf{v}i} \geq m_{\mathbf{y}i}$, that is, the dimension of its internal input vector $v(k,i)$ is not smaller than that of its external output vector $y(k,i)$. Obviously, this is not a condition that can be easily satisfied by every engineering plant.

A possible approach to remove these conditions is to adopt some appropriate decompositions of the TFMs $G_i^{[1]}(\lambda)$ and $\bar{G}_i^{[1]}(\lambda)$, $i = 1,\ 2,\ \cdots,\ N$, which has been pointed out in [9] and investigated in detail in [24].

When there are constraints on the subsystem external input vectors, state vectors, and so on, some results have been obtained in [25,26]. These results inherit the properties of those without constraints and are computationally attractive for analysis and synthesis of a large-scale networked system.

3.4.4 In/Out-degree and Controllability/Observability of a Networked System

Some necessary and sufficient conditions have been derived in the previous subsections respectively for the controllability and observability of a networked system, which are computationally attractive and have clarified to some extent relations among system controllability and observability, dynamics of each subsystem in the plant, and subsystem interconnections. In this section, relations among system structure and its controllability/observability are further investigated, which is helpful in the settlement of the problem of determining the minimal number of inputs/outputs for each subsystem under the requirements that a controllable/observable networked system can be constructed from these subsystems. For this purpose, the following property of the subsystem connection matrix Φ is at first introduced. This property is firstly observed in [22] and plays an important role in the analysis of its stability and robust stability, which will be discussed in Chapter 7.

Define $M_{\star}$ and $M_{\star,i}$ respectively as $M_{\star}=\sum_{k=1}^{N}m_{\star k}$ and $M_{\star,i}=\sum_{k=1}^{i}m_{\star k}$ with $M_{\star,0}=0$, where $\star=\mathbf{v},\mathbf{z}$. Let e_k denote the kth canonical basis vector of the complex space $\mathcal{C}^{M_{\mathbf{z}}}$, that is, the $M_{\mathbf{z}}$-dimensional column vector with its kth row element 1 and all other elements zero. Moreover, let $j(i)$, $i=1,2,\cdots,M_{\mathbf{v}}$, denote the position of the nonzero element of the ith row of the SCM Φ. Then, from the assumptions on this matrix we have that $\Phi=\mathbf{col}\left\{e_{j(i)}^{T}\Big|_{i=1}^{M_{\mathbf{v}}}\right\}$. Let $m(i)$ stand for the number of subsystems that are directly affected by the ith element of the vector $z(k)$. Denote $\mathbf{diag}\left\{\sqrt{m(i)}\Big|_{i=M_{\mathbf{z},j-1}+1}^{M_{\mathbf{z},j}}\right\}$ by $\Theta(j)$, $j=1,2,\cdots,N$. Moreover, denote $\mathbf{diag}\left\{\Theta(j)|_{j=1}^{N}\right\}$ by Θ. Note that $e_k e_k^T=\mathbf{diag}\left\{0_{k-1}^{T},\ 1,\ 0_{M_{\mathbf{z}}-k}^{T}\right\}$. It can be shown through straightforward algebraic manipulations that

$$\begin{aligned}
\Phi^T\Phi &= \mathbf{col}^T\left\{e_{j(i)}^{T}\Big|_{i=1}^{M_{\mathbf{v}}}\right\}\mathbf{col}\left\{e_{j(i)}^{T}\Big|_{i=1}^{M_{\mathbf{v}}}\right\} \\
&= \mathbf{diag}\left\{m(i)|_{i=1}^{M_{\mathbf{z}}}\right\} \\
&= \Theta^2.
\end{aligned} \tag{3.44}$$

Obviously, from the definition of $m(i)$ we have that $\sum_{i=M_{\mathbf{z},j-1}+1}^{M_{\mathbf{z},j}}m(i)$ equals the out-degree of the jth subsystem of the networked system $\mathbf{\Sigma}$.

On the basis of this relation and Theorem 3.5, we obtain a necessary condition for the observability of system $\mathbf{\Sigma}$.

Lemma 3.3. *The networked system $\mathbf{\Sigma}$ is observable only if for each $i=1,2,\cdots,N$, the matrix pair $(A_{\mathbf{xx}}(i),\mathbf{col}\{C_{\mathbf{x}}(i),\ A_{\mathbf{zx}}(i)\})$ is observable.*

Proof. Define the matrix-valued polynomials $M_1(\lambda)$ and $\hat{M}_1(\lambda)$ respectively as

$$M_1(\lambda) = \begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} \\ -C_{\mathbf{x}} \\ -\Phi A_{\mathbf{zx}} \end{bmatrix}, \quad \hat{M}_1(\lambda) = \begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} \\ C_{\mathbf{x}} \\ \Theta A_{\mathbf{zx}} \end{bmatrix}. \tag{3.45}$$

Assume that the system $\mathbf{\Sigma}$ is observable. Then, according to Theorem 3.5, it is necessary that for every complex scalar λ, the matrix-valued polynomial $M(\lambda)$ defined in Eq. (3.29) is of FCR. From the definitions of the matrix-valued polynomials $M(\lambda)$ and $M_1(\lambda)$ it is obvious that the matrix-valued polynomial $M_1(\lambda)$ must be of FCR at every complex scale λ, which is equivalent to

$$M_1^H(\lambda)M_1(\lambda) > 0. \tag{3.46}$$

On the basis of Eqs. (3.44) and (3.45), the following equality can be straightforwardly established for each $\lambda \in \mathcal{C}$:

$$\begin{aligned} M_1^H(\lambda)M_1(\lambda) &= \left(\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^H \left(\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right) + C_{\mathbf{x}}^T C_{\mathbf{x}} + A_{\mathbf{zx}}^T \Phi^T \Phi A_{\mathbf{zx}} \\ &= \left(\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right)^H \left(\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right) + C_{\mathbf{x}}^T C_{\mathbf{x}} + A_{\mathbf{zx}}^T \Theta^2 A_{\mathbf{zx}} \\ &= \hat{M}_1^H(\lambda)\hat{M}_1(\lambda). \end{aligned} \tag{3.47}$$

It can therefore be declared that to guarantee the observability of the system $\mathbf{\Sigma}$, it is necessary that the matrix-valued polynomial $\hat{M}_1(\lambda)$ is of FCR at each complex scale λ.

From the block diagonal structure of the matrices $A_{\mathbf{xx}}$, $A_{\mathbf{zx}}$, and $C_{\mathbf{x}}$ and from Eq. (3.44) it is obvious that

$$\hat{M}_1(\lambda) = \begin{bmatrix} \mathbf{diag}\left\{\lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i)|_{i=1}^N\right\} \\ \mathbf{diag}\left\{C_{\mathbf{x}}(i)|_{i=1}^N\right\} \\ \mathbf{diag}\left\{\Theta(i)A_{\mathbf{zx}}(i)|_{i=1}^N\right\} \end{bmatrix}. \tag{3.48}$$

Define the matrix-valued polynomials $\hat{M}_{1i}(\lambda)$ and $\tilde{M}_{1i}(\lambda)$ with $i = 1, 2, \cdots, N$ as

$$\hat{M}_{1i}(\lambda) = \begin{bmatrix} \lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{x}}(i) \\ \Theta(i)A_{\mathbf{zx}}(i) \end{bmatrix}, \quad \tilde{M}_{1i}(\lambda) = \begin{bmatrix} \lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{x}}(i) \\ A_{\mathbf{zx}}(i) \end{bmatrix}.$$

Straightforward matrix manipulations show that for each fixed complex λ, the complex-valued matrix $\hat{M}_1(\lambda)$ is of FCR if and only if for each $i = 1, 2, \cdots, N$, the complex-valued matrix $\hat{M}_{1i}(\lambda)$ is of FCR. Moreover, clearly from the definitions of the matrix-valued polynomials $\hat{M}_{1i}(\lambda)$ and $\tilde{M}_{1i}(\lambda)$ we have that

$$\hat{M}_{1i}(\lambda) = \mathbf{diag}\left\{I_{m_{\mathbf{x}i}},\ I_{m_{\mathbf{y}i}},\ \Theta(i)\right\} \tilde{M}_{1i}(\lambda). \tag{3.49}$$

Note that the matrix $\Theta(i)$ is positive definite by its definition. It is clear that the matrix-valued polynomial $\hat{M}_{1i}(\lambda)$ is of FCR at every complex scale λ if and only if the matrix-valued polynomial $\tilde{M}_{1i}(\lambda)$ is.

The proof can now be completed through a direct application of Theorem 3.2. □

From the state space model of the subsystem $\mathbf{\Sigma}_i$ it is clear that both vectors $y(k,i)$ and $z(k,i)$ are its output vectors. In other words, when this subsystem is isolated from the influences of other subsystems and its influences to other subsystems are also completely removed, then the observability of the subsystem $\mathbf{\Sigma}_i$ is equivalent to that of the matrix pair $(A_{\mathbf{xx}}(i), \mathbf{col}\{C_{\mathbf{x}}(i),\ A_{\mathbf{zx}}(i)\})$. Hence, the results of Lemma 3.3 imply that to construct an observable networked system, each its subsystem should be observable.

Note that observability of the matrix pair $(A_{\mathbf{xx}}(i), \mathbf{col}\{C_{\mathbf{x}}(i),\ A_{\mathbf{zx}}(i)\})$ is not equivalent to that of the matrix pair $(A_{\mathbf{xx}}(i),\ C_{\mathbf{x}}(i))$. In fact, from Theorem 3.2 it is clear that if the matrix pair $(A_{\mathbf{xx}}(i),\ C_{\mathbf{x}}(i))$ is observable, then the matrix pair $(A_{\mathbf{xx}}(i), \mathbf{col}\{C_{\mathbf{x}}(i),\ A_{\mathbf{zx}}(i)\})$ is also observable; but the converse is in general not true. The results of Lemma 3.3 therefore also imply that even when there exist subsystems that are not observable through *only* their own external outputs, the whole networked system may still be observable by means of subsystem connections.

Similar results have been observed in [6] for system controllability. However, the conclusions there depend on the subsystem connection matrix Φ. This makes them difficult to be applied in constructing a networked system that is controllable, as an appropriate subsystem connection is usually not known before system designs. On the other hand, note that $\mathbf{col}\{\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}},\ -C_{\mathbf{x}},\ -\Phi A_{\mathbf{zx}}\} = \mathbf{diag}\{I_{M_{\mathbf{x}}},\ -I_{M_{\mathbf{y}}},\ -\Phi\}\mathbf{col}\{\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}},\ C_{\mathbf{x}},\ A_{\mathbf{zx}}\}$. This means that to guarantee that the matrix $\mathbf{col}\{\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}},\ -C_{\mathbf{x}},\ -\Phi A_{\mathbf{zx}}\}$ is of FCR, it is necessary that the matrix $\mathbf{col}\{\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}},\ C_{\mathbf{x}},\ A_{\mathbf{zx}}\}$ is. Based on these observation, similar arguments as those in the proof of Lemma 3.3 show that the associated conclusions in [6] are in fact valid for an arbitrary subsystem connection matrix Φ.

To establish a relation between system observability and its subsystem out-degrees, define TFMs $G^{[1]}(\lambda)$ and $G^{[2]}(\lambda)$ respectively as $G^{[1]}(\lambda) = \mathbf{diag}\{G_i^{[1]}(\lambda)|_{i=1}^{N}\}$ and $G^{[2]}(\lambda) = \mathbf{diag}\{G_i^{[2]}(\lambda)|_{i=1}^{N}\}$, where $G_i^{[1]}(\lambda) = C_{\mathbf{v}}(i) + C_{\mathbf{x}}(i)[\lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i)]^{-1}A_{\mathbf{xv}}(i)$ and $G_i^{[2]}(\lambda) = A_{\mathbf{zv}}(i) + A_{\mathbf{zx}}(i)[\lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i)]^{-1}A_{\mathbf{xv}}(i)$ for each $i = 1, 2, \cdots, N$. From the block diagonal structure of the TFM $G^{[1]}(\lambda)$ it is clear that this TFM is of FCNR if and only if each of the TFMs $G_i^{[1]}(\lambda)$, $i \in \{1, 2, \cdots, N\}$, is.

Assume that the TFMs $G^{[1]}(\lambda)$ and $G_i^{[1]}(\lambda)$ have respectively m and m_i distinctive transmission zeros. Then, under the condition that the TFM $G^{[1]}(\lambda)$ is of FCNR, it is obvious from

Lemma 3.1 and from $G^{[1]}(\lambda) = \mathbf{diag}\{G_i^{[1]}(\lambda)|_{i=1}^N\}$ that, for each $i = 1, \cdots, N$, every transmission zero of $G_i^{[1]}(\lambda)$ is also a transmission zero of $G^{[1]}(\lambda)$. As argued in [9], we generally only have that $\max_{1 \le i \le N} m_i \le m \le \sum_{i=1}^N m_i$. Moreover, for each of the transmission zeros of the TFM $G^{[1]}(\lambda)$, there exists at least one integer i belonging to the set $\{\,1,\ 2,\ \cdots,\ N\,\}$ such that it is also a transmission zero of the TFM $G_i^{[1]}(\lambda)$.

Let $\lambda_0^{[k]}$ denote the kth transmission zero of the TFM $G^{[1]}(\lambda)$, $k = 1, 2, \cdots, m$. Assume that in the TFM set $\{G_1^{[1]}(\lambda),\ G_2^{[1]}(\lambda),\ \cdots,\ G_N^{[1]}(\lambda)\}$, there are $s^{[k]}$ TFMs that have this transmission zero. Denote them by $G_{k(s)}^{[1]}(\lambda)$, $s = 1, \cdots, s^{[k]}$. Clearly, both $s^{[k]}$ and $k(s)$ belong to the set $\{1,\ 2,\ \cdots,\ N\}$. As in [9] and in the previous section, we once again assume, without any loss of generality, that $k(1) < k(2) < \cdots < k(s^{[k]})$. Let $Y_s^{[k]}$ denote the matrix constructed from a set of linear independent vectors that span the null space of $G_{k(s)}^{[1]}(\lambda_0^{[k]})$, and let $p(k, s)$ denote the dimension of this null space. Obviously, the matrix $Y_s^{[k]}$ is of FCR, which further leads to that the matrix $Y_s^{[k]H} Y_s^{[k]}$ is positive definite. Hence, the matrix

$$\Gamma_s^{[k]} = G_{k(s)}^{[2]}(\lambda_0^{[k]}) Y_s^{[k]} \left(Y_s^{[k]H} Y_s^{[k]}\right)^{-1/2} \tag{3.50}$$

is well defined for each $s = 1, 2, \cdots, s^{[k]}$ and each $k = 1, 2, \cdots, m$.

Using these matrices, we derive the following conclusion, which gives a sufficient condition for the observability of the networked system $\mathbf{\Sigma}$. Their proof is deferred to the appendix.

Theorem 3.8. *Assume that all the TFMs $G_i^{[1]}(\lambda)|_{i=1}^N$ are of FCNR. Let $\{\lambda_0^{[k]}|_{k=1}^m\}$ denote the set of distinctive transmission zero of the TFM $G^{[1]}(\lambda)$. If the matrix Θ satisfies simultaneously the inequality*

$$I_{p(k,s)} - \Gamma_s^{[k]H} \Theta^2(k(s)) \Gamma_s^{[k]} > 0 \tag{3.51}$$

or

$$I_{p(k,s)} - \Gamma_s^{[k]H} \Theta^2(k(s)) \Gamma_s^{[k]} < 0 \tag{3.52}$$

for each $s = 1, 2, \cdots, s^{[k]}$ and $k = 1, 2, \cdots, m$, then the dynamic system $\mathbf{\Sigma}$ is observable.

Compared with the results given in Theorem 3.4, which is originally derived in [9], the conditions of Theorem 3.8 are only sufficient. On the other hand, these conditions can be verified individually for each subsystem and therefore have a much lower computational complexity, and the computation results are generally more numerically reliable. In particular, in the above conditions, the dimension of the involved matrix is $p(k, s) \times p(k, s)$, whereas that in Theorem 3.4 is $\sum_{i=1}^N m_{\mathbf{v}i} \times \sum_{i=1}^{s^{[k]}} p(k, i)$. Obviously, the latter is usually significantly greater than the former for a large-scale system, which is less attractive from the viewpoint of computations.

Note that the matrix Θ is closely related to the out-degrees of the networked system $\mathbf{\Sigma}$. Theorem 3.8 in fact establishes some relations between the observability of a networked system and its subsystem out-degrees. This theorem, together with the following Theorem 3.9, which is the counterpart of this theorem in controllability verifications, is essential in obtaining the minimal input/output number for each subsystem such that a controllable/observable system can be constructed.

Remark 3.2. *Note that, by definition, $\Theta(j) \geq I_{m_{\mathbf{z}j}}$ for each $j = 1, 2, \cdots, N$. It can be easily understood that if there is an integer pair (k, s) with $k \in \{1, 2, \cdots, m\}$ and $s \in \{1, 2, \cdots, s^{[k]}\}$ such that the associated matrix $\Gamma_s^{[k]}$ is not of FCR, then for any subsystem connection matrix Φ, the associated inequality $I_{p(k,s)} - \Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]} < 0$ cannot be satisfied. Hence, to satisfy the conditions of Theorem 3.8, one possible approach is to meet the inequality $I_{p(k,s)} - \Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]} > 0$. This might be achieved by reducing the number of subsystems that an internal output straightforwardly affects. These observations mean that under such a situation, sparse subsystem connections might be helpful to make a networked system observable.*

On the contrary, if for each $s = 1, 2, \cdots, s^{[k]}$ and each $k = 1, 2, \cdots, m$, the associated matrix $\Gamma_s^{[k]}$ is always of FCR, then the minimal eigenvalue of the matrix $\Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]}$ can be made large by increasing the number of subsystems that an internal output directly influences, which implies that the inequality $I_{p(k,s)} - \Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]} < 0$ might be satisfied through simply increasing the number of subsystem connections; that is, dense subsystem connections are appreciated from the viewpoint of system observability.

Remark 3.3. *Although the matrix $Y_s^{[k]}$ is not unique for each integer pair $(k,\ s)$, its selection does not have any influence on the satisfaction of the conditions of Eqs. (3.51) and (3.52), which can be straightforwardly proven from relations among different basis vectors of a subspace.*

When controllability is to be investigated by means of the duality between controllability and observability of an LTI system, which has already been adopted in [9] and in the previous section, similar results can be derived through completely the same arguments. More precisely, based on this duality and the state space model of the whole system given in [9], it can be directly declared that when the networked system $\mathbf{\Sigma}$ is well-posed, it is controllable if and only if for each complex scale λ, the following matrix-valued polynomial $\bar{M}(\lambda)$ is of FRR [6,9]:

$$\bar{M}(\lambda) = \begin{bmatrix} \lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}} & -B_{\mathbf{x}} & -A_{\mathbf{xv}}\Phi \\ -A_{\mathbf{zx}} & -B_{\mathbf{v}} & I_{M_{\mathbf{v}}} - A_{\mathbf{zv}}\Phi \end{bmatrix}.$$

Note that the transpose of the matrix-valued polynomial $\bar{M}(\lambda)$ has completely the same form as that of the matrix-valued polynomial $M(\lambda)$. It is not out of imaginations that necessary/sufficient conditions similar to those of Lemma 3.3 and Theorem 3.8 can be derived for controllability verifications of a networked system.

However, to achieve these conclusions, it appears necessary to assume that every column of the subsystem connection matrix Φ only has one nonzero element. Although this condition can be satisfied in general through augmenting the subsystem internal output vectors $z(k,i)|_{i=1}^{N}$ with repeated elements, the augmentation usually violates an associated FCNR condition and therefore greatly restricts applicability of the associated results.

In this paper, we derive another necessary/sufficient condition for system controllability without that assumption.

For this purpose, define TFMs $\bar{G}^{[1]}(\lambda)$ and $\bar{G}^{[2]}(\lambda)$ respectively as $\bar{G}^{[1]}(\lambda)=\mathbf{diag}\{\bar{G}_i^{[1]}(\lambda)|_{i=1}^{N}\}$ and $\bar{G}^{[2]}(\lambda)=\mathbf{diag}\{\bar{G}_i^{[2]}(\lambda)|_{i=1}^{N}\}$, where $\bar{G}_i^{[1]}(\lambda)=B_{\mathbf{S}}^{T}(i)+B_{\mathbf{T}}^{T}(i)[\lambda I_{m_{\mathbf{Ti}}}-A_{\mathbf{TT}}^{T}(i)]^{-1}A_{\mathbf{ST}}^{T}(i)$ and $\bar{G}_i^{[2]}(\lambda)=(G_i^{[2]}(\lambda))^{T}$. Assume that the TFM $\bar{G}^{[1]}(\lambda)$ has $\bar{m}$ distinctive transmission zeros, which are denoted by $\bar{\lambda}_0^{[k]}|_{k=1}^{\bar{m}}$. Moreover, let $\bar{G}_{\bar{k}(s)}^{[1]}(\lambda)|_{s=1}^{\bar{s}^{[k]}}$ represent the TFMs that have $\bar{\lambda}_0^{[k]}$ as its transmission zero, and let $\bar{k}(1)<\bar{k}(2)<\cdots<\bar{k}(\bar{s}^{[k]})$. Furthermore, let $\bar{p}(k,s)$ denote the dimension of the null space of the matrix $\bar{G}_{\bar{k}(s)}^{[1]}(\bar{\lambda}_0^{[k]})$, and let $\bar{Y}_s^{[k]}$ be the matrix constructed from a set of linear independent vectors that span this null space. Define the matrix

$$\bar{\Gamma}_s^{[k]}=\bar{G}_{\bar{k}(s)}^{[2]}(\bar{\lambda}_0^{[k]})\bar{Y}_s^{[k]}\left(\bar{Y}_s^{[k]H}\Theta^{-2}(\bar{k}(s))\bar{Y}_s^{[k]}\right)^{-1/2}. \tag{3.53}$$

Then, we have the following results, whose proof is included in the appendix.

Theorem 3.9. *Assume that the TFM $\bar{G}^{[1]}(\lambda)$ is of FCNR. Then, System $\mathbf{\Sigma}$ is controllable only when the matrix pair $(A_{\mathbf{xx}}(i),\ [B_{\mathbf{x}}(i)\ \ A_{\mathbf{xv}}(i)])$ is controllable for every $i=1,2,\cdots,N$. Moreover, if for each integer pair (k,s) with $k\in\{1,2,\cdots,\bar{m}\}$ and $s\in\{1,2,\cdots,\bar{s}^{[k]}\}$, the matrix inequality*

$$I_{\bar{p}(k,s)}-\bar{\Gamma}_s^{[k]H}\bar{\Gamma}_s^{[k]}>0 \tag{3.54}$$

is satisfied, then this system is controllable.

It is interesting to notice that although the necessary condition of Theorem 3.9 is dual to that of Lemma 3.3, its sufficient condition differs significantly from that of Theorem 3.8. Moreover, their proofs are also not completely dual to each other. These are due to that in order to apply the duality between controllability and observability, the subsystem connection matrix Φ must satisfy the condition that $\Phi\Phi^{T}$ is a diagonal matrix, which cannot be met in general.

By the definition of the matrix $\bar{\Gamma}_s^{[k]}$, careful comparisons between Eqs. (3.54) and (3.51) show that some qualitative relations exist between in-degrees and controllability of a networked system, which are similar to those between its out-degrees and observability.

3.5 Construction of Controllable/Observable Networked Systems

In the design of a networked system, it is often necessary to know that to meet the required performances for the whole system, what specifications each subsystem must satisfy [10,11, 16–18,27]. Note that controllability and observability are essential for a system to work properly, recalling that to reconstruct the states of a system from measured input–output data, it is necessary that the system is observable. Moreover, controllability is necessary for a system to perform satisfactorily [2,9]. On the other hand, in actual engineering, it is generally appreciative to have inputs/outputs directly and separately affecting/measuring the states of each individual subsystem and/or their functions [9–11,18,27]. It is therefore natural to ask how many sensors are required to monitor the states for each of its subsystem to guarantee that a controllable/observable networked system can be constructed, as well as how many actuators are required to maneuver the states for each of its subsystem.

In this section, we investigate the minimal number of outputs/inputs required for each subsystem that guarantee construction of an observable/controllable networked system. To avoid possible confusions, an actuator that *directly* affects some states of *only* one subsystem is called a local actuator, whereas an output that *directly* depends on some states of *only* one subsystem is called a local external output.

The following theorem gives an answer to this minimal input/output problem. Its proof is provided in the appendix.

Theorem 3.10. *Let $p_{\max}(i)$ denote the maximum geometric multiplicity of the matrix $A_{\mathbf{xx}}(i)$, $i = 1, 2, \cdots, N$. Then an observable networked system $\mathbf{\Sigma}$ can be constructed with local external outputs if and only if*

$$m_{\mathbf{y}i} + m_{\mathbf{z}i} \geq p_{\max}(i), \quad \forall i \in \{1,\ 2,\ \cdots,\ N\}.$$

Moreover, a controllable networked system $\mathbf{\Sigma}$ can be constructed with local actuators if and only if

$$m_{\mathbf{u}i} + m_{\mathbf{v}i} \geq p_{\max}(i), \quad \forall i \in \{1,\ 2,\ \cdots,\ N\}.$$

Here m_{*i} stands for the dimension of the column vector $*(k,i)$ with $* = u,\ v,\ y,\ z$.

Remark 3.4. *This theorem reveals that to reduce the required number of external inputs/outputs, it is better to design a subsystem with its STM having distinctive eigenvalues. This is in a good agreement with the results on a lumped system reported in [5].*

Corollary 3.3. *To be able to build a controllable/observable networked system from several subsystems, it is necessary and sufficient that each subsystem is controllable/observable.*

Proof. This is an immediate result of Theorem 3.5 and Corollary 3.1 together with Theorem 3.9. □

Note that the matrices $A_{\mathbf{zx}}(i)$, $A_{\mathbf{zv}}(i)$, and $A_{\mathbf{xv}}(i)$ represent connection strengths among subsystems of the system $\mathbf{\Sigma}$. The bigger the magnitude of the elements of these matrices, the tighter the subsystems are connected [9]. On the other hand, it is clear from the proof of Theorem 3.10 that when each subsystem is observable/controllable, through reducing subsystem connection strengths, it is always possible to construct an observable/controllable networked system. In the extreme situation, when all the subsystems are disconnected, the networked system becomes a collection of isolated individual observable/controllable subsystems, which is obviously observable/controllable.

On the other hand, when these matrices are appropriately selected such that the corresponding matrices $\Gamma_s^{[k]}$ are of FCR for each integer pair $(k,\ s)$, it can be easily seen from Eq. (3.52) that through increasing magnitudes of the elements of these matrices, that is, through increasing subsystem connection strengths, it is also possible to build an observable networked system using observable subsystems. Similar conclusions can be obtained for building a controllable networked system by means of the duality between observability and controllability.

However, when there are some restrictions on the subsystem connection matrix Φ and/or on subsystem connection strengths, which is often required in practical engineering [9,18,19], further efforts are still necessary to find the minimal number of inputs/outputs for each subsystem in the construction of a controllable/observable system.

3.6 Bibliographic Notes

Controllability and observability are respectively related to system properties about capabilities of maneuvering and estimating plant states. Originally, these issues are investigated without any constraints on the plant input vector and state vector, which leads to the now extensively known PBH test, rank conditions on the controllability matrix, rank conditions on the observability matrix, and so on. A preliminary assumption for these investigations are that the plant parameters are known. Studies along this line have produced many important results of system theories, such as the controllable canonical form, observable canonical form of a state space model, division of a plant state space into four subspaces, which are respectively controllable and observable, controllable but unobservable, observable but uncontrollable, and uncontrollable and unobservable [2,3].

When the plant parameters are not provided, a concept called structural controllability/observability is developed, which depends only on the positions of the plant inputs and the directed connections among the states of a plant. In these studies, graph theory has played an important role. According to the requirement that whether there exists a group of parameters that make the system have this property or all the parameters can make the system possess this property, structural controllability/observability is further divided into weak and strong controllability/observability [18,28]. Especially, these concepts are adopted in the analysis and synthesis of a large-scale system in which influences among subsystems are through their states [12,13].

In practical engineering, there are usually some constraints on plant inputs and/or states. Under these constraints, controllability and observability verification becomes much more mathematically involved, and there are still no general results [29–31]. However, for some specific situations, solid results have been established. For example, when plant inputs are constrained in magnitude, it has been proved that controllability of an LTI system is equivalent to its controllability without constraints and all its eigenvalues are antistable [28]. These results have partially been extended to networked systems that are computationally attractive for a system with a large number of subsystems [25,26].

The problem of finding the smallest number of inputs such that a lumped controllable system can be constructed appears to be firstly investigated in [32]. This problem has also been discussed in [28,33] afterward. In [28], however, only a necessary condition is given, and its proof is left as an exercise. On the other hand, the Jordan form of the system state transition matrix is used in [32] to decompose the system state space into controllable and uncontrollable subspaces. This decomposition is not very appropriate, as system states usually take only real values that cannot be guaranteed by this decomposition in general. Moreover, although the necessity of a condition is established in [33] through a direct application of the PBH test, its sufficiency is illustrated only through a numerical example, and in this illustration, the Jordan form was straightforwardly used once again, and the constructed input matrix cannot be guaranteed to be real. A complete settlement of this problem appears to be first given in [5]. The problem of searching the minimum number of inputs and outputs for each subsystem in a networked system seems to be originally investigated in [34].

Appendix 3.A

3.A.1 Proof of Theorem 3.4

For each $* = r, c$ and $i = 1, 2, \cdots, k_*$, define the matrix

$$Z_{*,i} = \mathbf{diag}\left\{ \left[\begin{array}{c} 1 \\ 0_{m_{*,i,j}-1} \end{array} \right] \Bigg|_{j=1}^{p_*(i)} \right\}. \tag{3.A.1}$$

From Eq. (3.17) it can be straightforwardly proven that the matrix $X_{*,i}$ of Eq. (3.12) can be represented as

$$X_{*,i} = T^{-H} \begin{bmatrix} 0_{(M_{*,i}-1)\times p_*(i)} \\ Z_{*,i} \\ 0_{(n-M_{*,i})\times p_*(i)} \end{bmatrix}, \tag{3.A.2}$$

in which $M_{r,i} = \sum_{j=1}^{i-1} m_{r,j}$ and $M_{c,i} = M_{r,k_r} + \sum_{j=1}^{i-1} m_{c,j}$. Hence

$$B^T X_{*,i} = (T^{-1}B)^H \begin{bmatrix} 0_{(M_{*,i}-1)\times p_*(i)} \\ Z_{*,i} \\ 0_{(n-M_{*,i})\times p_*(i)} \end{bmatrix}. \tag{3.A.3}$$

Assume now that the input matrix B is given by Eq. (3.21). From the definition of the matrix T it is obvious that this matrix is real. Moreover, for each $i = 1, 2, \cdots, k_r$ with $* = r$ and each $i = 1, 2, \cdots, \frac{k_c}{2}$ with $* = c$, direct algebraic manipulations show that

$$\begin{aligned} B^T X_{*,i} &= \begin{bmatrix} \mathbf{col}\left\{\hat{B}_{r,i}|_{i=1}^{k_r}\right\} \\ \mathbf{col}\left\{\hat{B}_{c,i}|_{i=1}^{k_c}\right\} \end{bmatrix}^H \begin{bmatrix} 0_{(M_{*,i}-1)\times p_*(i)} \\ Z_{*,i} \\ 0_{(n-M_{*,i})\times p_*(i)} \end{bmatrix} \\ &= \hat{B}_{*,i}^H Z_{*,i} \\ &= \bar{\tilde{B}}_{*,i}. \end{aligned} \tag{3.A.4}$$

Furthermore, for each $i = \frac{k_c}{2} + 1, \ \cdots, k_c$, similar arguments show that

$$B^T X_{*,i} = \hat{B}_{c,i-k_c/2}^H Z_{c,i} = \tilde{B}_{c,i-k_c/2}. \tag{3.A.5}$$

Note that when a matrix is of FCR, its conjugate is also of FCR. It can therefore be declared that when the matrix $\tilde{B}_{*,i}$ is of FCR for each $i = 1, 2, \cdots, k_r$ with $* = r$ and each $i = 1, 2, \cdots, \frac{k_c}{2}$ with $* = c$, Lemma 3.1 implies that the system is controllable.

On the contrary, assume that the system is controllable. Then, according to Lemma 3.1, the matrix $B^T X_{*,i}$ is necessarily of FCR for each feasible $*$ and i. Construct the matrices $\hat{B}_{*,i} \in \mathcal{C}^{m_{*,i}\times p_{max}}$, $* = r, c$, $i = 1, 2, \cdots, k_*$, satisfying the equation

$$\mathbf{col}\left\{\mathbf{col}\left\{\hat{B}_{r,i}|_{i=1}^{k_r}\right\}, \ \mathbf{col}\left\{\hat{B}_{c,i}|_{i=1}^{k_c}\right\}\right\} = T^{-1}B. \tag{3.A.6}$$

As the matrix T is invertible, this construction is obviously always possible. Moreover, by Eq. (3.A.18) the corresponding $\tilde{B}_{*,i}$ is always of FCR.

Denote the real and imaginary parts of the matrices $\hat{B}_{*,i}$ and $\left[T_{c,i,j}\right]_{j=1,i=1}^{j=p_c(i),i=k_c/2}$ respectively by $\hat{B}_{*,i,r}$ and $\hat{B}_{*,i,j}$ and by $\hat{T}_{c,r}$, and $\hat{T}_{c,j}$. Using the matrix T_r of Eq. (3.16), define the matrix $\hat{T}$ as $\hat{T} = [T_r\ \hat{T}_{c,r}\ \hat{T}_{c,j}]$. Recall that the matrix T is nonsingular, and it can be straightforwardly proven that the matrix $\hat{T}$ is also invertible. On the basis of these matrices, direct algebraic manipulations show that

$$\Im\left\{T\begin{bmatrix}\mathbf{col}\left\{\hat{B}_{r,i}|_{i=1}^{k_r}\right\}\\ \mathbf{col}\left\{\hat{B}_{c,i}|_{i=1}^{k_c}\right\}\end{bmatrix}\right\} = \hat{T}\begin{bmatrix}\mathbf{col}\left\{\hat{B}_{r,i,j}|_{i=1}^{k_r}\right\}\\ \mathbf{col}\left\{\hat{B}_{c,i,j}+\hat{B}_{c,i+k_c/2,j}|_{i=1}^{k_c/2}\right\}\\ \mathbf{col}\left\{\hat{B}_{c,i,r}-\hat{B}_{c,i+k_c/2,r}|_{i=1}^{k_c/2}\right\}\end{bmatrix} \tag{3.A.7}$$

Note that the matrix B is real. From this relation and Eq. (3.A.6) it can be declared that it is necessary that

$$\hat{B}_{r,i,j} = 0 \tag{3.A.8}$$

for each $i = 1, 2, \cdots, k_r$ and

$$\hat{B}_{c,i,j} = -\hat{B}_{c,i+k_c/2,j}, \qquad \hat{B}_{c,i,r} = \hat{B}_{c,i+k_c/2,r} \tag{3.A.9}$$

for each $i = 1, 2, \cdots, \frac{k_c}{2}$. Hence, $\hat{B}_{r,i} \in \mathcal{R}^{m_{r,i}\times p_{max}}$ for each $i = 1, 2, \cdots, k_r$, and $\hat{B}_{c,i+k_c/2} = \bar{\hat{B}}_{c,i}$ for each $i = 1, 2, \cdots, \frac{k_c}{2}$. This completes the proof. □

3.A.2 Proof of Theorem 3.8

From Theorem 3.7 it can be easily seen that system $\mathbf{\Sigma}$ is observable if and only if for each nonzero vector $x \in \mathcal{C}^{M_\mathbf{x}+M_\mathbf{v}}$, there exists $\lambda \in \mathcal{C}$ such that

$$\begin{bmatrix}\lambda I_{M_\mathbf{x}} - A_{\mathbf{xx}} & -A_{\mathbf{xv}}\\ -C_\mathbf{x} & -C_\mathbf{v}\end{bmatrix}x = 0, \tag{3.A.10}$$

then with the same complex number λ, the following inequality is valid:

$$\left[-\Phi A_{\mathbf{zx}}\ \ I_{M_\mathbf{v}} - \Phi A_{\mathbf{zv}}\right]x \neq 0. \tag{3.A.11}$$

Partition the vector x as $x = \left[\, x_1^T\ \ x_2^T \,\right]$ where $x_1 \in \mathcal{C}^{M_\mathbf{x}}$ and $x_2 \in \mathcal{C}^{M_\mathbf{v}}$. Then, according to Eq. (3.A.10), we have that

$$\left[\lambda I_{M_\mathbf{x}} - A_{\mathbf{xx}}\right]x_1 - A_{\mathbf{xv}}x_2 = 0, \tag{3.A.12}$$

$$C_\mathbf{x}x_1 + C_\mathbf{v}x_2 = 0. \tag{3.A.13}$$

When λ is not an eigenvalue of the matrix $A_{\mathbf{xx}}$, the matrix $\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}$ is invertible. In this case, Eq. (3.A.12) implies that $x_1 = \left[\lambda I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right]^{-1} A_{\mathbf{xv}} x_2$. After substituting this relation into Eqs. (3.A.11) and (3.A.13), direct algebraic manipulations show that

$$G^{[1]}(\lambda)x_2 = 0, \tag{3.A.14}$$

$$\left[I_{M_{\mathbf{v}}} - \Phi G^{[2]}(\lambda)\right] x_2 \neq 0. \tag{3.A.15}$$

In these derivations, the definitions of the TFMs $G^{[1]}(\lambda)$ and $G^{[2]}(\lambda)$ have been utilized.

When λ is an eigenvalue of the matrix $A_{\mathbf{xx}}$, a pseudo-inverse must be taken, and the treatments are completely the same as those in [9,24]. In particular, note that the dimension of the matrix $A_{\mathbf{xx}}$ is finite, which means that all its eigenvalues can only take an isolated value. Hence, there exists $\varepsilon > 0$ that in general depends on the value of λ such that for each $\delta \in (-\varepsilon,\ \varepsilon)/\{0\}$, the matrix $(\lambda - \delta)I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}$ is invertible. These imply that the vector x_1 satisfying Eq. (3.A.12) can be formally expressed as

$$x_1 = \lim_{\delta \to 0} \left[(\lambda - \delta) I_{M_{\mathbf{x}}} - A_{\mathbf{xx}}\right]^{-1} A_{\mathbf{xv}} x_2. \tag{3.A.16}$$

Using this expression, conclusions can be obtained which are completely the same as those for the case where λ is not an eigenvalue of the matrix $A_{\mathbf{xx}}$.

Note that every TFM $G_i^{[1]}(\lambda)$, $i = 1, 2, \cdots, N$, is assumed to be of FCNR, and the TFM $G^{[1]}(\lambda)$ is block diagonal with its ith diagonal block being $G_i^{[1]}(\lambda)$. It is obvious that the TFM $G^{[1]}(\lambda)$ is also of FCNR. It can therefore be declared from Lemma 3.1 and Eq. (3.A.14) that λ is a transmission zero of the TFM $G^{[1]}(\lambda)$. These results imply that when the TFMs $G_i^{[1]}(\lambda)|_{i=1}^{N}$ are of FCNR, verifications of the conditions in Theorem 3.7 are necessary only for all transmission zeros of the TFM $G^{[1]}(\lambda)$.

Assume that $\lambda = \lambda_0^{[k]}$. Then, according to the definition of the number $\lambda_0^{[k]}$, it is also a transmission zero of the TFM $G_{k(s)}^{[1]}(\lambda)$, $s = 1, 2, \cdots, s^{[k]}$. Moreover, from the definition of the matrix $Y_s^{[k]}$ we have that, for every nonzero complex valued vector $\alpha_s \in \mathcal{C}^{p(k,s)}$,

$$G_{k(s)}^{[1]}(\lambda_0^{[k]}) Y_s^{[k]} \alpha_s = 0. \tag{3.A.17}$$

Define the matrix

$$Y^{[k]} = \begin{bmatrix} 0_{M_{\mathbf{v},k(1)-1} \times p(k,1)} & \cdots & 0_{M_{\mathbf{v},k(s^{[k]})-1} \times p(k,s^{[k]})} \\ Y_1^{[k]} & \cdots & Y_{s^{[k]}}^{[k]} \\ 0_{(M_{\mathbf{v}} - M_{\mathbf{v},k(1)}) \times p(k,1)} & \cdots & 0_{(M_{\mathbf{v}} - M_{\mathbf{v},k(s^{[k]})}) \times p(k,s^{[k]})} \end{bmatrix}.$$

Then from the block diagonal structure of the TFM $G^{[1]}(\lambda)$ and Eq. (3.A.17) it can be directly proven that, for each nonzero vector $x_2 \in \mathcal{C}^{M_{\mathbf{v}}}$ satisfying $G^{[1]}(\lambda_0^{[k]})x = 0$, there exists a unique nonzero $\alpha \in \mathcal{C}^{\sum_{j=1}^{s^{[k]}} p(k,j)}$ such that

$$x_2 = Y^{[k]}\alpha. \tag{3.A.18}$$

On the other hand, based on the block diagonal structures of the TFM $G^{[2]}(\lambda)$ and the matrix Θ, direct algebraic manipulations show that for each complex valued vector x_2 satisfying Eq. (3.A.18), we have that

$$\begin{aligned}\Theta G^{[2]}(\lambda_0^{[k]})x_2 &= \mathbf{diag}\{\Theta(i)|_{i=1}^{N}\}\mathbf{diag}\{G_i^{[2]}(\lambda)|_{i=1}^{N}\}Y^{[k]}\alpha \\ &= \begin{bmatrix} 0_{M_{\mathbf{v},k(1)-1}\times p(k,1)} & \cdots & 0_{M_{\mathbf{v},k(s^{[k]})-1}\times p(k,s^{[k]})} \\ \Theta(k(1))G_{k(1)}^{[2]}(\lambda)Y_1^{[k]} & \cdots & \Theta(k(s^{[k]}))G_{k(s^{[k]})}^{[2]}(\lambda)Y_{s^{[k]}}^{[k]} \\ 0_{(M_{\mathbf{v}}-M_{\mathbf{v},k(1)})\times p(k,1)} & \cdots & 0_{(M_{\mathbf{v}}-M_{\mathbf{v},k(s^{[k]})})\times p(k,s^{[k]})} \end{bmatrix}\alpha. \end{aligned} \tag{3.A.19}$$

Hence,

$$x_2^H x_2 = \alpha^H \mathbf{diag}\left\{Y_j^{[k]H} Y_j^{[k]}|_{j=1}^{s^{[k]}}\right\}\alpha. \tag{3.A.20}$$

Moreover, from Eq. (3.44) we have that

$$\begin{aligned}\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)^H \left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right) &= x_2^H G^{[2]H}(\lambda_0^{[k]})\Theta^2 G^{[2]}(\lambda_0^{[k]})x_2 \\ &= \left(\Theta G^{[2]}(\lambda_0^{[k]})x_2\right)^H \left(\Theta G^{[2]}(\lambda_0^{[k]})x_2\right). \end{aligned} \tag{3.A.21}$$

Substituting the right-hand side of Eq. (3.A.19) into that of Eq. (3.A.21), it can be directly proven that

$$\begin{aligned}&\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)^H \left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right) \\ &= \alpha^H \mathbf{diag}\left\{\left(\Theta(k(j))G_{k(j)}^{[2]}(\lambda_0^{[k]})Y_j^{[k]}\right)^H \left(\Theta(k(j))G_{k(j)}^{[2]}(\lambda_0^{[k]})Y_j^{[k]}\right)\Big|_{j=1}^{s^{[k]}}\right\}\alpha. \end{aligned} \tag{3.A.22}$$

Denote the vector $\mathbf{diag}\{(Y_j^{[k]H} Y_j^{[k]})^{1/2}|_{j=1}^{s^{[k]}}\}\alpha$ by $\hat{\alpha}$. It can be declared from the FCR property of the matrices $Y_j^{[k]}|_{j=1}^{s^{[k]}}$ that the vector $\hat{\alpha}$ is not equal to zero if and only if the vector α is. On the other hand, from Eqs. (3.A.20) and (3.A.22) and from the definitions of the matrices $\Gamma_j^{[k]}|_{j=1}^{s^{[k]}}$ straightforward algebraic manipulations show that

$$x_2^H x_2 - \left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)^H \left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right) = \hat{\alpha}^H \mathbf{diag}\left\{ I_{p(k,s)} - \Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]}\Big|_{s=1}^{s^{[k]}}\right\}\hat{\alpha}. \tag{3.A.23}$$

Therefore, if the inequality of Eq. (3.51) is satisfied for each $s=1,2,\cdots,s^{[k]}$, then the matrix $\mathbf{diag}\{I_{p(k,s)}-\Gamma_s^{[k]H}\Theta^2(k(s))\Gamma_s^{[k]}|_{s=1}^{s^{[k]}}\}$ is positive definite. This means that for an arbitrary nonzero complex vector x_2 satisfying Eq. (3.A.14), we have that

$$x_2^H x_2-\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)^H\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)>0. \tag{3.A.24}$$

On the other hand, if for every $s\in\left\{1,2,\cdots,s^{[k]}\right\}$, the inequality (3.52) is satisfied, then similar arguments show that for each nonzero complex vector x_2 satisfying Eq. (3.A.14), the following inequality is satisfied:

$$x_2^H x_2-\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)^H\left(\Phi G^{[2]}(\lambda_0^{[k]})x_2\right)<0. \tag{3.A.25}$$

Therefore, under both of these situations,

$$x_2\neq\Phi G^{[2]}(\lambda_0^{[k]})x_2. \tag{3.A.26}$$

Hence, the matrix-valued polynomial $M(\lambda)$ is of FCR at each $\lambda=\lambda_0^{[k]}$. This completes the proof. □

3.A.3 Proof of Theorem 3.9

To prove the condition for the necessity, assume that there exists a subsystem, denoted $\mathbf{\Sigma}_i$, such that the associated matrix pair $(A_{\mathbf{xx}}(i),\ [B_{\mathbf{x}}(i)\ \ A_{\mathbf{xv}}(i)])$ is not controllable. Then, according to Theorem 3.1, there exist at least one $\lambda_0\in\mathcal{C}$ and one nonzero vector $x_i\in\mathcal{C}^{m_{\mathbf{x}i}}$ such that

$$x_i^H\left[\lambda_0 I_{m_{\mathbf{x}i}}-A_{\mathbf{xx}}(i)\ \ B_{\mathbf{x}}(i)\ \ A_{\mathbf{xv}}(i)\right]=0. \tag{3.A.27}$$

Define the $M_{\mathbf{x}}$-dimensional vector $x=\mathbf{col}\{0_{M_{\mathbf{x},i-1}},\ x_i,\ 0_{M_{\mathbf{x}}-M_{\mathbf{x},i}}\}$. Then, $x\neq 0$. Moreover, from Eq. (3.A.27) and the block diagonal structure of the matrices $A_{\mathbf{xx}}$, $B_{\mathbf{x}}$, and $A_{\mathbf{xv}}$ direct matrix algebraic manipulations show that

$$x^H\left[\lambda_0 I_{M_{\mathbf{x}}}-A_{\mathbf{xx}}\ \ -B_{\mathbf{x}}\ \ -A_{\mathbf{xv}}\right]=0. \tag{3.A.28}$$

Note that

$$\left[\lambda_0 I_{M_{\mathbf{x}}}-A_{\mathbf{xx}}\ \ -B_{\mathbf{x}}\ \ -A_{\mathbf{xv}}\Phi\right]=\left[\lambda_0 I_{M_{\mathbf{x}}}-A_{\mathbf{xx}}\ \ -B_{\mathbf{x}}\ \ -A_{\mathbf{xv}}\right]\mathbf{diag}\left\{I_{m_{\mathbf{x}}},\ I_{m_{\mathbf{u}}},\ \Phi\right\}.$$

We therefore have that the matrix $[\lambda_0 I_{M_{\mathbf{x}}}-A_{\mathbf{xx}}\ -B_{\mathbf{x}}\ -A_{\mathbf{xv}}\Phi]$ can never be of FRR, no matter how the subsystem connection matrix Φ is designed. Hence it can be claimed further from the definition of the matrix-valued polynomial $\bar{M}(\lambda)$ that it is also never of FRR at $\lambda=\lambda_0$. According to Theorem 3.5, system $\mathbf{\Sigma}$ is not controllable.

To prove the condition for the sufficiency, note that the matrix-valued polynomials $\bar{M}^T(\lambda)$ and $M(\lambda)$ have completely the same form. Similar arguments as those for the derivations of Eqs. (3.A.14) and (3.A.15) in the proof of Theorem 3.8 show that, the matrix-valued polynomial $\bar{M}(\lambda)$ is of FRR at each complex number λ if and only if for each pair $(\lambda,\ x_2)$ satisfying

$$\bar{G}^{[1]}(\lambda)x_2 = 0 \tag{3.A.29}$$

where $\lambda \in \mathcal{C}$ and $x_2 \in \mathcal{C}^{M_{\mathbf{z}}}$, $x_2 \neq 0$, the following inequality is satisfied:

$$\left[I_{M_{\mathbf{z}}} - \Phi^T \bar{G}^{[2]}(\lambda)\right] x_2 \neq 0. \tag{3.A.30}$$

From the assumption that the TFM $\bar{G}^{[1]}(\lambda)$ is of FCNR, its block diagonal structure, and from the definitions of the matrices $\bar{Y}_s^{[k]}|_{s=1}^{\bar{s}^{[k]}}$ it can be straightforwardly shown that every λ satisfying Eq. (3.A.29) must be a transmission zero of the TFM $\bar{G}^{[1]}(\lambda)$. Moreover, all the nonzero x_2 satisfying Eq. (3.A.29) with $\lambda = \bar{\lambda}_0^{[k]}$ can be expressed as

$$x_2 = \bar{Y}^{[k]}\alpha, \tag{3.A.31}$$

where α is a nonzero $\sum_{s=1}^{\bar{s}^{[k]}} \bar{p}(k,s)$-dimensional complex vector, and

$$\bar{Y}^{[k]} = \begin{bmatrix} 0_{M_{\mathbf{z},\bar{k}(1)-1}\times \bar{p}(k,1)} & \cdots & 0_{M_{\mathbf{z},\bar{k}(\bar{s}^{[k]})-1}\times \bar{p}(k,\bar{s}^{[k]})} \\ \bar{Y}_1^{[k]} & \cdots & \bar{Y}_{\bar{s}^{[k]}}^{[k]} \\ 0_{(M_{\mathbf{z}}-M_{\mathbf{z},\bar{k}(1)})\times \bar{p}(k,1)} & \cdots & 0_{(M_{\mathbf{z}}-M_{\mathbf{z},\bar{k}(\bar{s}^{[k]})})\times \bar{p}(k,\bar{s}^{[k]})} \end{bmatrix}.$$

On the other hand, from Eq. (3.44) and singular value decompositions for a matrix [8] it can be declared that there exist $U_1 \in \mathcal{R}^{M_{\mathbf{v}}\times M_{\mathbf{z}}}$ and $U_2 \in \mathcal{R}^{M_{\mathbf{v}}\times(M_{\mathbf{v}}-M_{\mathbf{z}})}$ such that

$$\Phi = U_1\Theta, \quad [U_1\ U_2]^T[U_1\ U_2] = [U_1\ U_2][U_1\ U_2]^T = I_{M_{\mathbf{v}}}. \tag{3.A.32}$$

Hence, for each x_2 satisfying Eq. (3.A.31), we have that

$$\left[I_{M_{\mathbf{z}}} - \Phi^T \bar{G}^{[2]}(\lambda_0^{[k]})\right] x_2 = \Theta\left[\Theta^{-1}\bar{Y}^{[k]} - U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\right]\alpha, \tag{3.A.33}$$

which means that $\left[I_{M_{\mathbf{z}}} - \Phi^T \bar{G}^{[2]}(\lambda_0^{[k]})\right] x_2 \neq 0$ if and only if

$$\left[\Theta^{-1}\bar{Y}^{[k]} - U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\right]\alpha \neq 0. \tag{3.A.34}$$

Note that

$$\left|\left|\Theta^{-1}\bar{Y}^{[k]}\alpha\right|\right|_2^2 = \alpha^H \mathbf{diag}\left\{\left.\bar{Y}_s^{[k]H}\Theta^{-2}(\bar{k}(s))\bar{Y}_s^{[k]}\right|_{s=1}^{\bar{s}^{[k]}}\right\}\alpha. \tag{3.A.35}$$

Moreover, from Eq. (3.A.32) we have that $U_1U_1^T = I_{M_{\mathbf{v}}} - U_2U_2^T \leq I_{M_{\mathbf{v}}}$. Hence,

$$\begin{aligned} \left|\left|U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\alpha\right|\right|_2^2 &= \alpha^H\bar{Y}^{[k]H}\bar{G}^{[2]}(\lambda_0^{[k]H})U_1U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\alpha \\ &\leq \alpha^H\bar{Y}^{[k]H}\bar{G}^{[2]H}(\lambda_0^{[k]})\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\alpha \\ &= \alpha^H\mathbf{diag}\left\{\bar{Y}_s^{[k]H}\bar{G}_{\bar{k}(s)}^{[2]H}(\lambda_0^{[k]})\bar{G}_{\bar{k}(s)}^{[2]}(\lambda_0^{[k]})\bar{Y}_s^{[k]}\Big|_{s=1}^{\bar{s}^{[k]}}\right\}\alpha, \end{aligned} \tag{3.A.36}$$

which further leads to

$$\begin{aligned} &\left|\left|\Theta^{-1}\bar{Y}^{[k]}\alpha\right|\right|_2^2 - \left|\left|U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\alpha\right|\right|_2^2 \\ \geq\; & \alpha^H\mathbf{diag}\left\{\left(\bar{Y}_s^{[k]H}\Theta^{-2}(\bar{k}(s))\bar{Y}_s^{[k]} - \bar{Y}_s^{[k]H}\bar{G}_{\bar{k}(s)}^{[2]H}(\lambda_0^{[k]})\bar{G}_{\bar{k}(s)}^{[2]}(\lambda_0^{[k]})\bar{Y}_s^{[k]}\right)\Big|_{s=1}^{\bar{s}^{[k]}}\right\}\alpha \\ =\; & \hat{\alpha}^H\mathbf{diag}\left\{\left(I_{\bar{p}(k,s)} - \bar{\Gamma}_s^{[k]H}\bar{\Gamma}_s^{[k]}\right)\Big|_{s=1}^{\bar{s}^{[k]}}\right\}\hat{\alpha}, \end{aligned} \tag{3.A.37}$$

where $\hat{\alpha} = \mathbf{diag}\{(\bar{Y}_s^{[k]H}\Theta^{-2}(\bar{k}(s))\bar{Y}_s^{[k]})^{1/2}|_{s=1}^{\bar{s}^{[k]}}\}\alpha$.

Note that the matrix $\bar{Y}_s^{[k]H}\Theta^{-2}(\bar{k}(s))\bar{Y}_s^{[k]}$ is invertible for each feasible integer pair $(k,\ s)$. It is obvious that the vector α is nonzero if and only if the vector $\hat{\alpha}$ is. Therefore, if the condition of Eq. (3.54) is satisfied, then for any nonzero $\sum_{s=1}^{\bar{s}^{[k]}}\bar{p}(k,s)$-dimensional complex vector α, we have that

$$\left|\left|\Theta^{-1}\bar{Y}^{[k]}\alpha\right|\right|_2^2 - \left|\left|U_1^T\bar{G}^{[2]}(\lambda_0^{[k]})\bar{Y}^{[k]}\alpha\right|\right|_2^2 > 0. \tag{3.A.38}$$

Hence, the condition of Eq. (3.A.34) is satisfied, which means that the system $\mathbf{\Sigma}$ is controllable. This completes the proof. □

3.A.4 *Proof of Theorem 3.10*

From Lemma 3.3, we have that to guarantee the observability of the networked system $\mathbf{\Sigma}$, it is necessary that for each $i = 1, 2, \cdots, N$, the matrix pair $(A_{\mathbf{xx}}(i), [C_{\mathbf{T}}^T(i)\ A_{\mathbf{ST}}^T(i)]^T)$ is observable. It can therefore be declared from Corollary 3.2 that to construct an observable $\mathbf{\Sigma}$, it is necessary that $m_{\mathbf{y}i} + m_{\mathbf{z}i} \geq p_{\max}(i)$.

Now, assume that $m_{\mathbf{y}i} + m_{\mathbf{z}i} = p_{\max}(i)$ for every $1 \leq i \leq N$. Then, according to Theorem 3.6 and the duality between system controllability and system observability, there always exist a matrix $C_{\mathbf{T}}(i)$ and a matrix $A_{\mathbf{ST}}(i)$ for each $i \in \{1, 2, \cdots, N\}$ such that the matrix pair $(A_{\mathbf{xx}}(i), [C_{\mathbf{T}}^T(i)\ A_{\mathbf{ST}}^T(i)]^T)$ is observable.

Note that, for an arbitrary real number κ_i, we have that

$$\begin{bmatrix} \lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{T}}(i) \\ \kappa_i A_{\mathbf{ST}}(i) \end{bmatrix} = \mathbf{diag}\left\{I_{m_{\mathbf{x}i}},\ I_{m_{\mathbf{y}i}},\ \kappa_i I_{m_{\mathbf{z}i}}\right\} \begin{bmatrix} \lambda I_{m_{\mathbf{x}i}} - A_{\mathbf{xx}}(i) \\ C_{\mathbf{T}}(i) \\ A_{\mathbf{ST}}(i) \end{bmatrix}.$$

It is clear from Theorem 3.2 that observability of the matrix pair $(A_{\mathbf{xx}}(i), \mathbf{col}\{C_{\mathbf{x}}(i), \kappa_i A_{\mathbf{zx}}(i)\})$ is equivalent to that of the matrix pair $(A_{\mathbf{xx}}(i), \mathbf{col}\{C_{\mathbf{x}}(i),\ A_{\mathbf{zx}}(i)\})$, provided that $\kappa_i \neq 0$.

For each $j \in \{\,1,\ 2,\ \cdots,\ N\,\}$, define the set

$$\mathcal{J}(j) = \left\{ (k,s) \,\middle|\, k(s) = j,\ \begin{array}{l} s \in \{\,1,\ 2,\ \cdots,\ s^{[k]}\,\} \\ k \in \{\,1,\ 2,\ \cdots,\ m\,\} \end{array} \right\}.$$

This set is associated with all the transmission zeros of the TFM $G^{[1]}(\lambda)$, that is, also a transmission zero of the TFM $G_j^{[1]}(\lambda)$ with $j \in \{\,1,\ 2,\ \cdots,\ N\,\}$. Then, obviously, the satisfaction of Eq. (3.51) can be equivalently expressed as the satisfaction, for each $j = 1, 2, \cdots, N$, of the inequality

$$I_{p(k,s)} - \Gamma_s^{[k]H}\Theta^2(j)\Gamma_s^{[k]} > 0 \tag{3.A.39}$$

for every pair (k,s) of the set $\mathcal{J}(j)$.

For a fixed subsystem connection matrix Φ, define

$$\gamma_i = \max\left\{\sigma_{max}\left(\Theta(i)A_{\mathbf{SS}}(i)\right),\ \max_{(k,s)\in\mathcal{J}(i)} \sigma_{max}\left(\Theta(i)\Gamma_s^{[k]}\right)\right\}, \tag{3.A.40}$$

where $\sigma_{\max}(\cdot)$ stands for the maximal singular value of a matrix. Moreover, for each subsystem of system $\mathbf{\Sigma}$, define the matrices

$$\hat{A}_{\mathbf{ST}}(i) = \kappa_i A_{\mathbf{ST}}(i) \quad \text{and} \quad \hat{A}_{\mathbf{SS}}(i) = \kappa_i A_{\mathbf{SS}}(i), \tag{3.A.41}$$

where κ_i is an arbitrary number belonging to $(0,\ 1/\gamma_i)$.

Using these two matrices, construct a new networked system $\hat{\mathbf{\Sigma}}$ through simply replacing the system matrices $A_{\mathbf{zx}}(i)$ and $A_{\mathbf{zv}}(i)$ respectively by $\hat{A}_{\mathbf{ST}}(i)$ and $\hat{A}_{\mathbf{SS}}(i)$ keeping the other system matrices unchanged. Moreover, define the matrices $\hat{A}_{\mathbf{SS}}$, $\hat{A}_{\mathbf{ST}}$, and so on and the TFMs $\hat{G}^{[1]}(\lambda)$, $\hat{G}^{[2]}(\lambda)$, and so on, respectively as their counterparts associated with system $\mathbf{\Sigma}$.

Based on the block diagonal structure of the matrix $\hat{A}_{\mathbf{SS}}$ and Eq. (3.44), it can be straightforwardly proven that $(\Phi\hat{A}_{\mathbf{SS}})^T(\Phi\hat{A}_{\mathbf{SS}}) = \mathbf{diag}\{\kappa_i^2 A_{\mathbf{SS}}^T(i)\Theta^2(i)A_{\mathbf{zv}}(i)|_{i=1}^N\}$. Hence it can be claimed from Eqs. (3.A.40) and (3.A.41) that

$$\sigma_{\max}\left(\Phi\hat{A}_{\mathbf{SS}}\right) \;=\; \max_{1\le i\le N}\left\{\sigma_{\max}\left(\Theta(i)\hat{A}_{\mathbf{SS}}(i)\right)\right\}$$

$$\begin{aligned} &= \max_{1\le i\le N} \{\kappa_i \times \sigma_{\max}(\Theta(i) A_{\mathbf{SS}}(i))\} \\ &< 1. \end{aligned} \tag{3.A.42}$$

Note that the absolute value of each eigenvalue of a square matrix is not greater than its maximal singular value [8]. It can therefore be declared that the matrix $I - \Phi\hat{A}_{\mathbf{SS}}$ is invertible, and hence the reconstructed networked system $\hat{\mathbf{\Sigma}}$ is well-posed.

On the other hand, note that in system $\hat{\mathbf{\Sigma}}$, only the matrices $\hat{A}_{\mathbf{ST}}(i)$ and $\hat{A}_{\mathbf{SS}}(i)$ are different from those of system $\mathbf{\Sigma}$. This implies that the TFMs $G^{[1]}(\lambda)$ and $\hat{G}^{[1]}(\lambda)$, their transmission zeros, and the associated matrices $Y_s^{[k]}$ are completely the same. It can therefore be declared from the definition of the matrix $\Gamma_s^{[k]}$ that for each integer pair $(k,\ s)$ with $k \in \{\,1,\ 2,\ \cdots,\ m\,\}$ and $s \in \{\,1,\ 2,\ \cdots,\ s^{[k]}\,\}$, there certainly exists a unique $j \in \{\,1,\ 2,\ \cdots,\ N\,\}$ such that their pair (k, s) belongs to the set $\mathcal{J}(j)$. This further leads to

$$\hat{\Gamma}_s^{[k]} = \kappa_j \Gamma_s^{[k]}. \tag{3.A.43}$$

Hence, we have from Eqs. (3.A.40) and (3.A.41) that

$$\sigma_{max}\left(\hat{\Gamma}_s^{[k]}\Theta(j)\right) = \kappa_j \sigma_{max}\left(\Gamma_s^{[k]}\Theta(j)\right) < 1, \tag{3.A.44}$$

which further implies the satisfaction of the condition of Eq. (3.A.39) for each element of the set $\mathcal{J}(j)$ and each $j \in \{\,1,\ 2,\ \cdots,\ N\,\}$, and hence the system $\hat{\mathbf{\Sigma}}$ is observable.

The results on minimal input selection for system controllability can be established directly using duality between controllability and observability of a dynamic system, as well as the sufficient condition of Theorem 3.9.

This completes the proof. □

References

[1] T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[2] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[3] R.E. Kalman, Canonical structure of linear dynamical systems, Proceedings of the National Academy of Science, USA 46 (1962) 596–600.

[4] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[5] T. Zhou, Minimal inputs/outputs for a networked system, IEEE Control Systems Letters 1 (2017) 298–303.

[6] Y. Zhang, T. Zhou, Controllability analysis for a networked dynamic system with autonomous subsystems, IEEE Transactions on Automatic Control 62 (2017) 3408–3415.

[7] M. de Wal, B. Jager, A review of methods for input/output selection, Automatica 37 (2001) 487–510.

[8] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[9] T. Zhou, On the controllability and observability of networked dynamic systems, Automatica 52 (2015) 63–75.

[10] S. Pequito, S. Kar, A.P. Aguiar, A framework for structural input/output and control configuration selection in large-scale systems, IEEE Transactions on Automatic Control 61 (2016) 303–318.
[11] T.H. Summers, F.L. Cortesi, J. Lygeros, On submodularity and controllability in complex dynamical networks, IEEE Transactions on Control of Network Systems 3 (2016) 91–101.
[12] M. Egerstedt, S. Martini, M. Cao, K. Camlibel, A. Bicchi, Interacting with networks, how does structure relate to controllability in single-leader, consensus networks?, IEEE Control Systems Magazine 32 (2012) 66–73.
[13] Y.Y. Liu, J.J. Slotine, A.L. Barabasi, Controllability of complex networks, Nature 473 (2011) 167–173.
[14] C.T. Lin, Structural controllability, IEEE Transactions on Automatic Control 19 (1974) 201–208.
[15] J.M. Dion, C. Commault, J. der Woude, Generic properties and control of linear structured systems: a survey, Automatica 39 (2003) 1125–1144.
[16] F. Pasqualetti, S. Zampieri, F. Bullo, Controllability metrics, limitations and algorithms for complex networks, IEEE Transactions on Control of Network Systems 1 (2014) 40–52.
[17] V. Tzoumas, M.A. Rahimian, G.J. Pappas, A. Jadbabaie, Minimal actuator placement with bounds on control effort, arXiv:1409.3289v5 [math.OC], 2016.
[18] D.D. Siljak, Large-Scale Dynamic Systems: Stability and Structure, North-Holland Books, New York, USA, 1978.
[19] J. Schuppen, O. Boutin, P.L. Kempker, J. Komenda, T. Masopust, N. Pambakian, A.C.M. Ran, Control of distributed systems: tutorial and overview, European Journal of Control 17 (2011) 579–602.
[20] M.S. Andersen, S.K. Pakazad, A. Hansson, A. Rantzer, Robust stability of sparsely interconnected uncertain systems, IEEE Transactions on Automatic Control 59 (2014) 2151–2156.
[21] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Review 51 (2009) 661–703.
[22] T. Zhou, Y. Zhang, On the stability and robust stability of networked dynamic systems, IEEE Transactions on Automatic Control 61 (2016) 1595–1600.
[23] T. Zhou, Coordinated one-step optimal distributed state prediction for a networked dynamical system, IEEE Transactions on Automatic Control 58 (2013) 2756–2771.
[24] Y. Zhang, T. Zhou, A reinvestigation on the controllability and observability of networked dynamic systems, in: Proceedings of the 34th Chinese Control Conference, Hanzhou, Zhejiang Province, China, pp. 6740–6746.
[25] T. Zhou, On the controllability of networked dynamic systems with bounded inputs, in: Proceedings of the 2015 American Control Conference, Chicago, Illinois, USA, pp. 3404–3409.
[26] T. Zhou, Controllability of a networked system with input and state constraints, Science China: Mathematics 46 (2016) 1603–1616 (in Chinese).
[27] M. de Wal, B. Jager, A review of methods for input/output selection, Automatica 37 (2001) 487–510.
[28] E.D. Sontag, Mathematical Control Theory: Deterministic Finite Dimensional Systems, second edition, Springer-Verlag, New York, Inc., New York, USA, 1998.
[29] E.D. Sontag, An algebraic approach to bounded controllability of linear systems, International Journal of Control 39 (1984) 181–188.
[30] W.P.M.H. Heemels, M.K. Camlibel, Controllability of linear systems with input and state constraints, in: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, Louisiana, USA, pp. 536–541.
[31] M. Helwa, P. Caines, In-block controllability of affine systems on polytopes, in: Proceedings of the 53th IEEE Conference on Decision and Control, Los Angles, California, USA, pp. 3936–3942.
[32] J.D. Simon, S.K. Mitter, A theory of modal control, Information and Control 13 (1968) 316–353.
[33] Z.Z. Yuan, C. Zhao, Z.R. Du, W.X. Weng, Y.C. Lai, Exact controllability of complex networks, Nature Communications (2013), https://doi.org/10.1038/ncomms3447.
[34] T. Zhou, Relations among out-degree, controllability and observability of a networked system, arXiv:1610.02192v1 [math.OC], 2016.

CHAPTER 4

Kalman Filtering and Robust Estimation

4.1 Introduction

In many engineering systems, biological systems, social systems, and so on, there exist variables that cannot be directly measured. A key issue here is therefore the possibilities of estimating these variables from measurements of accessible variables and how to estimate these variables when these possibilities arise. The study of these problems can be traced back to Gauss's invention of the now most widely known least squares method. The theory of probability and statistics, linear algebra, a so on have now provided solid mathematical foundations for many estimation methods [1]. What distinguishes a particular estimation method for a dynamical system from the general results on estimations in probability and statistics is that it is developed for a special kind of problems that have some particular structure. More precisely, the output of a linear dynamical system depends on its input in a particular way, called convolution.

It is extensively accepted that a systematic investigation on state estimation for a dynamical system was initialized by R. Kalman, D.G. Luenberger et al. and started in the 1960s when the so-called modern control theory began to emerge [2–4]. In this theory, rather than a differential equation or a transfer function, the input–output relation of a linear dynamical system is described by a set of coupled first-order differential equations, which is now usually called a state space model.

In this chapter, we first introduce the Luenberger observer and discuss its design procedure. Afterward, we consider how to deal with measurement errors in state estimations, which naturally leads to the Kalman filter. A derivation using likelihood maximization is provided. Finally, we study the problem of state estimations with parametric modeling errors and develop a procedure that recursively estimates the states of a plant and is robust against these model errors.

4.2 State Estimation and Observer Design

In a finite-dimensional discrete-time system, assume that its outputs depend linearly on its inputs and its parameters are time invariant. For such a system, its input–output relation can be described by a set of first-order difference equations. In particular, let $y(k)$ and $u(k)$ represent

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00004-X

respectively the plant p- and q-dimensional output and input vectors. Then, the following two equations are extensively used to describe its dynamic properties:

$$x(k+1) = Ax(k) + Bu(k), \tag{4.1a}$$

$$y(k) = Cx(k) + Du(k), \tag{4.1b}$$

where $x(k)$ is called the plant state vector, which is assumed to be an n-dimensional column vector throughout this chapter. The first equation is often called the plant state transition equation, whereas the second one is called the plant output equation. The matrices A, B, C, and D are usually real and have compatible dimensions.

As its name indicates, the problem of state estimation is to estimate the plant state vector $x(k)$ at each time instant k using its currently available input vectors $u(s)$ and output vectors $y(s)$, $s = 0, 1, 2, \ldots, k$.

When the system matrices A, B, C, and D are known, one of the well-known state estimators is the Luenberger observer [3]. In this state estimator, a matrix L is sought such that the state vector $\hat{x}(t)$ of the dynamic system

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) - L[\hat{y}(k) - y(k)], \tag{4.2a}$$

$$\hat{y}(k) = C\hat{x}(k) + Du(k) \tag{4.2b}$$

satisfies

$$\lim_{k\to\infty} ||\hat{x}(k) - x(k)||_2 = 0 \tag{4.3}$$

for arbitrary initial values $x(0)$ and $\hat{x}(0)$ of the plant and observer state vectors.

The structure of the Luenberger observer is given in Fig. 4.1. Clearly, except the term related to their input vectors, the state space models of the observer and the plant have the same structure. Conceptually, the “extra” term in the observer can be regarded as a feedback based on the prediction error about the plant output vector that forces the state vector of the observer to be equal to that of the plant.

Denote $\hat{x}(k) - x(k)$ by $e(k)$, which is in fact the estimation error of the estimator at the time instant k. From Eqs. (4.1a) and (4.2a) it is straightforward to prove that

$$e(k+1) = (A - LC)e(k). \tag{4.4}$$

Clearly, to satisfy the requirement (4.3), it is necessary and sufficient that the dynamic system of equations (4.4) is stable. This implies that in the design of the Luenberger observer, we should find an appropriate matrix L such that all the eigenvalues of the matrix $A - LC$ have the magnitude smaller than 1. In other words, to move all the eigenvalues of the matrix

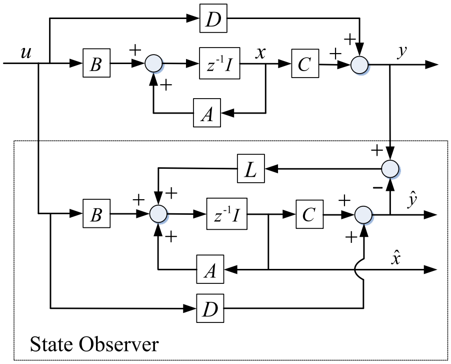

Figure 4.1: The Luenberger observer.

A into the interior of the unit disk on the complex plane through the feedback matrix L and the measurement matrix C.

A standard result on linear systems is that for a prescribed matrix pair A and C, the existence of a matrix L such that the magnitude of each eigenvalue of the matrix $A - LC$ is bounded by 1 is equivalent to the detectability of the plant. The latter can be simply verified through many available criteria, such as the PBH test, the rank condition, and so on [1].

This state estimator is in fact a one-step ahead predictor; noting that rather than the measurement data $y(k+1)$, it is $y(k)$ that is utilized in estimating the state vector $x(k+1)$. In other words, in the estimation of this vector, only $y(0)$, $y(1)$, $\ldots$, $y(k)$ have been used. To use the information contained in the measurement $y(k+1)$ in this estimation, it is necessary to modify the state space model of the observer given by Eq. (4.3) into the following form:

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) - L[\hat{y}(k+1) - y(k+1)], \tag{4.5a}$$

$$\hat{y}(k) = C\hat{x}(k) + Du(k), \tag{4.5b}$$

where $\hat{y}(k+1)$ is an estimate of $y(k+1)$ on the basis of the measurements $y(0)$, $y(1)$, $\cdots$, $y(k)$ and the input sequence $u(0)$, $u(1)$, $\cdots$, $u(k)$, $u(k+1)$. Generally, it can be calculated from $\hat{x}(k)$ using $\hat{y}(k+1) = C(k+1)[A(k)\hat{x}(k) + B(k)u(k)] + D(k+1)u(k+1)$. Here $A(k)\hat{x}(k) + B(k)u(k)$ is adopted to perform a one-step ahead prediction on $x(k+1)$.

Once again, let $e(k)$ denote the state estimation error. Then, when the above estimate of $y(k+1)$ is utilized, we have that

$$\begin{aligned} e(k+1) &= \hat{x}(k+1) - x(k+1) \\ &= \left\{A\hat{x}(k) + Bu(k) - L[\hat{y}(k+1) - y(k+1)]\right\} - [A(k)x(k) + B(k)u(k)] \\ &= A[\hat{x}(k) - x(k)] - L\left(\left\{C\left[A\hat{x}(k) + Bu(k)\right] + Du(k+1)\right\}\right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \{C\,[Ax(k)+Bu(k)] + Du(k+1)\}) \\
&= (A - LCA)[\hat{x}(k) - x(k)] \\
&= (I - LC)Ae(k). \qquad (4.6)
\end{aligned}
$$

Clearly, to make the state estimate asymptotically converge to its actual value, it is necessary and sufficient that the matrix $(I - LC)A$ is stable. The existence of a matrix L such that the matrix $(I - LC)A$ is stable is guaranteed by the detectability of the matrix pair $(A,\ CA)$. In fact, detectability of this matrix pair is a necessary and sufficient condition for the existence of the desirable L matrix.

In addition to the above full-order observer, there are also reduced-order observer in which the dimension of the state vector of the observer is strictly smaller than that of the plant. Such an observer can be designed through some modifications of the above procedure. An interested reader may refer to [1] for details.

4.3 Kalman Filter as a Maximum Likelihood Estimator

When the matrix pair $(A,\ CA)$ is detectable, there are infinitely many matrices L that make the matrix $(I - LC)A$ become a Horwitz or stable matrix, that is, all its eigenvalues have a magnitude smaller than 1. A natural question is that, among these matrices, which one is optimal. When external disturbances are Gaussian and optimality is measured by the covariance matrix of estimation errors, the answer is the Kalman filter, which is valid even when the plant is time varying.

In particular, for a linear time-varying finite-dimensional plant, assume that its dynamics is described by the following state space model:

$$x(k+1) = A(k)x(k) + B(k)u(k) + G(k)w(k), \qquad (4.7a)$$

$$y(k) = C(k)x(k) + D(k)u(k) + v(k). \qquad (4.7b)$$

Once again, $x(k)$ represents the plant n-dimensional state vector, $y(k)$ and $u(k)$ respectively represent the plant p- and q-dimensional output and input vectors, and $w(k)$ and $v(k)$ are respectively the process noise and measurement error vectors. To simplify expressions, we assume that the process noise vector $w(k)$ and the measurement error vector $v(k)$ are independent of each other and are also independent of all the plant state vectors before or at that time instant, that is, $x(s)$ with $0 \le s \le k$. We also assume that these vectors are white, that is, the values of these two vectors at different time instants are independent of themselves. Moreover, assume that the mathematical expectations of these two random processes are constantly equal to 0 and that their covariance matrices are respectively $Q(k)$ and $R(k)$. It is further assumed for brevity that the covariance matrices $Q(k)$ and $R(k)$ are invertible at each time instant k.

In the following discussions, when the plant is time invariant, which means that both the parameter matrices of the plant, as well as the statistic properties of the process noise and measurement error vectors, do not depend on the temporal variable k, the matrices $A(k)$, $B(k)$, and so on are usually simplified to A, B, and so on, respectively.

When the system matrices $A(k)$, $B(k)$, $C(k)$, $D(k)$, and $G(k)$ are known functions of the sampling time variable k, assume that the plant initial state vector $x(0)$ is normally distributed with mathematical expectation $\bar{x}(0)$ and covariance matrix $P(0)$. Then, the optimal state estimator, which is now extensively known as the Kalman filter, has completely the same structure as that of the Luenberger observer given in Eqs. (4.5a) and (4.5b). More precisely, denote the state vector of the estimator by $\hat{x}(k)$. Then, from an arbitrary estimate about the plant initial state vector, denote it by $\hat{x}(0)$, the optimal estimate of the plant state vector at the time instant $k+1$, on the basis of the measured plant output vector $y(s)$ with $0 \le s \le k+1$, which is now represented by $\hat{x}(k+1)$, can be recursively computed as follows:

$$\hat{x}(k+1) = A(k)\hat{x}(k) + B(k)u(k) - L(k+1)[\hat{y}(k+1) - y(k+1)], \tag{4.8a}$$

$$\hat{y}(k) = C(k)\hat{x}(k) + D(k)u(k), \tag{4.8b}$$

where $\hat{y}(k+1) = C(k+1)\left[A(k)\hat{x}(k) + B(k)u(k)\right] + D(k+1)u(k+1)$. Moreover, the matrix-valued function $L(k)$ is called the gain matrix of the Kalman filter, which has the following explicit and recursive computation formula:

$$P_{-}(k+1) = A(k)P(k)A^{T}(k) + G(k)Q(k)G^{T}(k), \tag{4.9a}$$

$$P(k+1) = \left[P_{-}^{-1}(k+1) + C^{T}(k+1)R^{-1}(k+1)C(k+1)\right]^{-1}, \tag{4.9b}$$

$$L(k+1) = P(k+1)C^{T}(k+1)R^{-1}(k+1). \tag{4.9c}$$

Recall that when a linear system is subject to an input signal that can be expressed as a linear combination of several other input signals, its output signal is simply the same linear combination of the output signals due to each individual input signal. The above state estimator can be directly modified to situations in which the system is simultaneously affected by deterministic and stochastic signals [1].

This estimator can be extended to situations in which the process noises and measurement errors are not white, the process noises and measurement errors are correlated, and so on. It has also been proved that as long as the plant is linear and the process noise, measurement error, and the plant initial state vector are normally distributed, the above estimator is optimal among all estimation procedures in the sense that the covariance matrix of the estimation error vector is minimal at each time instant.

The above procedure can be derived through various methods. For example, the geometric formulation and projection method adopted in Kalman's original work, the method with Markovian representations, dynamic programming-based method, invariant embedding-based method, method using deterministic least squares estimations, and method based on likelihood maximization [5]. In this book, we adopt the likelihood maximization-based method, which is more convenient in extending the Kalman filter to situations in which there exist parametric errors in the plant state space model.

4.3.1 Derivation of the Kalman Filter

Note that a linear combination of normally distributed random variables is still normally distributed. It is clear from the plant state space model of Eqs. (4.7a) and (4.7b) that when its initial state vector $x(0)$, the process noise vector $w(s)$, and the measurement error vector $v(s)$ are normally distributed, all the plant state vectors and its measured output vectors until a time instant, say, k, that is, $x(0),\ x(1),\ \cdots,\ x(k)$ and $y(0),\ y(1),\ \cdots,\ y(k)$ are jointly normally distributed. Hence, with the availability of the measurement vectors $y(s)|_{s=0}^{k}$, the optimal estimate on $x(k)$ is the conditional expectation $\mathbf{E}\left[x(k)\,|y(s),\ s=0,\ 1,\ \cdots,\ k\right]$. Recall that the conditional expectation of jointly normally distributed random variables still has a normal distribution [6]. Let $\hat{x}(k|k)$ represent the optimal estimate of the state vector $x(k)$ based on the measurements $y(0),\ y(1),\ \cdots$, and $y(k)$, and let $\hat{x}(k|k+1)$ be the one based on the measurements $y(0),\ y(1),\ \cdots$, and $y(k+1)$. Moreover, let $P(k)$ denote the covariance matrix of estimation errors of $\hat{x}(k|k)$. Assume that the estimate $\hat{x}(k|k)$ is unbiased and the covariance matrix $P(k)$ is positive definite. Then, from the normal distribution assumptions on the process noises $w(k)$ and the measurement errors $v(k)$ we have that the probability density functions (PDFs) of $v(k+1)$ and $w(k)$ are respectively

$$f_{v(k+1)}(v(k+1)) = \frac{1}{(2\pi\ \det(R(k+1)))^{p/2}} e^{-v^T(k+1)R^{-1}(k+1)v(k+1)}, \tag{4.10a}$$

$$f_{w(k)}(w(k)) = \frac{1}{(2\pi\ \det(Q(k)))^{q/2}} e^{-w^T(k)Q^{-1}(k)w(k)}. \tag{4.10b}$$

Moreover, denote the estimation error of the estimate $\hat{x}(k|k)$ by $e(k|k)$. Then, the state vector of the plant at the time instant k, that is, $x(k)$, can also be expressed as $x(k)=\hat{x}(k|k)-e(k|k)$. On the basis of the assumptions on the unbiasedness of the estimator and the normal distribution of its estimation errors, we have that the PDF of $x(k)$ can be expressed as

$$f_{x(k)}(x(k)) = \frac{1}{(2\pi\ \det(P(k)))^{n/2}} e^{-(x(k)-\hat{x}(k|k))^T P^{-1}(k)(x(k)-\hat{x}(k|k))}. \tag{4.10c}$$

On the other hand, from Eq. (4.7) we have that

$$\begin{aligned} v(k+1) &= y(k+1) - C(k+1)\left[A(k)x(k) + B(k)u(k) + G(k)w(k)\right] - D(k+1)u(k+1) \\ &= \bar{y}(k+1) - C(k+1)A(k)x(k) - C(k+1)G(k)w(k), \end{aligned} \tag{4.11}$$

where $\bar{y}(k+1)$ is defined as

$$\bar{y}(k+1) = y(k+1) - C(k+1)B(k)u(k) - D(k+1)u(k+1).$$

Recall that $w(k)$ and $v(k)$ are white and independent of each other and independent of the plant initial states. It is obvious from Eqs. (4.7a) and (4.7b) that $x(k)$, $w(k)$, and $v(k+1)$ are independent of each other. Hence, when the measurement data $y(k+1)$ become available, through substituting Eq. (4.11) into Eq. (4.10a) we have that the joint PDF of $x(k)$, $w(k)$, and $v(k+1)$ is equivalent to

$$\begin{aligned} & f_{x(k)}(x(k)) \times f_{w(k)}(w(k)) \times f_{v(k+1)}(v(k+1)) \\ = \; & Const \times \exp\Big\{-\Big[||w(k)||^2_{Q^{-1}(k)} + ||x(k) - \hat{x}(k|k)||^2_{P^{-1}(k)} \\ & + ||\bar{y}(k+1) - C(k+1)A(k)x(k) - C(k+1)G(k)w(k)||^2_{R^{-1}(k+1)}\Big]\Big\}, \end{aligned} \tag{4.12}$$

where *Const* is a constant that can be explicitly expressed as

$$Const = \left(\sqrt{2\pi}\right)^{p+q+n} \times \det{}^{n/2}(P(k)) \times \det{}^{q/2}(Q(k)) \times \det{}^{p/2}(R(k+1)).$$

Hence, the logarithm of the likelihood function of the random vectors $x(k)$ and $w(k)$ after the arrival of the measurement $y(k+1)$, denote it by $l(x(k), w(k))$, can be expressed as

$$\begin{aligned} l(x(k), w(k)) &= \log(Const) - ||w(k)||^2_{Q^{-1}(k)} - ||x(k) - \hat{x}(k|k)||^2_{P^{-1}(k)} \\ & \quad - ||\bar{y}(k+1) - C(k+1)A(k)x(k) - C(k+1)G(k)w(k)||^2_{R^{-1}(k+1)}. \end{aligned} \tag{4.13}$$

To simplify mathematical expressions, denote $\mathbf{col}\{x(k),\ w(k)\}$ and $\mathbf{col}\{\hat{x}(k|k),\ 0\}$ respectively by $\alpha(k)$ and $\alpha_0(k)$. Moreover, denote $C(k+1)\left[A(k)\hat{x}(k|k) + B(k)u(k)\right] + D(k+1)u(k+1)$ by $\hat{y}(k+1|k)$. Then, we have that

$$\begin{aligned} y(k+1) - \hat{y}(k+1|k) &= y(k+1) - \big\{C(k+1)\left[A(k)\hat{x}(k|k) + B(k)u(k)\right] \\ & \quad + D(k+1)u(k+1)\big\} \\ &= \bar{y}(k+1) - C(k+1)A(k)\hat{x}(k|k). \end{aligned} \tag{4.14}$$

Hence

$$\begin{aligned}
&\bar{y}(k+1) - C(k+1)A(k)x(k) - C(k+1)G(k)w(k) \\
=\ & [\bar{y}(k+1) - C(k+1)A(k)\hat{x}(k|k)] - [C(k+1)A(k)x(k) + C(k+1)G(k)w(k) \\
& - C(k+1)A(k)\hat{x}(k|k)] \\
=\ & y(k+1) - \hat{y}(k+1|k) - C(k+1)[A(k)\ \ G(k)][\alpha(k) - \alpha_0(k)]. \qquad (4.15)
\end{aligned}$$

Substituting these relations into Eq. (4.13), we can straightforwardly prove that

$$\begin{aligned}
l(x(k), w(k)) =\ & \log(Const) - ||\alpha(k) - \alpha_0(k)||^2_{\mathbf{diag}^{-1}(P(k),\ Q(k))} \\
& - ||y(k+1) - \hat{y}(k+1|k) - C(k+1)[A(k)\ \ G(k)][\alpha(k) - \alpha_0(k)]||^2_{R^{-1}(k+1)}. \qquad (4.16)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{dl(x(k), w(k))}{d\alpha(k)} \\
=\ & -2\Big\{\mathbf{diag}^{-1}(P(k),\ Q(k))[\alpha(k) - \alpha_0(k)] + (C(k+1)[A(k)\ \ G(k)])^T\, R^{-1}(k+1) \\
& \times \big\{C(k+1)[A(k)\ \ G(k)][\alpha(k) - \alpha_0(k)] - \big[y(k+1) - \hat{y}(k+1|k)\big]\big\}\Big\}. \qquad (4.17)
\end{aligned}$$

At the maximum likelihood estimate of $x(k)$ and $w(k)$, the likelihood function and therefore its logarithm achieve maximal values. Note that the likelihood function $l(x(k), w(k))$ is a strictly concave function of both vectors $x(k)$ and $w(k)$. This is equivalent to that, at this estimate,

$$\frac{dl(x(k), w(k))}{d\alpha(k)} = 0. \qquad (4.18)$$

Denote $\mathbf{col}\{\hat{x}(k|k+1),\ \hat{w}(k|k+1)\}$ by $\hat{\alpha}(k|k+1)$. Then Eqs. (4.17) and (4.18) imply that

$$\begin{aligned}
\hat{\alpha}(k|k+1) - \alpha_0(k) =\ & \Big\{\mathbf{diag}^{-1}(P(k),\ Q(k)) + (C(k+1)[A(k)\ \ G(k)])^T\, R^{-1}(k+1) \\
& \times (C(k+1)[A(k)\ \ G(k)])\Big\}^{-1} \\
& \times (C(k+1)[A(k)\ \ G(k)])^T\, R^{-1}(k+1)\big[y(k+1) - \hat{y}(k+1|k)\big]. \qquad (4.19)
\end{aligned}$$

From the well-known matrix formula $G(I + HG)^{-1} = (I + GH)^{-1}G$ we further have that

$$\begin{aligned}
&\hat{\alpha}(k|k+1) - \alpha_0(k) \\
=\ & \mathbf{diag}(P(k),\ Q(k))\Big\{I + [A(k)\ \ G(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\ \ G(k)]
\end{aligned}$$

$$
\begin{aligned}
&\times \mathbf{diag}(P(k),\ Q(k))\}^{-1}[A(k)\ \ G(k)]^T C^T(k+1)R^{-1}(k+1)\left[y(k+1)-\hat{y}(k+1|k)\right]\\
&= \mathbf{diag}(P(k),\ Q(k))[A(k)\ \ G(k)]^T \Big\{ I + C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\ \ G(k)]\\
&\quad \times \mathbf{diag}(P(k),\ Q(k))[A(k)\ \ G(k)]^T \Big\}^{-1} C^T(k+1)R^{-1}(k+1)\left[y(k+1)-\hat{y}(k+1|k)\right]\\
&= \begin{bmatrix} P(k)A^T(k) \\ Q(k) \end{bmatrix} \Big\{ I + C^T(k+1)R^{-1}(k+1)C(k+1)\\
&\quad \times \left[A(k)P(k)A^T(k) + G(k)Q(k)G^T(k)\right]\Big\}^{-1}\\
&\quad \times C^T(k+1)R^{-1}(k+1)\left[y(k+1)-\hat{y}(k+1|k)\right]\\
&= \begin{bmatrix} P(k)A^T(k) \\ Q(k)G^T(k) \end{bmatrix} P_-^{-1}(k+1)P(k+1)C^T(k+1)R^{-1}(k+1)\left[y(k+1)-\hat{y}(k+1|k)\right].
\end{aligned}
\tag{4.20}
$$

In these derivations, the definitions of the matrices $P_-(k+1)$ and $P(k+1)$ have been used, which are respectively given by Eqs. (4.9a) and (4.9b). Hence

$$
\begin{aligned}
\hat{x}(k+1|k+1) &= A(k)\hat{x}(k|k+1) + B(k)u(k) + G(k)\hat{w}(k|k+1)\\
&= [A(k)\ \ G(k)]\hat{\alpha}(k|k+1) + B(k)u(k)\\
&= [A(k)\ \ G(k)]\Bigg\{\begin{bmatrix} P(k)A^T(k) \\ Q(k)G^T(k) \end{bmatrix} P_-^{-1}(k+1)P(k+1)C^T(k+1)\\
&\qquad \times R^{-1}(k+1)\left[y(k+1)-\hat{y}(k+1|k)\right] + \alpha_0(k)\Bigg\} + B(k)u(k)\\
&= A(k)\hat{x}(k|k) + B(k)u(k) - P(k+1)C^T(k+1)R^{-1}(k+1)\\
&\quad \times \left[\hat{y}(k+1|k) - y(k+1)\right].
\end{aligned}
\tag{4.21}
$$

The last equality has completely the same form as that of Eq. (4.8a) if we replace $\hat{x}(k|k)$, $\hat{x}(k+1|k+1)$, and $\hat{y}(k|k+1)$ respectively by $\hat{x}(k)$, $\hat{x}(k+1)$, and $\hat{y}(k+1)$ and recall the definition of the Kalman gain matrix $L(k+1)$ given in Eq. (4.9c). These replacements are natural, since they are completely the same quantities from their definitions. We use different symbols in the derivations only to emphasize the available information based on which an estimate is performed.

In the derivations, it is assumed that the estimate $\hat{x}(k|k)$ is unbiased. Now, we discuss reasonability of this assumption. After this discussion, we will investigate engineering significance for the matrix $P(k+1)$.

Let $e(k)$ denote the state estimation error of the Kalman filter at the time instant k, that is, $e(k)=\hat{x}(k)-x(k)$. Then, from Eqs. (4.7a), (4.7b), and (4.8a) we have that

$$\begin{aligned}\hat{x}(k+1) &= A(k)\hat{x}(k)+B(k)u(k)-L(k+1)\Big\{\big(C(k+1)[A(k)\hat{x}(k)+B(k)u(k)] \\ &\quad +D(k+1)u(k+1)\big)-\big(C(k+1)[A(k)x(k)+B(k)u(k)+G(k)w(k)] \\ &\quad +D(k+1)u(k+1)\big)+D(k+1)u(k+1)+v(k+1)\big)\Big\} \\ &= A(k)\hat{x}(k)+B(k)u(k)-L(k+1)\left\{C(k+1)A(k)e(k)\right. \\ &\quad \left.-C(k+1)G(k)w(k)-v(k+1)\right\}. \end{aligned} \tag{4.22}$$

Combing this equation with Eq. (4.7a), we establish the following recursive formula for the estimation error $e(k)$:

$$\begin{aligned}e(k+1) &= \hat{x}(k+1)-x(k+1) \\ &= \Big\{A(k)\hat{x}(k)+B(k)u(k)-L(k+1)C(k+1)A(k)e(k) \\ &\qquad +L(k+1)C(k+1)G(k)w(k)+L(k+1)v(k+1)\Big\} \\ &\quad -[A(k)x(k)+B(k)u(k)+G(k)w(k)] \\ &= [I-L(k+1)C(k+1)]A(k)e(k)+[L(k+1)C(k+1)-I]G(k)w(k) \\ &\quad +L(k+1)v(k+1). \end{aligned} \tag{4.23}$$

Hence, from the assumptions on the process noise $w(k)$ and the measurement error $v(k)$ we can be declare that

$$\begin{aligned}\mathbf{E}[e(k+1)] &= [I-L(k+1)C(k+1)]A(k)\mathbf{E}[e(k)]+[L(k+1)C(k+1)-I]G(k)\mathbf{E}[w(k)] \\ &\quad +L(k+1)\mathbf{E}[v(k+1)] \\ &= [I-L(k+1)C(k+1)]A(k)\mathbf{E}[e(k)]. \end{aligned} \tag{4.24}$$

Therefore, if $\hat{x}(k)$ is an unbiased estimate, then the estimate $\hat{x}(k+1)$ obtained from the Kalman filter of Eqs. (4.8a) and (4.8b) is also unbiased. As the Kalman filter is a recursive estimator and Eq. (4.24) is valid for every time instant, it is obvious that the state estimate given by the Kalman filter is always unbiased, provided that the initial estimate, that is, the estimate on the plant state vector at $k=0$, is unbiased, which is usually not a very restrictive condition.

In addition, if the associated dynamic system, which can be described by $x(k+1)=[I-L(k+1)C(k+1)]A(k)x(k)$, is stable, then, even if an initial estimate on the plant state vector is biased, it is clear from Eq. (4.24) that this estimate is asymptotically unbiased, noting that

$$\lim_{k\to\infty}\mathbf{E}[e(k)]=0.$$

On the other hand, from the assumptions on $x(0)$, $w(k)$, and $v(k)$, it is clear that $e(k)$, $w(k)$ and $v(k+1)$ are independent of each other at every time instant $k \geq 0$. We can therefore declare from Eq. (4.23) that

$$\begin{aligned}
\mathbf{Cov}[e(k+1)] &= [I - L(k+1)C(k+1)]A(k)\mathbf{Cov}[e(k)]A^T(k)[I - L(k+1)C(k+1)]^T \\
&\quad + [I - L(k+1)C(k+1)]G(k)\mathbf{Cov}[w(k)]G^T(k)[I - L(k+1)C(k+1)]^T \\
&\quad + L(k+1)\mathbf{Cov}[v(k+1)]L^T(k+1) \\
&= [I - L(k+1)C(k+1)]A(k)P(k)A^T(k)[I - L(k+1)C(k+1)]^T \\
&\quad + [I - L(k+1)C(k+1)]G(k)Q(k)G^T(k)[I - L(k+1)C(k+1)]^T \\
&\quad + L(k+1)R(k+1)L^T(k+1) \\
&= [I - L(k+1)C(k+1)]P_-(k+1)[I - L(k+1)C(k+1)]^T \\
&\quad + L(k+1)R(k+1)L^T(k+1).
\end{aligned} \tag{4.25}$$

Note that

$$\begin{aligned}
I - L(k+1)C(k+1) &= I - \left[P_-^{-1}(k+1) + C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1} \\
&\quad \times C^T(k+1)R^{-1}(k+1)C(k+1) \\
&= I - \left[I + P_-(k+1)C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1} \\
&\quad \times P_-(k+1)C^T(k+1)R^{-1}(k+1)C(k+1) \\
&= \left[I + P_-(k+1)C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1} \\
&= P(k+1)P_-^{-1}(k+1).
\end{aligned} \tag{4.26}$$

Moreover, note that all the matrices $P_-(k+1)$, $P(k+1)$, $R(k+1)$ are symmetric from their definitions. We therefore have that

$$\begin{aligned}
\mathbf{Cov}[e(k+1)] &= \left[P(k+1)P_-^{-1}(k+1)\right]P_-(k+1)\left[P(k+1)P_-^{-1}(k+1)\right]^T \\
&\quad + \left[P(k+1)C^T(k+1)R^{-1}(k+1)\right]R(k+1) \\
&\quad \times \left[P(k+1)C^T(k+1)R^{-1}(k+1)\right]^T \\
&= P(k+1)\left[P_-^{-1}(k+1) + C^T(k+1)R^{-1}(k+1)C(k+1)\right]P(k+1) \\
&= P(k+1).
\end{aligned} \tag{4.27}$$

Thus the matrix $P(k+1)$ defined in Eq. (4.9c) of the Kalman filter is in fact the covariance matrix of its estimation errors at the time constant $k+1$, which is one of the most important indices in assessing the quality of an estimator. Hence, the estimation procedure of

Eqs. (4.8a)–(4.9c) provides not only a recursive estimate of the plant state vector, but also a recursive method for assessing its estimation quality.

Although we do not provide detailed discussions on state predictions here, it is worth mentioning that $A(k)\hat{x}(k) + B(k)u(k)$ is in fact the optimal one-step ahead prediction on the plant state vector at the time instant $k+1$, that is, $x(k+1)$, on the basis of $(u(s),\ y(s))|_{s=0}^{k}$, and the covariance matrix of its prediction error is equal to $P_{-}(k+1)$ [1,5]. From the definition of the matrix $P(k+1)$ given in Eq. (4.9b) it is obvious that

$$\begin{aligned} P^{-1}(k+1) &= P_{-}^{-1}(k+1) + C^{T}(k+1)R^{-1}(k+1)C(k+1) \\ &\geq P_{-}^{-1}(k+1). \end{aligned} \tag{4.28}$$

Then, we can declare from standard results in matrix theories [7–9] (some of them are included in Lemma 2.1) that

$$P(k+1) \leq P_{-}(k+1). \tag{4.29}$$

Hence, the quality of filtering in the Kalman filter is always not worse than its prediction. This is a clear engineering requirement, as filtering uses more plant input–output data, and therefore more information about the plant, than a predictor.

4.3.2 Convergence Property of the Kalman Filter

Due to its optimality, recursive implementability, and many other attractive properties, the Kalman filter derived in the previous subsection has been extensively applied in engineering systems, social systems, biological systems, and so on. In this subsection, we discuss one of its important properties, the asymptotic convergence to a constant gain observer. To avoid awkward discussions, we only investigate situations under which the system matrix $A(k)$ is always invertible. More general results can be found, for example, in [1]. For this purpose, define the matrix

$$\Phi(k) = \begin{bmatrix} A(k) & G(k)Q(k)G^{T}(k)A^{-T}(k) \\ C^{T}(k+1)R^{-1}(k+1)C(k+1)A(k) & [I + C^{T}(k+1)R^{-1}(k+1)C(k+1)G(k)Q(k)G^{T}(k)]A^{-T}(k) \end{bmatrix}. \tag{4.30}$$

Then, straightforward algebraic manipulations show that the relation between the matrices $P(k+1)$ and $P(k)$, which is given by Eqs. (4.9a) and (4.9b), can be rewritten as

$$P(k+1) = \mathbf{H}_{m}(\Phi(k),\ P(k)). \tag{4.31}$$

Moreover,

$$\Phi^T(k) J \Phi(k) = J \quad \text{with} \quad J = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & 0_{n\times n} \end{bmatrix}, \tag{4.32}$$

that is, the matrix $\Phi(k)$ is Hamiltonian.

From the cascade property of the homographic transformation given by Lemma 2.6 we obtain the following equality:

$$\begin{aligned} P(k) &= \mathbf{H}_m\left(\Phi(k),\ \mathbf{H}_m\left(\Phi(k-1),\ \cdots,\ \mathbf{H}_m\left(\Phi(1),\ P(0)\right)\cdots\right)\right) \\ &= \mathbf{H}_m(\Phi(k)\Phi(k-1),\ \mathbf{H}_m\left(\Phi(k-2),\ \cdots,\ \mathbf{H}_m\left(\Phi(1),\ P(0)\right)\cdots\right)) \\ &= \cdots \\ &= \mathbf{H}_m\left(\prod_{j=k}^{1} \Phi(j),\ P(0)\right), \end{aligned} \tag{4.33}$$

which provides a very simple relation between the covariance matrix of the estimation errors of the Kalman filter at any time instant k and that at the initial time instant $k = 0$. In fact, similar relations can be established for this covariance matrix between any two time instants. These relations are very helpful in investigating convergence properties of the Kalman filter.

On the basis of this relation and Definition 2.7, which gives a Riemannian distance between two positive definite matrices, the following asymptotic properties are established through some straightforward algebraic operations.

Theorem 4.1. *Assume that the plant described by Eq. (4.7) is time invariant and its state transition matrix is invertible. Then, when the matrix pair $(A,\ G)$ is controllable and the matrix pair $(A,\ C)$ is observable with the increment of the time instant k, the covariance matrix $P(k)$ of the Kalman filter given by Eqs. (4.8a) and (4.8b) converges to a constant matrix.*

A proof of these results is given in the appendix of this chapter.

When the matrix A is not invertible, and/or the plant is not controllable, and/or the plant is not observable, some results have also been established for the convergence of the covariance matrix $P(k)$ of the estimation errors of the Kalman filter. The derivations, however, are quite lengthy, although the essential ideas are still borrowed from linear algebra. An interested reader is recommended to refer to [1,5].

Note that a normal distribution is completely determined by its mathematical expectation and covariance matrix. It is clear from Theorem 4.1 that when the plant is time invariant, controllable, and observable, the estimation error of the Kalman filter converges to a stationary process. On the other hand, from Eq. (4.9c) it can also be declared that, under these conditions, the update gain matrix $L(k)$ of the Kalman filter also converges to a constant matrix. This result is quite attractive in engineering applications, as it implies that without sacrificing steady-state estimation accuracy, the Kalman filter can be replaced by a state observer with a constant update gain matrix, which can significantly reduce online computation burdens of the Kalman filter.

4.4 Recursive Robust State Estimation Through Sensitivity Penalization

Due to its simplicity and optimality, the Kalman filter is quite attractive in actual applications. A potential pitfall of applying the Kalman filter to real-world problems lies in its explicit dependence on the parameters of the plant state space model. In many real-world problems, plant parameters are estimated from experimental data, which unavoidably introduce errors into the estimated parameters. Moreover, a model usually can only capture major features of the plant dynamics, which means that there often exists unmodeled dynamics in the errors of a plant state space model. In addition, a plant may need to work in various distinctive environments that differ from each other significantly. Hence, an essential issue about the Kalman filter, which is in fact valid for any state estimation procedure, is about the degeneration of its estimation accuracy caused by modeling errors.

Several cases have been reported in which the Kalman filter failed to work very well [1,5]. To reduce the sensitivity of estimation accuracy of an estimator to modeling errors, various approaches have been suggested, such as the H_∞-norm-based method, the guaranteed cost-based method, the set-membership-based approach, and so on [10]. In this section, we introduce a sensitivity penalization-based approach. The basic idea here is to reduce the sensitivity of the cost function to modeling errors, which is adopted in the derivations of a state estimator and therefore increases the robustness in estimation accuracy of the resulting state estimator to modeling errors. A prominent characteristic of this robust state estimator is that it has a similar form as that of the Kalman filter and can be recursively implemented. Its computational complexity is comparable to that of the Kalman filter, and no extra conditions are required to be verified in its realizations. The last property significantly distinguishes it from other robust state estimators, which usually require a verification of some matrix-based conditions that cannot be performed online very easily in general.

4.4.1 Estimation Algorithm

When there exist parametric errors in a plant state space model, the input–output relation of the plant can generally be expressed as the following a modification of the state space model of Eqs. (4.7a) and (4.7b), which is extensively adopted to describe the dynamics of a linear time-varying finite-dimensional plant:

$$x(k+1) = A(k, \varepsilon(k))x(k) + B(k, \varepsilon(k))u(k) + G(k, \varepsilon(k))w(k), \tag{4.34a}$$

$$y(k) = C(k, \varepsilon(k))x(k) + D(k, \varepsilon(k))u(k) + v(k), \tag{4.34b}$$

where $\varepsilon(k)$ represent the deviations of the nominal values of the plant parameters from their actual values. The value of this vector may be time varying, but it is assumed that at every sampled time instant, all the system parameter matrices, that is, $A(k, \varepsilon(k))$, $B(k, \varepsilon(k))$, $C(k, \varepsilon(k))$, $D(k, \varepsilon(k))$, and $G(k, \varepsilon(k))$, are differentiable with respect to each of its elements. Moreover, its dimension, denoted by n_e, is assumed to be time invariant for avoiding complicated mathematical expressions.

Generally, there is a vector, say p_0, in the model that consists of the nominal values of all the physical parameters, chemical parameters, biological parameters, and so on in the system. For a concise presentation, this vector has been dropped out, but all the results must be understood as valid with system parameters varying in a neighborhood of the vector p_0, that is, each system parameter is implicitly assumed to be in an interval centered at its nominal value.

Similarly to the derivations of the Kalman filter, assume that from the plant output measurements $y(s)|_{s=0}^{k}$, an estimate about the plant state vector $x(k)$ has already been obtained. Denote it by $\hat{x}(k|k)$. Moreover, assume that there is a positive definite matrix $P(k)$ that is used to weight the difference between any other estimate on $x(k)$ and the available estimate.

Once again, let $\hat{x}(k|k+1)$ and $\hat{w}(k|k+1)$ represent respectively an estimate for the state vector $x(k)$ and the process noise $w(k)$ from the plant output measurements $y(s)|_{s=0}^{k+1}$. Moreover, let $e(k, \varepsilon(k), \varepsilon(k+1))$ denote the prediction error on the plant output $y(k+1)$ using these estimates, that is,

$$\begin{aligned} e(k, \varepsilon(k), \varepsilon(k+1)) \;=\;& y(k+1) - C(k+1, \varepsilon(k+1))\big[A(k, \varepsilon(k))\hat{x}(k|k+1) + B(k, \varepsilon(k))u(k) \\ &+ G(k, \varepsilon(k))\hat{w}(k|k+1)\big] - D(k+1, \varepsilon(k+1))u(k+1). \end{aligned}$$

To avoid lengthy mathematical expressions, denote

$$y(k+1) - C(k+1, \varepsilon(k+1))B(k, \varepsilon(k))u(k) - D(k+1, \varepsilon(k+1))u(k+1)$$

by $\bar{y}(k+1, \varepsilon(k), \varepsilon(k+1))$. Then $e(\varepsilon(k), \varepsilon(k+1))$ can be reexpressed as

$$\begin{aligned} e(k, \varepsilon(k), \varepsilon(k+1)) \;=\;& \bar{y}(k+1, \varepsilon(k), \varepsilon(k+1)) - C(k+1, \varepsilon(k+1))A(k, \varepsilon(k))\hat{x}(k|k+1) - \\ & C(k+1, \varepsilon(k+1))G(k, \varepsilon(k))\hat{w}(k|k+1). \end{aligned}$$

When there do not exist modeling errors in the plant state space model, $e(k, \varepsilon(k), \varepsilon(k+1))$ is usually called the innovation process in state estimations and plays an essential role in developing an estimation algorithm and in analyzing properties of an estimator. Loosely speaking, it represents new information about the plant state vector $x(k)$ carried in the plant output measurement $y(k+1)$ [1].

Denote the jth row element of the parametric error vector $\varepsilon(k)$ by $\varepsilon_j(k)$, $j = 1, 2, \ldots, n_e$. With a given nonnegative number $\lambda(k)$, define the cost function

$$
\begin{aligned}
& J(\hat{x}(k|k+1), \hat{w}(k|k+1)) \\
= \; & \frac{1}{2}\Bigg\{||\hat{w}(k|k+1)||^2_{Q^{-1}(k)} + ||\hat{x}(k|k+1) - \hat{x}(k|k)||^2_{P^{-1}(k)} + ||e(0,0)||^2_{R^{-1}(k+1)} \\
& + \lambda(k) \sum_{j=1}^{n_e} \left|\left| \frac{\partial e(k,0,0}{\partial \varepsilon_j(k)} \right|\right|_2^2 + \left|\left| \frac{\partial e(k,0,0)}{\partial \varepsilon_j(k+1)} \right|\right|_2^2 \Bigg\}
\end{aligned}
\tag{4.35}
$$

Comparing Eq. (4.35) with Eq. (4.13), it is clear that maximizing the cost function of Eq. (4.13) is equivalent to minimizing that of Eq. (4.35) with $\varepsilon(k) = \varepsilon(k+1) = 0$ and $\lambda(k) = 0$. Note also that $e(\varepsilon(k), \varepsilon(k+1))$ is the only factor in the cost function of Eq. (4.13) that may be affected by modeling errors. We can declare that through introducing the factors $\frac{\partial e(0,0)}{\partial \varepsilon_j(k)}$ and $\frac{\partial e(0,0)}{\partial \varepsilon_j(k+1)}$ into the cost function of Eq. (4.35), we can reduce influences of parametric modeling errors on the cost functions and therefore increase robustness of the resulted state estimator.

Note that for an arbitrary finite number $\lambda(k) \geq 0$, there always exists a scalar $\mu(k)$ belonging to $(0,\ 1]$ such that $\lambda(k) = \frac{1-\mu(k)}{\mu(k)}$. From this expression of the penalization factor $\lambda(k)$ it is obvious that the cost function of Eq. (4.35) is proportional to a convex combination of the cost function of Eq. (4.13) and the sum of the squares of the Euclidean norm of the partial derivatives of $e(k, \varepsilon(k), \varepsilon(k+1))$ with respect to each parametric error. This means that an appropriate selection of the factor $\lambda(k)$ can reflect a good trade-off between the nominal value of the cost function adopted in state estimations and its sensitivity to parametric modeling errors. From these respects, the factor $\lambda(k)$ can also be explained as a penalization on the sensitivity of the cost function in Eq. (4.13) to modeling errors. This explanation leads to the name of sensitivity penalization-based robust state estimator for the resulted estimation procedure [11].

Through minimizing the cost function of Eq. (4.35), a robust state estimator can be derived, which has a similar form as that of the Kalman filter and can be recursively realized. This robust state estimator consists of three steps given in the following procedure.

Algorithm 4.4.1. The Robust State Estimator

- *Initialize the state estimator with* $\hat{x}(0)$ *and* $P(0)$.
- *Assume that both matrices* $P(k)$ *and* $Q(k)$ *are invertible. Then, at the time instant* $k+1$, *modify the system parameters as*

$$
\begin{aligned}
&\bar{P}(k) = \left[P^{-1}(k) + \lambda(k) S^T(k) S(k)\right]^{-1},\\
&\bar{Q}(k) = \left\{Q^{-1}(k) + \lambda(k) T^T(k)\left[I + \lambda(k) S(k) P(k) S^T(k)\right]^{-1} T(k)\right\}^{-1},\\
&\bar{G}(k,0) = G(k,0) - \lambda(k) A(k,0) \bar{P}(k) S^T(k) T(k),\\
&\bar{A}(k,0) = [A(k,0) - \lambda(k) \bar{G}(k,0) \bar{Q}(k) T^T(k) S(k)][I - \lambda(k) \bar{P}(k) S^T(k) S(k)],\\
&S(k) = \mathbf{col}\begin{bmatrix} C(k+1,0)\frac{\partial A(k,0)}{\partial \varepsilon_j(k)} \\ \frac{\partial C(k+1,0)}{\partial \varepsilon_j(k+1)} A(k,0) \end{bmatrix}_{j=1}^{n_e}, \quad T(k) = \mathbf{col}\begin{bmatrix} C(k+1,0)\frac{\partial G(k,0)}{\partial \varepsilon_j(k)} \\ \frac{\partial C_{k+1,0}}{\partial \varepsilon_j(k+1)} G(k,0) \end{bmatrix}_{j=1}^{n_e},\\
&z(k) = \begin{bmatrix} C(k+1,0)\frac{\partial B(k,0)}{\partial \varepsilon_j(k)} u(k) \\ \frac{\partial C(k+1,0)}{\partial \varepsilon_j(k+1)} B(k,0) u(k) + \frac{\partial D(k+1,0)}{\partial \varepsilon_j(k+1)} u(k+1) \end{bmatrix}_{j=1}^{n_e}.
\end{aligned}
$$

- *Update the matrices* $P(k+1)$, $L(k+1)$, *and* $K(k+1)$ *respectively as follows:*

$$P_-(k+1) = A(k,0) \bar{P}(k) A^T(k,0) + \bar{G}(k,0) \bar{Q}(k) \bar{G}^T(k,0), \tag{4.36a}$$

$$P(k+1) = \left\{P_-^{-1}(k+1) + C^T(k+1,0) R^{-1}(k+1) C(k+1,0)\right\}^{-1}, \tag{4.36b}$$

$$L(k+1) = P(k+1) C^T(k+1,0) R^{-1}(k+1), \tag{4.36c}$$

$$
\begin{aligned}
K(k+1) = {} & \lambda(k) P(k+1) P_-^{-1}(k+1)\left[A(k,0) P(k) S^T(k) + G(k,0) Q(k) T^T(k)\right]\\
& \times \left\{I + \lambda(k)\left[S(k) P(k) S^T(k) + T(k) Q(k) T^T(k)\right]\right\}^{-1}.
\end{aligned} \tag{4.36d}
$$

Moreover, update the state estimate as follows:

$$
\begin{aligned}
\hat{x}(k+1) = {} & \bar{A}(k,0)\hat{x}(k) + [I - L(k+1) C(k+1,0)]\, B(k,0) u(k) + K(k+1) z(k)\\
& - L(k+1)\left\{C(k+1,0) \bar{A}(k,0) \hat{x}(k) + D(k+1,0) u(k+1) - y(k+1)\right\}.
\end{aligned} \tag{4.37}
$$

Replace the time index k *with* $k+1$ *and return to the second step.*

Comparison of this estimation procedure with the Kalman filter shows that, except the 2nd step in which system parameters are modified and the additional term $z(k)$ is included in

Eq. (4.37), these two-state estimators have almost the same form. This implies that the computational complexity of this robust state estimator is comparable to that of the Kalman filter, and through simply modifying parameters of the Kalman filter, its robustness against parametric modeling errors can be improved.

It is worth mentioning that although the matrix $P(k+1)$ in the robust state estimator is updated at every time instant in a recursive way similarly to that of the Kalman filter, it is not the covariance matrix of estimation errors. As a matter of fact, from an engineering point of view, the matrix $P(k)$ in this robust state estimator differs significantly from its counterpart in the Kalman filter. It will be made clear in Subsection 3.4.3 that in the robust state estimator, although this matrix is not equal to the covariance matrix of estimation errors, these two matrices are closely related to each other. To avoid possible confusions, this matrix is called the pseudo-covariance matrix in robust estimations.

4.4.2 Derivation of the Robust Estimator

In this subsection, we provide a derivation for the robust estimation procedure given in the previous subsection. For brevity, abbreviate $A(k,0)$, $\bar{A}(k,0)$, $B(k,0)$, $C(k,0)$, $G(k,0)$, and $\bar{G}(k,0)$ respectively as $A(k)$, $\bar{A}(k)$, $B(k)$, $C(k)$, $G(k)$, and $\bar{G}(k)$. From the definition of $e(k,\varepsilon(k),\varepsilon(k+1))$ we have that, for every $j\in\{1,2,\ldots,n_e\}$,

$$\begin{aligned}\frac{\partial e(k,\varepsilon(k),\varepsilon(k+1))}{\partial\varepsilon_j(k)} &= -C(k+1,\varepsilon(k+1))\frac{\partial B(k,\varepsilon(k))}{\partial\varepsilon_j(k)}u(k)\\ &\quad -C(k+1,\varepsilon(k+1))\frac{\partial A(k,\varepsilon(k))}{\partial\varepsilon_j(k)}\hat{x}(k|k+1)\\ &\quad -C(k+1,\varepsilon(k+1))\frac{\partial G(k,\varepsilon(k))}{\partial\varepsilon_j(k)}\hat{w}(k|k+1)\\ &= -C(k+1,\varepsilon(k+1))\frac{\partial B(k,\varepsilon(k))}{\partial\varepsilon_j(k)}u(k)\\ &\quad -\left[C(k+1,\varepsilon(k+1))\frac{\partial A(k,\varepsilon(k))}{\partial\varepsilon_j(k)}\;\; C(k+1,\varepsilon(k+1))\frac{\partial G(k,\varepsilon(k))}{\partial\varepsilon_j(k)}\right]\alpha(k),\end{aligned}\tag{4.38}$$

$$\begin{aligned}\frac{\partial e(k,\varepsilon(k),\ \varepsilon(k+1))}{\partial\varepsilon_j(k+1)} &= -\frac{\partial C(k+1,\varepsilon(k+1))}{\partial\varepsilon_{k+1,k}}B(k,\varepsilon(k))u(k)\\ &\quad -\frac{\partial C(k+1,\varepsilon(k+1))}{\partial\varepsilon_{k+1,k}}A(k,\varepsilon(k))\hat{x}(k|k+1)\\ &\quad -\frac{\partial D(k+1,\varepsilon(k+1))}{\partial\varepsilon_j(k+1)}u(k+1)\end{aligned}$$

$$
\begin{aligned}
&\qquad - \frac{\partial C(k+1, \varepsilon(k+1))}{\partial \varepsilon_j(k+1)} G(k, \varepsilon(k)) \hat{w}(k|k+1) \\
&= -\frac{\partial C(k+1, \varepsilon(k+1))}{\partial \varepsilon_j(k+1)} B(k, \varepsilon(k)) u(k) - \frac{\partial D(k+1, \varepsilon(k+1))}{\partial \varepsilon_j(k+1)} u(k+1) \\
&\qquad - \left[\frac{\partial C(k+1, \varepsilon(k+1))}{\partial \varepsilon_{k+1,k}} A(k, \varepsilon(k)) \quad \frac{\partial C(k+1, \varepsilon(k+1))}{\partial \varepsilon_j(k+1)} G(k, \varepsilon(k)) \right] \alpha(k),
\end{aligned}
\tag{4.39}
$$

where $\alpha(k) = \mathbf{col}\{\hat{x}(k|k+1),\ \hat{w}(k|k+1)\}$, which is consistent with the symbol adopted in the derivation of the Kalman filter. Then, from the definition of the vector $z(k)$ and those of the matrices $S(k)$ and $T(k)$ we can straightforwardly prove that

$$
\sum_{j=1}^{n_e} \left|\left| \frac{\partial e(k,0,0)}{\partial \varepsilon_j(k)} \right|\right|_2^2 + \left|\left| \frac{\partial e(k,0,0)}{\partial \varepsilon_j(k+1)} \right|\right|_2^2 = ||z(k) - [S(k)\ \ T(k)]\alpha(k)||_2^2 . \tag{4.40}
$$

As in Subsection 4.3.1, denote $y(k+1) - D(k+1)u(k+1) - C(k+1)B(k)u(k)$ by $\bar{y}(k+1)$. Then, from the definition of the cost function $J(\hat{x}(k|k+1), \hat{w}(k|k+1))$ and Eq. (4.40) direct algebraic manipulations show that

$$
\begin{aligned}
J(\alpha(k)) = \frac{1}{2} \Big\{ &(\alpha(k) - \alpha_0(k))^T \mathbf{diag}\Big\{ P^{-1}(k),\ Q^{-1}(k) \Big\} (\alpha(k) - \alpha_0(k)) \\
&+ \{\bar{y}(k+1) - C(k+1)[A(k)\ G(k)]\alpha(k)\}^T\ R^{-1}(k+1) \\
&\times \{\bar{y}(k+1) - C(k+1)[A(k)\ G(k)]\alpha(k)\} \\
&+ \lambda(k)\,\{z(k) - [S(k)\ T(k)]\,\alpha(k)\}^T\ \{z(k) - [S(k)\ T(k)]\,\alpha(k)\} \Big\},
\end{aligned}
\tag{4.41}
$$

where $\alpha_0(k) = \mathbf{col}\{\hat{x}(k|k),\ 0\}$. Here, with a little abuse of notation, the cost function $J(\hat{x}(k|k+1), \hat{w}(k|k+1))$ is written as $J(\alpha(k))$. We adopt this expression in the remaining of this subsection. Therefore,

$$
\begin{aligned}
\frac{\partial J(\alpha(k))}{\partial \alpha(k)} &= \mathbf{diag}\Big\{ P^{-1}(k),\ Q^{-1}(k) \Big\} [\alpha(k) - \alpha_0(k)] + \{C(k+1)[A(k)\ G(k)]\}^T\ R^{-1}(k+1) \\
&\quad \times \{C(k+1)[A(k)\ G(k)]\alpha(k) - \bar{y}(k+1)\} + \lambda(k)[S(k)\ T(k)]^T \\
&\quad \times [[S(k)\ T(k)]\alpha(k) - z(k)] \\
&= \Big\{ \mathbf{diag}\Big\{ P^{-1}(k),\ Q^{-1}(k) \Big\} + \lambda(k)[S(k)\ T(k)]^T [S(k)\ T(k)] \\
&\quad + [A(k)\ G(k)]^T C^T(k+1) R^{-1}(k+1) C(k+1)[A(k)\ G(k)] \Big\} \alpha(k) \\
&\quad - \mathbf{diag}\Big\{ P^{-1}(k),\ Q^{-1}(k) \Big\} \alpha_0(k) \\
&\quad - [A(k)\ G(k)]^T C^T(k+1) R^{-1}(k+1) \bar{y}(k+1) \\
&\quad - \lambda(k)[S(k)\ T(k)]^T z(k).
\end{aligned}
\tag{4.42}
$$

Note that when all the three matrices $P(k)$, $Q(k)$, and $R(k+1)$ are positive definite, their inverses are also positive definite. It can be straightforwardly proven from Eq. (4.41) that the cost function $J(\alpha(k))$ is a strictly convex function of the vector $\alpha(k)$. It can therefore be declared that the cost function $J(\alpha(k))$ achieves its minimum value at the optimal $\alpha(k)$, denote it by $\hat{\alpha}(k)$, if and only if at this $\hat{\alpha}(k)$, its first-order derivative with respect to $\alpha(k)$ is equivalent to zero. On the basis of this observation and Eq. (4.42), we have that

$$\begin{aligned}\hat{\alpha}(k) \;=\; & \Big(\mathbf{diag}\Big\{P^{-1}(k),\; Q^{-1}(k)\Big\} + \lambda(k)[S(k)\;T(k)]^T[S(k)\;T(k)] \\ & \quad + [A(k)\;G(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\;G(k)]\Big)^{-1} \\ & \times \Big(\mathbf{diag}\Big\{P^{-1}(k),\; Q^{-1}(k)\Big\}\alpha_0(k) + [A(k)\;G(k)]^T C^T(k+1)R^{-1}(k+1)\bar{y}(k+1) \\ & \quad + \lambda(k)[S(k)\;T(k)]^T z(k)\Big). \end{aligned} \tag{4.43}$$

On the other hand, we can prove by direct algebraic manipulations that

$$\begin{aligned} & T^T(k)T(k) - \lambda(k)T^T(k)S(k)[P^{-1}(k) + \lambda(k)S^T(k)S(k)]^{-1}S^T(k)T(k) \\ = \; & T^T(k)[I + \lambda(k)S(k)P(k)S^T(k)]^{-1}T(k). \end{aligned} \tag{4.44}$$

Then from Lemma 2.2 and the definitions of the matrices $\bar{P}(k)$ and $\bar{Q}(k)$ we can immediately obtain the following equality:

$$\begin{aligned} & \mathbf{diag}\Big\{P^{-1}(k),\; Q^{-1}(k)\Big\} + \lambda(k)[S(k)\;T(k)]^T[S(k)\;T(k)] \\ = & \begin{bmatrix} P^{-1}(k) + \lambda(k)S^T(k)S(k) & \lambda(k)S^T(k)T(k) \\ \lambda(k)T^T(k)S(k) & Q^{-1}(k) + \lambda(k)T^T(k)T(k) \end{bmatrix} \\ = & \begin{bmatrix} I & 0 \\ \lambda(k)T^T(k)S(k)\bar{P}(k) & I \end{bmatrix} \begin{bmatrix} \bar{P}^{-1}(k) & 0 \\ 0 & \bar{Q}^{-1}(k) \end{bmatrix} \begin{bmatrix} I & \lambda(k)\bar{P}(k)S^T(k)T(k) \\ 0 & I \end{bmatrix} \end{aligned} \tag{4.45}$$

Substituting this relation into Eq. (4.43), we can further prove that

$$\begin{aligned} \hat{\alpha}(k) \;=\; & \begin{bmatrix} I & -\lambda(k)\bar{P}(k)S^T(k)T(k) \\ 0 & I \end{bmatrix} \\ & \times \left\{\begin{bmatrix} \bar{P}^{-1}(k) & 0 \\ 0 & \bar{Q}^{-1}(k) \end{bmatrix} + [A(k)\;\bar{G}(k)]^T C^T(k+1)R^{-1}(k+1) \right. \\ & \left. \times\; C(k+1)[A(k)\;\bar{G}(k)] \right\}^{-1} \end{aligned}$$

$$\begin{aligned}\times \Big[\mathbf{col}\left\{I,\ -\lambda(k)T^T(k)S(k)\bar{P}(k)\right\} P^{-1}(k)\hat{x}(k|k) + [A(k)\ \bar{G}(k)]^T \\ \times C^T(k+1)R^{-1}(k+1)\bar{y}_{k+1} + \lambda(k)[S(k)\ \bar{T}(k)]^T z(k)\Big],\end{aligned} \tag{4.46}$$

where $\bar{T}(k) = T(k) - \lambda(k)S(k)\bar{P}(k)S^T(k)T(k)$.

Hence, from the state transition equation of the plant state space model given by Eq. (4.34a) we can construct an estimate about the state vector $x(k+1)$ from $y(s)|_{s=0}^{k+1}$:

$$\hat{x}(k+1) = A(k)\hat{x}(k|k+1) + B(k)u(k) + G(k)\hat{w}(k|k+1). \tag{4.47}$$

Clearly, this estimate satisfies the following equation:

$$\begin{aligned}
&\hat{x}(k+1) - B(k)u(k) \\
&= [A(k)\ G(k)]\hat{\alpha}(k) \\
&= [A(k)\ \bar{G}(k)]\Big\{\mathbf{diag}\left\{\bar{P}^{-1}(k),\ \bar{Q}^{-1}(k)\right\} \\
&\quad + [A(k)\ \bar{G}(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\ \bar{G}(k)]\Big\}^{-1} \\
&\quad \times \left\{\begin{bmatrix} I \\ -\lambda(k)T^T(k)S(k)\bar{P}(k)\end{bmatrix} P^{-1}(k)\hat{x}(k|k)\right. \\
&\quad \left. + \begin{bmatrix} A^T(k) \\ \bar{G}^T(k)\end{bmatrix} C^T(k+1)R^{-1}(k+1)\bar{y}(k+1) + \lambda(k)\begin{bmatrix} S^T(k) \\ \bar{T}^T(k)\end{bmatrix} z(k)\right\} \\
&= [A(k)\ \bar{G}(k)]\Big\{I + \mathbf{diag}\left\{\bar{P}(k),\, \bar{Q}(k)\right\}[A(k)\ \bar{G}(k)]^T C^T(k+1)R^{-1}(k+1) \\
&\quad \times C(k+1)[A(k)\ \bar{G}(k)]\Big\}^{-1}\mathbf{diag}\left\{\bar{P}(k),\, \bar{Q}(k)\right\} \\
&\quad \times \left\{\begin{bmatrix} I \\ -\lambda(k)T^T(k)S(k)\bar{P}(k)\end{bmatrix} P^{-1}(k)\hat{x}(k|k) + \begin{bmatrix} A^T(k) \\ \bar{G}^T(k)\end{bmatrix} C^T(k+1)\right. \\
&\quad \left. \times R^{-1}(k+1)\bar{y}(k+1) + \lambda(k)\begin{bmatrix} S^T(k) \\ \bar{T}^T(k)\end{bmatrix} z(k)\right\} \\
&= \Big\{I + [A(k)\ \bar{G}(k)]\mathbf{diag}\left\{\bar{P}(k),\, \bar{Q}(k)\right\}[A(k)\ \bar{G}(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)\Big\}^{-1} \\
&\quad \times [A(k)\ \bar{G}(k)]\mathbf{diag}\left\{\bar{P}(k),\, \bar{Q}(k)\right\} \\
&\quad \times \left\{\begin{bmatrix} I \\ -\lambda(k)T^T(k)S(k)\bar{P}(k)\end{bmatrix} P^{-1}(k)\hat{x}(k|k) + \begin{bmatrix} A^T(k) \\ \bar{G}^T(k)\end{bmatrix} C^T(k+1)\right. \\
&\quad \left. \times R^{-1}(k+1)\bar{y}(k+1) + \lambda(k)\begin{bmatrix} S^T(k) \\ \bar{T}^T(k)\end{bmatrix} z(k)\right\}
\end{aligned}$$

$$= \left[I + P_{-}(k+1)C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1} \left\{\bar{A}(k)\hat{x}(k|k) + P_{-}(k+1)C^T(k+1)\right.$$
$$\left. \times R^{-1}(k+1)\bar{y}(k+1) + \lambda(k)\left[A(k)\bar{P}(k)S^T(k) + \bar{G}(k)\bar{Q}(k)\bar{T}^T(k)\right]z(k)\right\}, \qquad (4.48)$$

where $P_{-}(k+1) = A(k)\bar{P}(k)A^T(k) + \bar{G}(k)\bar{Q}(k)\bar{G}^T(k)$. In the derivation of the last equality, we used the relation $\bar{P}(k)P^{-1}(k) = I - \lambda(k)\bar{P}(k)S^T(k)S(k)$, which is a direct result of the definition of the matrix $\bar{P}(k)$.

On the other hand, from the definitions of the matrices $\bar{T}(k)$ and $\bar{P}(k)$ we have that

$$\begin{aligned} \bar{T}(k) &= T(k) - \lambda(k)S(k)\bar{P}(k)S^T(k)T(k) \\ &= \left\{I - \lambda(k)S(k)\left[P_{-}^{-1}(k) + \lambda(k)S^T(k)S(k)\right]^{-1} S^T(k)\right\} T(k) \\ &= \left[I + \lambda(k)S(k)P(k)S^T(k)\right]^{-1} T(k). \end{aligned}$$

Based on this equality and the definitions of the matrices $\bar{G}(k)$, $\bar{P}(k)$, and $\bar{Q}(k)$, direct but tedious algebraic manipulations show that

$$\begin{aligned} & A(k)\bar{P}(k)S^T(k) + \bar{G}(k)\bar{Q}(k)\bar{T}^T(k) \\ = & \left[A(k)P(k)S^T(k) + G(k)Q(k)T^T(k)\right] \\ & \times \left\{I + \lambda(k)\left[S(k)P(k)S^T(k) + T(k)Q(k)T^T(k)\right]\right\}^{-1}. \end{aligned} \qquad (4.49)$$

Substituting Eq. (4.49) and the definition of $\bar{y}(k+1)$ into Eq. (4.48), we have that

$$\begin{aligned} \hat{x}(k+1) &= \bar{A}(k)\hat{x}(k|k) + B(k)u(k) + \left\{\left[I + P_{-}(k+1)C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1}\right. \\ &\quad \left. - I\right\} \bar{A}(k)\hat{x}(k|k) + \left[I + P_{-}(k+1)C^T(k+1)R^{-1}(k+1)C(k+1)\right]^{-1} \\ &\quad \times \left\{P_{-}(k+1)C^T(k+1)R^{-1}(k+1)\bar{y}(k+1)\right. \\ &\quad \left. + \lambda(k)\left[A(k)\bar{P}(k)S^T(k) + \bar{G}(k)\bar{Q}(k)\bar{T}^T(k)\right]z(k)\right\} \\ &= \bar{A}(k)\hat{x}(k|k) + B(k)u(k) - L(k+1)C(k+1)\bar{A}(k)\hat{x}(k|k) \\ &\quad + L(k+1)\left[y(k+1) - D(k+1)u(k+1) - C(k+1)B(k)u(k)\right] \\ &\quad + \lambda(k)P(k+1)P_{-}^{-1}(k+1)\left[A(k)P(k)S^T(k) + G(k)Q(k)T^T(k)\right] \\ &\quad \times \left\{I + \lambda(k)\left[S(k)P(k)S^T(k) + T(k)Q(k)T^T(k)\right]\right\}^{-1} z(k) \\ &= \bar{A}(k)\hat{x}(k|k) + \left[I - L(k+1)C(k+1)\right]B(k)u(k) + K(k+1)z(k) \end{aligned}$$

$$- L(k+1)\left\{C(k+1)\bar{A}(k)\hat{x}(k|k) + D(k+1)u(k+1) - y(k+1)\right\}. \qquad (4.50)$$

This is completely the same as that of Eq. (4.37) if we replace $\hat{x}(k|k)$ with $\hat{x}(k)$ and recall the definitions of the involved matrices, such as $\bar{A}(k)$, $B(k)$, and so on.

A detailed comparison of the above derivation process with that of the Kalman filter shows that, due to the introduction of the penalizing factor on the sensitivity of $e(k, \varepsilon(k), \varepsilon(k+1))$ to parametric modeling errors, the optimal estimates about the plant state vector $x(k)$ and the process noise vector $w(k)$ from the plant measurements $y(0)$, $y(1)$, $\cdots$, $y(k+1)$ have been changed. The magnitude of these changes depends on both the value of the penalizing factor and how sensitive the prediction error is to parametric errors. Owing to the quadratic form of the penalization, the changes of the optimal estimates can be realized through only modifying system parameters, that is, there is no need to change the structure of the estimates itself. This is of great significance from both theoretical analysis and implementation. This also means that most of the major attractive characteristics of the Kalman filter, especially those in realizations and computations, have been inherited by the above robust state estimator.

4.4.3 Asymptotic Properties of the Robust State Estimator

Similarly to the Kalman filter, the robust state estimator obtained in the previous subsection also has many attractive convergence properties. In this subsection, we investigate some asymptotic behaviors of its update gain matrices $L(k+1)$ and $K(k+1)$. It is shown that, under some conditions that can usually be satisfied, these two matrices converge respectively to a constant matrix.

To perform this analysis, we at first establish another recursive expression for the pseudo-covariance matrix $P(k)$. Compared to that of the estimation procedure, this expression gives a more explicit relation between the pseudo-covariance matrices of two successive sampled time constants.

Theorem 4.2. *Define the matrices $\check{Q}(k) = (Q^{-1}(k) + \lambda(k)T^T(k)T(k))^{-1}$ and $\check{A}(k) = A(k,0) - \lambda(k)G(k,0)\check{Q}(k)T^T(k)S(k)$ and assume that the matrix $\check{A}(k)$ is invertible. Moreover, define the matrices*

$$\tilde{A}(k) = \check{A}(k) + G(k,0)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k), \quad \tilde{G}(k) = \check{A}^{-1}(k)G(k,0),$$
$$\tilde{Q}(k) = \check{Q}(k) + \check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\tilde{G}(k)\check{Q}(k),$$
$$\tilde{S}(k) = \sqrt{\lambda(k)}\left[I + \lambda(k)T(k)Q(k)T^T(k)\right]^{-1/2} S(k),$$
$$\tilde{C}(k+1) = \begin{bmatrix} \tilde{S}(k)\check{A}^{-1}(k) \\ C(k+1,0) \end{bmatrix},$$

$$\tilde{R}(k+1)=\begin{bmatrix} I+\tilde{S}(k)\tilde{B}(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k) & 0 \\ 0 & R(k+1) \end{bmatrix}.$$

Then, for each positive integer k, we have

$$\begin{aligned} P^{-1}(k+1) &= \left[\tilde{A}(k)P(k)\tilde{A}^T(k)+G(k,0)\tilde{Q}(k)G^T(k,0)\right]^{-1} \\ &\quad + \tilde{C}^T(k+1)\tilde{R}^{-1}(k+1)\tilde{C}(k+1). \end{aligned} \tag{4.51}$$

Proof. To simplify mathematical expressions, in this proof, $A(k,0)$, $G(k,0)$, and $C(k+1,0)$ are once again respectively abbreviated as $A(k)$, $G(k)$, and $C(k+1)$. From the derivation of the robust state estimator given in Subsection 3.4.1 it is clear that the state estimate at the $(k+1)$th sampled time instant is

$$\begin{aligned} \hat{x}(k+1) &= B(k)u(k)+[A(k)\ G(k)]\left\{\mathbf{diag}\left\{P^{-1}(k),\ Q^{-1}(k)\right\}\right. \\ &\quad +\lambda(k)[S(k)\ T(k)]^T[S(k)\ T(k)] \\ &\quad \left.+\ [A(k)\ G(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\ G(k)]\right\}^{-1} \\ &\quad \times\left\{\mathbf{diag}\left\{P^{-1}(k),\ Q^{-1}(k)\right\}\mathbf{col}\left\{\hat{x}(k),\ 0\right\}[A(k)\ G(k)]^T C^T(k+1)\right. \\ &\quad \left.\times\ R^{-1}(k+1)\bar{y}(k+1)+\lambda(k)[S(k)\ T(k)]^T z(k)\right\}. \end{aligned} \tag{4.52}$$

On the other hand, according to Eq. (4.50), this state estimate can also be expressed as

$$\begin{aligned} \hat{x}(k+1) &= \bar{A}(k)\hat{x}(k)+P(k+1)C^T(k+1)R^{-1}(k+1)[\bar{y}(k+1)-C(k+1)\bar{A}(k)\hat{x}(k)] \\ &\quad +B(k)u(k)+K(k)z(k). \end{aligned} \tag{4.53}$$

Note that Eqs. (4.52) and (4.53) are just two different expressions for the same state estimate $\hat{x}(k+1)$. It is necessary that the coefficient matrices in these two expressions, which are respectively for $\hat{x}(k)$, $\bar{y}(k+1)$, and $z(k)$, are equal to each other. Equalizing the coefficient matrices of $\bar{y}(k+1)$ leads to the equality

$$\begin{aligned} P(k+1) &= [A(k)\ G(k)]\left\{\mathbf{diag}\left\{P^{-1}(k),\ Q^{-1}(k)\right\}+\lambda(k)[S(k)\ T(k)]^T[S(k)\ T(k)]\right. \\ &\quad \left.+\ [A(k)\ G(k)]^T C^T(k+1)R^{-1}(k+1)C(k+1)[A(k)\ G(k)]\right\}^{-1}[A(k)\ G(k)]^T. \end{aligned} \tag{4.54}$$

On the other hand, direct algebraic operations show that

$$\lambda(k)S^T(k)S(k)-\lambda^2(k)S^T(k)T(k)[Q^{-1}(k)+\lambda(k)T^T(k)T(k)]^{-1}T^T(k)S(k)$$

$$
\begin{aligned}
&= \lambda(k)S^T(k)\left\{I - \lambda(k)T(k)[I + \lambda(k)Q(k)T^T(k)T(k)]^{-1}Q(k)T^T(k)\right\}S(k)\\
&= \lambda(k)S^T(k)[I + \lambda(k)T(k)Q(k)T^T(k)]^{-1}S(k)\\
&= \tilde{S}^T(k)\tilde{S}(k).
\end{aligned}
\tag{4.55}
$$

Then, from Lemma 2.2 and the definition of $\check{Q}(k)$ we can immediately obtain the following relation:

$$
\begin{aligned}
&\mathbf{diag}\left\{P^{-1}(k),\ Q^{-1}(k)\right\} + \lambda(k)[S(k)\ T(k)]^T[S(k)\ T(k)]\\
&= \begin{bmatrix} P^{-1}(k) + \lambda(k)S^T(k)S(k) & \lambda(k)S^T(k)T(k)\\ \lambda(k)T^T(k)S(k) & Q^{-1}(k) + \lambda(k)T^T(k)T(k)\end{bmatrix}\\
&= \begin{bmatrix} I & \lambda(k)S^T(k)T(k)\check{Q}(k)\\ 0 & I\end{bmatrix}\begin{bmatrix} P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k) & 0\\ 0 & \check{Q}^{-1}(k)\end{bmatrix}\\
&\quad\times\begin{bmatrix} I & 0\\ \lambda(k)\check{Q}(k)T^T(k)S(k) & I\end{bmatrix}.
\end{aligned}
\tag{4.56}
$$

Substituting Eq. (4.56) into Eq. (4.55), we have

$$
\begin{aligned}
P(k+1) &= \left([A(k)\ G(k)]\begin{bmatrix} I & 0\\ \lambda(k)\check{Q}(k)T^T(k)S(k) & I\end{bmatrix}^{-1}\right)\\
&\quad\times\left\{\begin{bmatrix} P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k) & 0\\ 0 & \check{Q}^{-1}(k)\end{bmatrix}\right.\\
&\quad+\left([A(k)\ G(k)]\begin{bmatrix} I & 0\\ \lambda(k)\check{Q}(k)T^T(k)S(k) & I\end{bmatrix}^{-1}\right)^T\\
&\quad\times C^T(k+1)R^{-1}(k+1)C(k+1)\,([A(k)\ G(k)]\\
&\quad\left.\times\begin{bmatrix} I & 0\\ \lambda(k)\check{Q}(k)T^T(k)S(k) & I\end{bmatrix}^{-1}\right)\Bigg\}^{-1}\\
&\quad\times\left([A(k)\ G(k)]\begin{bmatrix} I & 0\\ \lambda(k)\check{Q}(k)T^T(k)S(k) & I\end{bmatrix}^{-1}\right)^T\\
&= [\check{A}(k)\ G(k)]\left\{\mathbf{diag}\left\{P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k),\ \check{Q}^{-1}(k)\right\} + [\check{A}(k)\ G(k)]^T C^T(k+1)\right.\\
&\quad\left.\times R^{-1}(k+1)C(k+1)[\check{A}(k)\ G(k)]\right\}^{-1}[\check{A}(k)\ G(k)]^T
\end{aligned}
$$

$$\begin{aligned}
&= \Big\{I + [\check{A}(k)\ G(k)]\mathbf{diag}\Big\{(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1},\ \check{Q}(k)\Big\} \\
&\quad \times [\check{A}(k)\ G(k)]^T C^T(k+1)\ R^{-1}(k+1)C(k+1)\Big\}^{-1} [\check{A}(k)\ G(k)] \\
&\quad \times \mathbf{diag}\Big\{(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1},\ \check{Q}(k)\Big\}[\check{A}(k)\ G(k)]^T \\
&= \Big\{\Big[\check{A}(k)(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1}\check{A}^T(k) + G(k)\check{Q}(k)G^T(k)\Big]^{-1} \\
&\quad + C^T(k+1)R^{-1}(k+1)C(k+1)\Big\}^{-1}.
\end{aligned} \tag{4.57}$$

When $\check{A}(k)$ is invertible, from the definition of the matrix $\tilde{G}(k)$ we have that

$$\begin{aligned}
&\check{A}(k)(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1}\check{A}^T(k) + G(k)\check{Q}(k)G^T(k) \\
&= \check{A}(k)\Big\{(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1} + \tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\Big\}\check{A}^T(k).
\end{aligned} \tag{4.58}$$

In addition, direct algebraic manipulations show that

$$\begin{aligned}
&\Big\{\Big[P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k)\Big]^{-1} + \tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\Big\}^{-1} \\
&= \tilde{S}^T(k)\Big[I + \tilde{S}(k)\tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\Big]^{-1}\tilde{S}(k) \\
&\quad + \Big\{\tilde{G}(k)\Big[\check{Q}(k) + \check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\tilde{G}(k)\check{Q}(k)\Big]\tilde{G}^T(k) \\
&\quad + \Big[I + \tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\Big]P(k)\Big[I + \tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\Big]^T\Big\}^{-1}.
\end{aligned} \tag{4.59}$$

Substituting Eqs. (4.58) and (4.59) into Eq. (4.57), we obtain the following recursive expression for $P(k+1)$ when $\check{A}(k)$ is invertible:

$$\begin{aligned}
P^{-1}(k+1) &= \check{A}^{-T}(k)\Big[(P^{-1}(k) + \tilde{S}^T(k)\tilde{S}(k))^{-1} + \tilde{G}(k)\check{Q}(k)\tilde{G}^T(k)\Big]^{-1}\check{A}^{-1}(k) \\
&\quad + C^T(k+1)R^{-1}(k+1)C(k+1) \\
&= \Big\{G(k)\Big[\check{Q}(k) + \check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\tilde{G}(k)\check{Q}(k)\Big]G^T(k) \\
&\quad + \Big[\check{A}(k) + G(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\Big] \\
&\quad \times P(k)\Big[\check{A}(k) + G(k)\check{Q}(k)\tilde{G}^T(k)\tilde{S}^T(k)\tilde{S}(k)\Big]^T\Big\}^{-1} + \Big[\tilde{S}(k)\check{A}^{-1}(k)\Big]^T
\end{aligned}$$

$$
\begin{aligned}
&\times\left[I+\tilde{S}(k)\tilde{G}(k)\check{Q}(k)\tilde{G}^{T}(k)\tilde{S}^{T}(k)\right]^{-1}\left[\tilde{S}(k)\check{A}^{-1}(k)\right]\\
&+C^{T}(k+1)R^{-1}(k+1)C(k+1)\\
=\;&[\tilde{A}(k)P(k)\tilde{A}^{T}(k)+G(k)\tilde{Q}(k)G^{T}(k)]^{-1}+\tilde{C}^{T}(k+1)\tilde{R}^{-1}(k+1)\tilde{C}(k+1).
\end{aligned}
\tag{4.60}
$$

This completes the proof. □

The key step in the proof is the establishment of Eq. (4.56), which gives a decomposition of the matrix $\mathbf{diag}\left\{P^{-1}(k),\quad Q^{-1}(k)\right\}+\lambda(k)[S(k)\;T(k)]^{T}[S(k)\;T(k)]$, which is different from that of Eq. (4.45). It is this difference that leads to two different expressions for the matrix $P(k+1)$.

It is worth emphasizing that although the matrices $\tilde{A}(k)$ and $\tilde{Q}(k)$ have complicated expressions, they are completely determined by the nominal values of the plant parameters and the penalizing factor $\lambda(k)$ adopted in the cost function $J(\hat{x}(k|k+1),\hat{w}(k|k+1))$. On the other hand, note that the matrix $P(k+1)$ given by Eq. (4.51) has completely the same form as that of the Kalman filter given by Eq. (4.9b). This means that although the results of Theorem 4.2 are too complicated to be implemented in actual state estimations, they are convenient to be utilized in analyzing properties of the robust state estimator. In particular, from this expression the following asymptotic characteristics can be established for this estimator by the same token as that of Theorem 4.1. Only the associated results are stated here. The details of their mathematical derivations are omitted due to their close similarities with the Kalman filter. Another proof is given in [12] for the convergence of this matrix.

Corollary 4.1. *Assume that the plant nominal parameters are time invariant and all the derivatives of the plant parameters with respect to each parametric error at their nominal values are also time invariant. Moreover, assume that the matrix $\tilde{A}(k)$ defined in Theorem 4.2 is invertible. Then, when the penalizing factor $\lambda(k)$ in the cost function $J(\hat{x}(k|k+1),\hat{w}(k|k+1))$ is also time invariant, if the matrix pair $(\tilde{A}(k),\;G(k,0))$ is controllable and the matrix pair $(\tilde{A}(k),\;\tilde{C}(k))$ is observable, then, with the increment of the temporal variable k, the pseudo-covariance matrix $P(k)$ converges to a constant matrix.*

From Corollary 4.1 and the definitions of the matrices $L(k+1)$ and $K(k+1)$, we can declare that when all the conditions of this corollary are satisfied, these two matrices also converge respectively to a constant matrix. Therefore, the robust state estimator can also be replaced by an observer of constant gain matrices without sacrificing its steady-state estimation accuracy.

When the pseudo-covariance matrix $P(k)$ converges, denote its limit by the matrix P. On the basis of results on Riccati equations, we can prove through some algebraic manipulations that all the eigenvalues of the matrix $(I - PC^T R^{-1} C)\bar{A}$ have magnitudes smaller than 1. Here, the matrices $\bar{A}$, C, and R stand respectively for the time invariant values of the matrices $\bar{A}(k, 0)$, $C(k)$, and $R(k)$, that is, the robust state estimator converges to a stable time-invariant system. We can also prove that the convergence rate is exponential. A detailed discussion and related mathematical derivations can be found in [12]. We refer the interested reader to books like [1] for asymptotic properties of a Riccati recursion.

Now, we investigate the reasonability of the assumptions on positive definiteness of the pseudo-covariance matrix $P(k)$. Similarly to Eq. (4.30), define the matrix

$$\Phi(k) = \begin{bmatrix} \tilde{A}(k) & G(k,0)\tilde{Q}(k)G^T(k,0)\tilde{A}^{-T}(k) \\ \tilde{C}^T(k+1)\tilde{R}^{-1}(k+1)\tilde{C}(k+1)\tilde{A}(k) & [I + \tilde{C}^T(k+1)\tilde{R}^{-1}(k+1)\tilde{C}(k+1)G(k,0)\tilde{Q}(k)G^T(k,0)]\tilde{A}^{-T}(k) \end{bmatrix}. \tag{4.61}$$

Then, we can prove through straightforward algebraic manipulations that the matrix $\Phi(k)$ is Hamiltonian for each positive integer k. Moreover, similarly to Eq. (4.32), the relation between the matrices $P(k)$ and $P(0)$ given by Eq. (4.51) can be rewritten as

$$P(k) = \mathbf{H}_m \left(\prod_{s=k-1}^{0} \Phi(s),\ P(0) \right). \tag{4.62}$$

Note that from the definitions of the matrices $\tilde{Q}^{-1}(k)$ and $\tilde{R}^{-1}(k+1)$ it is obvious that when the matrices $Q(k)$ and $R(k+1)$ are positive definite, which is usually satisfied in actual applications, these two matrices are also positive definite. We therefore have that

$$\begin{aligned} &\left\{ G(k,0)\tilde{Q}(k)G^T(k,0)\tilde{A}^{-T}(k) \right\} \tilde{A}^T(k) = G(k,0)\tilde{Q}(k)G^T(k,0) \ge 0, \\ &\tilde{A}^T(k)\tilde{C}^T(k+1)\tilde{R}^{-1}(k+1)\tilde{C}(k+1)\tilde{A}(k) \\ &\quad = \left[\tilde{C}(k+1)\tilde{A}(k) \right]^T \tilde{R}^{-1}(k+1) \left[\tilde{C}(k+1)\tilde{A}(k) \right] \ge 0. \end{aligned}$$

Then, from the definition of the set $\mathcal{H}$ and Lemma 2.5 we have that $\Phi(k) \in \mathcal{H}$ for each k, and therefore

$$\prod_{s=k-1}^{0} \Phi(s) \in \mathcal{H}, \qquad k = 1, 2, 3, \ldots. \tag{4.63}$$

Eqs. (4.62) and (4.63) further imply that the matrix $P(k)$ is positive definite, provided that the matrix $P(0)$ is. Note that the assumption about the positive definiteness of the matrix $P(0)$ is a reasonable hypothesis and is generally satisfied in actual applications. We can therefore declare that the assumption on the matrix $P(k)$ adopted in the derivation of the robust state estimator is usually not very restrictive in actual applications.

4.4.4 Boundedness of Estimation Errors

As pointed out before, the matrix $P(k)$ in the robust state estimator is not the covariance matrix of estimation errors. Therefore, differently from the Kalman filter, even if this matrix converges, we still cannot immediately declare that the estimation errors of the robust state estimator are stochastically bounded. Intuitively, if both the plant and the estimator are stable, then the estimation error is bounded. On the other hand, for an unstable plant with modeling errors, development of a time-invariant robust state estimator with stochastically bounded estimation errors is in general impossible [1]. Based on these considerations, in this subsection, we investigate the boundedness for the estimation errors of the robust state estimator developed in the previous subsection under the condition that the plant itself is exponentially stable.

Based on this assumption and the results on the convergence of the pseudo-covariance matrix $P(k)$, the boundedness of the estimation bias of the robust state estimator can be established, as well as that of the covariance matrix of its estimation errors.

To simplify mathematical expressions in this boundedness analysis, we define the following matrices:

$$
\begin{aligned}
&A_f(k) = [I - L(k+1)C(k+1,0)]\bar{A}(k),\\
&F(k,\varepsilon(k),\varepsilon(k+1)) = L(k+1)C(k+1,\varepsilon(k+1))A(k,\varepsilon(k)),\\
&\bar{K}_1(k+1) = K(k+1)\mathbf{col}\begin{bmatrix} C(k+1,0)\frac{\partial B(k,0)}{\partial \varepsilon_j(k)} \\ \frac{\partial C(k+1,0)}{\partial \varepsilon_j(k+1)} B(k,0)u(k) \end{bmatrix}_{j=1}^{n_e},\\
&\bar{K}_2(k+1) = K(k+1)\mathbf{col}\begin{bmatrix} 0 \\ \frac{\partial D(k+1,0)}{\partial \varepsilon_j(k+1)} \end{bmatrix}_{j=1}^{n_e},\\
&\Phi(k,\varepsilon(k),\varepsilon(k+1)) = \begin{bmatrix} A(k,\varepsilon(k)) & 0 \\ F(k,\varepsilon(k),\varepsilon(k+1)) & A_f(k) \end{bmatrix},\\
&\Theta(k,\varepsilon(k),\varepsilon(k+1)) = \begin{bmatrix} G(k,\varepsilon(k)) & 0 \\ L(k+1)C(k+1,\varepsilon(k+1))G(k,\varepsilon(k)) & L(k+1) \end{bmatrix},\\
&\Gamma(k,\varepsilon(k),\varepsilon(k+1)) =\\
&\begin{bmatrix} B(k,\varepsilon(k)) & 0 \\ B(k,0) + \bar{K}_1(k+1) + L(k+1)[C(k+1,\varepsilon(k+1))B(k,\varepsilon(k)) - C(k+1,0)B(k,0)] & \bar{K}_2(k+1) + L(k+1)[D(k+1,\varepsilon(k+1)) - D(k+1,0)] \end{bmatrix}.
\end{aligned}
$$

Using these symbols, we can straightforwardly prove from Eqs. (4.34a), (4.34b), and (4.37) that

$$\begin{aligned}\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} &= \Theta(k, \varepsilon(k), \varepsilon(k+1)) \begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} + \Gamma(k, \varepsilon(k), \varepsilon(k+1)) \begin{bmatrix} u(k) \\ u(k+1) \end{bmatrix} \\ &\quad + \Phi(k, \varepsilon(k), \varepsilon(k+1)) \begin{bmatrix} w(k) \\ v(k+1) \end{bmatrix}. \end{aligned} \tag{4.64}$$

Without loss of generality, we can assume that the mathematical expectations of both the process noise vector $w(k)$ and the measurement error vector $v(k)$ are equal to zero. We also adopt in this subsection the hypothesis that these two random processes are independent of the parametric error vector $\varepsilon(k)$ of the plant. These hypotheses are usually satisfied in actual applications. Under these assumptions, direct algebraic manipulations from Eq. (4.64) show that

$$\mathbf{E}\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} = \Theta(k, \varepsilon(k), \varepsilon(k+1))\mathbf{E}\begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} + \Gamma(k, \varepsilon(k), \varepsilon(k+1)) \begin{bmatrix} u(k) \\ u(k+1) \end{bmatrix}, \tag{4.65}$$

$$\begin{aligned}\mathbf{Cov}\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} &= \Theta(k, \varepsilon(k), \varepsilon(k+1))\mathbf{Cov}\begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} \Theta^T(k, \varepsilon(k), \varepsilon(k+1)) \\ &\quad + \Phi(k, \varepsilon(k), \varepsilon(k+1)) \begin{bmatrix} Q(k) & 0 \\ 0 & R(k+1) \end{bmatrix} \Phi^T(k, \varepsilon(k), \varepsilon(k+1)). \end{aligned} \tag{4.66}$$

On the basis of these relations and the stability of the matrix $A(k, \varepsilon(k))$, the boundedness can be established for both the mathematical expectation and the covariance matrix of estimation errors of the robust state estimator.

Theorem 4.3. *Assume that both the plant input vector $u(k)$ and the parametric modeling error vector $\varepsilon(k)$ are elementwise bounded in magnitude. Moreover, assume that $\mathbf{E}(w(k)) = 0$, $\mathbf{E}(v(k)) = 0$, that all the conditions of Corollary 4.1 are satisfied, and that $w(k)$, $v(k)$, and $\varepsilon(k)$ are independent of each other. Then, at every time instant i, the robust state estimator of Eq. (4.39) has a bounded estimation bias, and its estimation errors have a bounded covariance matrix.*

A proof of this theorem is given in the appendix of this chapter. From that proof it appears that if $u(k) \equiv 0$, then $\mathbf{E}(x(k) - \hat{x}(k))$ converges to zero at least exponentially. However, when $u(k) \not\equiv 0$, estimation bias is usually unavoidable. On the other hand, note that the quadratic stability of a dynamic system implies its exponential stability, but the converse is generally not true. This means that conditions of the above theorem are in general weaker than those adopted in other robust state estimations, in which it is usually assumed that the plant under investigation is quadratically stable or asymptotically quadratically stable [1,5].

4.5 Bibliographic Notes

Removing noises from corrupted signals is essentially the basic task in signal processing. Many fundamental ideas in state estimations are motivated from signal processing, and the development of state estimation theory has also deepened understanding of signal processing. A common characteristic here is that in both of these fields, the squared error criterion is widely adopted, which traces back to the Gauss's estimation on the orbit parameters of Ceres. In his original work, Gauss intelligently recognized prominent and attractive characteristics of a squared error in estimations that are not possessed by the absolute value of an error. The latter was suggested by Laplace around the same time to deal with similar problems [13]. Adoption of this criterion in signal processing and state estimation has led to various milestone results; two of the most important ones are the Wiener and the Kalman filters [1,2,5].

Extensive influences of state estimation theory are recognized by the presentation of the 2008 Charles Stark Draper Prize to R. Kalman from the National Academy of Engineering, the United States of American, "for the development and dissemination of the optimal digital technique (known as the Kalman Filter) that is pervasively used to control a vast array of consumer, health, commercial and defense products." It is extensively recognized and widely accepted that "The Kalman Filter uses a mathematical technique that removes "noise" from series of data. From incomplete information, it can optimally estimate and control the state of a changing, complex system over time. The Kalman filter revolutionized the field of control theory and has become pervasive in engineering systems. It has been applied to systems and devices in nearly all engineering fields and continues to find new uses today. Applications include target tracking by radar, global positioning systems, hydrological modeling, atmospheric observations, time-series analyses in econometrics, and automated drug delivery." [14].

Various papers and books have now been published on state estimations. Among them, a general introduction of the major results in this field can be found in [5], in which estimations for both linear systems and nonlinear systems have been discussed, as well as robust estimations. An excellent book is [1], in which systematic and detailed discussions have been given for state estimations with linear systems. Optimality of the Kalman filter for a networked system has been revealed in [15], in which measurements are randomly lost due to imperfect communications. A recursive robust state estimator is originally derived through sensitivity penalization in [11] and extended in [16] to linear systems with intermittent data arrivals. The

Riemannian distance between two positive definite matrices was first introduced in [17] for studying the asymptotic properties of the Kalman filter. A regularized least-squares based method is suggested in [18] for robust state estimations. Except a parameter that requires online optimizations, the resulted estimator has a structure similar to that of the Kalman filter when some specific structure requirements are satisfied by parametric modeling errors.

Appendix 4.A

4.A.1 Proof of Theorem 4.1

When the plant is time invariant, the matrix $\Phi(k)$ defined previously is also a constant matrix. More precisely, we have that

$$\Phi(k) = \begin{bmatrix} A & GQG^TA^{-T} \\ C^TR^{-1}CA & [I + C^TR^{-1}CGQG^T]A^{-T} \end{bmatrix}. \tag{4.A.1}$$

Denote this matrix by Φ for brevity. Then, based on Eq. (4.32), we have that, for any time instant $k \geq 0$,

$$P(k) = \mathbf{H}_m\left(\Phi^k,\ P(0)\right), \tag{4.A.2}$$

where, as usual, Φ^0 is defined to be I_{2n}.

On the other hand, from the definition of the matrix $\Phi(k)$ and the time invariance of the plant parameter matrices we can straightforwardly prove that, for an arbitrary positive integer m,

$$\begin{aligned} &\Phi_{11}^T(1)\Phi_{21}(1) + \sum_{i=2}^{m}\left[\left(\prod_{k=1}^{i}\Phi_{11}^T(k)\right)\Phi_{21}(i)\left(\prod_{k=i-1}^{1}\Phi_{11}(k)\right)\right] \\ &\qquad = A^TC^TR^{-1}CA + (A^T)^2C^TR^{-1}CA^2 + \cdots \\ &\qquad\quad + (A^T)^mC^TR^{-1}CA^m, \end{aligned} \tag{4.A.3}$$

$$\begin{aligned} &\sum_{i=1}^{m-1}\left[\left(\prod_{k=m}^{i+1}\Phi_{11}(k)\right)\Phi_{12}(i)\left(\prod_{k=i}^{m}\Phi_{11}^T(k)\right)\right] + \Phi_{12}(m)\Phi_{11}^T(m) \\ &\qquad = GQG^T + AGQG^TA^T + A^2GQG^T(A^T)^2 + \cdots \\ &\qquad\quad + A^{m-1}GQG^T(A^T)^{m-1}, \end{aligned} \tag{4.A.4}$$

where $\Phi_{ij}(k)$ stands for the ith row jth column block submatrix of the matrix $\Phi(k)$. Note that

$$A^TC^TR^{-1}CA + (A^T)^2C^TR^{-1}CA^2 + \cdots + (A^T)^mC^TR^{-1}CA^m$$

$$
\begin{aligned}
&= A^T \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix}^T \left(I_m \otimes R^{-1}\right) \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix} A \\
&= A^T \left\{ \left(I_m \otimes R^{-1/2}\right) \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix} \right\}^T \left\{ \left(I_m \otimes R^{-1/2}\right) \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{m-1} \end{bmatrix} \right\} A. \qquad (4.A.5)
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
& GQG^T + AGQG^TA^T + \cdots + A^{m-1}GQG^T(A^T)^{m-1} \\
= \ & [G \ \ AG \ \ \cdots \ \ A^{m-1}G](I_n \otimes Q)[G \ \ AG \ \ \cdots \ \ A^{m-1}G]^T \\
= \ & \left\{[G \ \ AG \ \ \cdots \ \ A^{m-1}G]\left(I_n \otimes Q^{1/2}\right)\right\}\left\{[G \ \ AG \ \ \cdots \ \ A^{m-1}G]\left(I_n \otimes Q^{1/2}\right)\right\}^T. \qquad (4.A.6)
\end{aligned}
$$

Recall that each eigenvalue of the Kronecker product of two square matrices can be expressed as the product of the eigenvalues of these two matrices. We can declare that when the matrices R and Q are positive definite, both $I_n \otimes Q^{1/2}$ and $I_n \otimes R^{-1/2}$ are invertible. As the matrix A is also invertible by assumptions, this further implies that the matrix $\sum_{i=0}^{m-1} A^i GQG^T(A^T)^i$ is positive definite if and only if the matrix $[G \ \ AG \ \ \cdots \ \ A^{m-1}G]$ is of full row rank. Moreover, the matrix $\sum_{i=1}^{m}(A^T)^i C^T R^{-1} CA^i$ is positive definite if and only if the matrix $\mathbf{col}\left\{CA^i|_{i=0}^{m-1}\right\}$ is of full column rank.

When the matrix pair $(A, \ G)$ is controllable, we have that the matrix $[G \ \ AG \ \ \cdots \ \ A^{n-1}G]$ is of full row rank. Moreover, when the matrix pair $(A, \ C)$ is observable, we have that the matrix $\mathbf{col}\left\{CA^i|_{i=0}^{n-1}\right\}$ is of full column rank. Therefore, when $(A, \ G)$ and $(A, \ C)$ are respectively controllable and observable, we have that both matrices

$$
\Phi_{11}^T(1)\Phi_{21}(1) + \sum_{i=2}^{n}\left[\left(\prod_{k=1}^{i}\Phi_{11}^T(k)\right)\Phi_{21}(i)\right]
$$

and

$$
\sum_{i=1}^{n-1}\left[\left(\prod_{k=n}^{i+1}\Phi_{11}(k)\right)\Phi_{12}(i)\left(\prod_{k=i}^{n}\Phi_{11}^T(k)\right)\right] + \Phi_{12}(n)\Phi_{11}^T(n)
$$

are positive definite.

Hence, we can declare by Lemma 2.2 that when both the controllability condition and the observability condition are satisfied, the mapping $X \rightarrow \mathbf{H}_m(\Phi^n, X)$ is strictly contractive over the set of positive definite matrices under the Riemannian metric defined in Section 2.1, that is, there exists a scalar $\rho \in [0, 1)$ such that, for arbitrary positive definite matrices X and Y,

$$\delta\left\{\mathbf{H}_m(\Phi^n, X), \mathbf{H}_m(\Phi^n, Y)\right\} < \rho\delta(X, Y).$$

For an arbitrary positive integer k, let $I(k)$ denote the maximal integer not greater than $\frac{k}{n}$. Moreover, denote $k - nI(k)$ by $r(k)$. As n is a finite integer, it is clear that $I(k)$ tends to infinity with the increment of k. On the other hand, from Eq. (4.A.2) and Lemma 2.6 we have that

$$P(k) = \underbrace{\mathbf{H}_m\left(\Phi^n, \mathbf{H}_m\left(\Phi^n, \cdots \mathbf{H}_m\left(\Phi^n,\right.\right.\right.}_{I(k)\ \text{times}} \mathbf{H}_m\left(\Phi^{r(k)}, P(0)\right)\Big)\cdots\Big)\Big). \tag{4.A.7}$$

Recall that the mapping $X \rightarrow \mathbf{H}_m(\Phi^n, X)$ is strictly contractive for an arbitrary positive integer n. It can therefore be declared that, for arbitrary positive definite matrices X and Y,

$$\delta\left\{\mathbf{H}_m\left(\Phi^k, X\right), \mathbf{H}_m\left(\Phi^k, Y\right)\right\} < \rho^{I(k)}\delta(X, Y). \tag{4.A.8}$$

Hence

$$\lim_{k\to\infty}\delta\left\{\mathbf{H}_m\left(\Phi^k, X\right), \mathbf{H}_m\left(\Phi^k, Y\right)\right\} = 0. \tag{4.A.9}$$

This completes the proof. □

4.A.2 Proof of Theorem 4.3

To prove Theorem 4.3, we first discuss some related properties of the maximal singular value of a lower block triangular matrix.

Lemma 4.A.1. *Assume that A_{11}, A_{21}, and A_{22} are three matrices with compatible dimensions. Construct the lower block triangular matrix $A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$. Then $\bar{\sigma}(A) \leq \sqrt{\bar{\sigma}^2(A_{11}) + \bar{\sigma}^2(A_{21}) + \bar{\sigma}^2(A_{22})}$.*

Proof of Lemma 4.A.1. Let γ be an arbitrary nonnegative number satisfying $\gamma > \bar{\sigma}(A)$. Then, from the definition of the maximal singular value of a matrix we can straightforwardly declare that $\gamma^2 I - AA^T > 0$. According to the lower block triangular structure of the matrix A, this inequality is equivalent to

$$\begin{bmatrix} \gamma^2 I - A_{11}A_{11}^T & -A_{11}A_{21}^T \\ -A_{21}A_{11}^T & \gamma^2 I - A_{21}A_{21}^T - A_{22}A_{22}^T \end{bmatrix} > 0. \tag{4.A.10}$$

Here, to simplify expressions, the dimensions of the associated identity matrices are not included, as their actual dimensions are clear from the contents, and confusions can hardly be caused. It is worth mentioning that these identity matrices usually have a different size. These declarations remain valid for the remaining statements in this appendix.

On the basis of the well-known Schur complement theorem [8], the last inequality can be further expressed as

$$\gamma^2 I - A_{11}A_{11}^T > 0, \qquad \gamma^2 I - A_{22}A_{22}^T > 0, \tag{4.A.11}$$

$$\gamma^2 I - A_{21}A_{21}^T - A_{22}A_{22}^T - A_{21}A_{11}^T(\gamma^2 I - A_{11}A_{11}^T)^{-1}A_{11}A_{21}^T > 0. \tag{4.A.12}$$

On the other hand, direct algebraic manipulations show that

$$\begin{aligned} & \gamma^2 I - A_{21}A_{21}^T - A_{22}A_{22}^T - A_{21}A_{11}^T(\gamma^2 I - A_{11}A_{11}^T)^{-1}A_{11}A_{21}^T \\ = \; & \gamma^2 I - A_{22}A_{22}^T - \gamma^2 A_{21}(\gamma^2 I - A_{11}A_{11}^T)^{-1}A_{21}^T \\ \geq \; & \gamma^2 I - \bar{\sigma}^2(A_{22})I - \gamma^2 A_{21}(\gamma^2 I - \bar{\sigma}^2(A_{11})I)^{-1}A_{21}^T \\ \geq \; & \frac{1}{\gamma^2 - \bar{\sigma}^2(A_{11})}\left[(\gamma^2 - \bar{\sigma}^2(A_{11}))(\gamma^2 - \bar{\sigma}^2(A_{22}))I - \gamma^2\bar{\sigma}^2(A_{21})I\right] \\ \geq \; & \frac{\left(\gamma^2 - \frac{\bar{\sigma}^2(A_{11})+\bar{\sigma}^2(A_{21})+\bar{\sigma}^2(A_{22})}{2}\right)^2 - \frac{(\bar{\sigma}^2(A_{11})+\bar{\sigma}^2(A_{21})+\bar{\sigma}^2(A_{22}))^2}{4}}{\gamma^2 - \bar{\sigma}^2(A_{11})} I. \end{aligned} \tag{4.A.13}$$

This means that when the nonnegative number γ satisfies $\gamma > \sqrt{\bar{\sigma}^2(A_{11})+\bar{\sigma}^2(A_{21})+\bar{\sigma}^2(A_{22})}$, it must also satisfy $\gamma^2 I - A_{21}A_{21}^T - A_{22}A_{22}^T - A_{21}A_{11}^T(\gamma^2 I - A_{11}A_{11}^T)^{-1}A_{11}A_{21}^T > 0$. This completes the proof. □

Using this inequality, Theorem 4.3 is established.

From the definitions of the matrices $A_f(k)$ and A_f and from the exponential convergence of the matrix $P(k)$ to the matrix P, we can directly declare that there exist a finite nonnegative number M_1 and $\rho_1 \in [0, 1)$ such that

$$\bar{\sigma}(A_f(k) - A_f) \leq M_1\rho_1^k, \quad \forall k \geq 0. \tag{4.A.14}$$

Note that the matrix A_f is stable. We can claim from the Lyapunov stability theory that, for an arbitrary $\mu > 1$, there exists a positive definite matrix V such that

$$V - A_f V A_f^T = \mu I. \tag{4.A.15}$$

Denote the deviations of the matrix $A_f(k)$ from its steady value A_f by $\Delta(k)$, that is, $\Delta(k) = A_f(k) - A_f$. Then, on the basis of Eqs. (4.A.14) and (4.A.15), we obtain the following inequality:

$$\begin{aligned} V - A_f(k) V A_f^T(k) &= V - (A_f + \Delta(k)) V (A_f + \Delta(k))^T \\ &= \mu I - \Delta(k)(V A_f^T) - (V A_f^T)^T \Delta(k)^T - \Delta(k) V \Delta^T(k) \\ &\geq \mu I - [I + (\Delta(k) V A_f^T)(\Delta(k) V A_f^T)^T] - \Delta(k) V \Delta^T(k) \\ &= (\mu - 1) I - \Delta(k)[V + V A_f^T A_f V] \Delta^T(k) \\ &\geq (\mu - 1) I - \bar{\sigma}^2(\Delta(k)) \bar{\sigma}(V + V A_f^T A_f V) I \\ &\geq [\mu - 1 - \bar{\sigma}(V + V A_f^T A_f V) M_1^2 \rho_1^{2k}] I. \end{aligned} \tag{4.A.16}$$

This implies that there exists a finite integer N_1 such that, for each $k \geq N_1$,

$$\bar{\sigma}\left(V^{-1/2} A_f(k) V^{1/2}\right) < 1. \tag{4.A.17}$$

Define the scalar

$$\rho_2 = \sup_{k \geq N_1} \bar{\sigma}(V^{-1/2} A_f(k) V^{1/2}).$$

Inequality (4.A.17) implies that $0 < \rho_2 < 1$. On the other hand, note that, for arbitrary integers k_1 and k_2 satisfying $k_2 \geq k_1 \geq 0$, we have

$$\prod_{i=k_2}^{k_1} A_f(k) = V^{1/2} \left(\prod_{i=k_2}^{k_1} V^{-1/2} A_f(i) V^{1/2} \right) V^{-1/2}. \tag{4.A.18}$$

Based on this relation and the definition of ρ_2 and on the inequality $\bar{\sigma}(AB) \leq \bar{\sigma}(A)\bar{\sigma}(B)$, which is well known in linear algebra [8], we can prove that there exists a positive finite number M_2 such that

$$\bar{\sigma}\left(\prod_{i=k_2}^{k_1} A_f(i) \right) \leq M_2 \rho_2^{k_2 - k_1}. \tag{4.A.19}$$

Let $\mathcal{E}$ and $\mathcal{U}$ denote respectively the set of parametric modeling errors and the plant input vectors. Then by the adopted assumptions both these two sets are bounded. On the other hand, from the exponential stability of the plant we have that there exist a finite nonnegative number M_3 and ρ_3 belonging to $[0,\ 1)$ such that, for every integer pair $(k_1,\ k_2)$ satisfying $k_2 \geq k_1 \geq 0$ and each parametric error vector $\varepsilon(k) \in \mathcal{E}$, we have the following inequality:

$$\bar{\sigma}\left(\prod_{i=k_2}^{k_1} A(i, \varepsilon(i))\right) \leq M_3 \rho_3^{k_2-k_1}. \tag{4.A.20}$$

To avoid awkward expressions, in the following discussions, $\prod_{i=s_2}^{s_1} X_i$ is defined as the identity matrix when $s_2 < s_1$. With straightforward matrix multiplications and mathematical inductions, we can prove by the definition of the matrix $\Theta(k, \varepsilon(k), \varepsilon(k+1))$ that, for two arbitrary integers k_1 and k_2 satisfying $k_2 \geq k_1 \geq 0$,

$$\begin{aligned}
&\prod_{i=k_2}^{k_1} \Theta(i, \varepsilon(i), \varepsilon(i+1)) \\
&= \prod_{i=k_2}^{k_1} \begin{bmatrix} A(i, \varepsilon(i)) & 0 \\ F(i, \varepsilon(i), \varepsilon(i+1)) & A_f(i) \end{bmatrix} \\
&= \begin{bmatrix} \prod_{i=k_2}^{k_1} A(i, \varepsilon(i)) & 0 \\ \sum_{j=k_1}^{k_2} \left\{ \left(\prod_{i=k_2}^{j+1} A_f(i)\right) F(i, \varepsilon(i), \varepsilon(i+1)) \left(\prod_{i=j-1}^{k_1} A(i, \varepsilon(i))\right) \right\} & \prod_{i=k_2}^{k_1} A_f(i) \end{bmatrix}.
\end{aligned} \tag{4.A.21}$$

Note that $\bar{\sigma}(E + F) \leq \bar{\sigma}(E) + \bar{\sigma}(F)$ and $\bar{\sigma}(AB) \leq \bar{\sigma}(A)\bar{\sigma}(B)$ [8]. From Eqs. (4.A.19) and (4.A.20) we have that for all the feasible modeling errors and each integer pair $(k_1,\ k_2)$ satisfying $k_2 \geq k_1 \geq 0$,

$$\begin{aligned}
&\bar{\sigma}\left(\sum_{j=k_1}^{k_2} \left\{ \left(\prod_{i=k_2}^{j+1} A_f(i)\right) F(i, \varepsilon(i), \varepsilon(i+1)) \left(\prod_{i=j-1}^{k_1} A(i, \varepsilon(i))\right) \right\}\right) \\
&\leq \sum_{j=k_1}^{k_2} \left\{ \bar{\sigma}\left(\prod_{i=k_2}^{j+1} A_f(i)\right) \bar{\sigma}(F(i, \varepsilon(i), \varepsilon(i+1))) \bar{\sigma}\left(\prod_{i=j-1}^{k_1} A(i, \varepsilon(i))\right) \right\} \\
&\leq \sum_{j=k_1}^{k_2} \left\{ (M_2 \rho_2^{k_2-j-1}) \sup_{k \geq 0} \sup_{\varepsilon(i), \varepsilon(i+1) \in \mathcal{E}} \bar{\sigma}(F(i, \varepsilon(i), \varepsilon(i+1)))(M_3 \rho_3^{j-k_1-1}) \right\}
\end{aligned}$$

$$\begin{aligned} &\leq \sum_{j=k_1}^{k_2} \left\{ \left(M_2 M_3 \sup_{k \geq 0} \sup_{\varepsilon(i), \varepsilon(i+1) \in \mathcal{E}} \bar{\sigma}(F(i, \varepsilon(i), \varepsilon(i+1))) \right) (\rho_4^{k_2-j-1} \rho_4^{j-k_1-1}) \right\} \\ &= (k_2 - k_1 + 1) M_4 \rho_4^{k_2-k_1}, \end{aligned} \tag{4.A.22}$$

where ρ_4 and M_4 are respectively defined as

$$\rho_4 = \max\{\rho_2,\ \rho_3\}, \quad M_4 = M_2 M_3 \sup_{i \geq 0} \sup_{\varepsilon(i), \varepsilon(i+1) \in \mathcal{E}} \bar{\sigma}(F(i, \varepsilon(i), \varepsilon(i+1))).$$

From the definition of the matrix $F(i, \varepsilon(i), \varepsilon(i+1))$, the assumption that every element of the system matrices is differentiable with respect to each parametric modeling error, and from the assumption that each parametric modeling error is magnitude bounded it is clear that at every sampled time instant k, each element of the matrix $F(i, \varepsilon(i), \varepsilon(i+1))$ is also magnitude bounded, provided that both vectors $\varepsilon(i)$ and $\varepsilon(i+1)$ belong to the set $\mathcal{E}$. Therefore, the number M_4 has a nonnegative finite value.

On the basis of Lemma 4.A.1 and Eqs. (4.A.20)–(4.A.22), we can declared that, for arbitrary $k_2 \geq k_1 \geq 0$ and $\varepsilon(i),\ \varepsilon(i+1) \in \mathcal{E}$,

$$\begin{aligned} &\prod_{i=k_2}^{k_1} \Theta(i, \varepsilon(i), \varepsilon(i+1)) \\ &\leq \sqrt{\bar{\sigma}^2 \left(\prod_{i=k_2}^{k_1} A(i, \varepsilon(i)) \right) + \bar{\sigma}^2 \left(\sum_{j=k_1}^{k_2} \left\{ \left(\prod_{i=k_2}^{j+1} A_f(i) \right) F(i, \varepsilon(i), \varepsilon(i+1)) \left(\prod_{i=j-1}^{k_1} A(i, \varepsilon(i)) \right) \right\} \right) + \bar{\sigma}^2 \left(\prod_{i=k_2}^{k_1} A_f(i) \right)} \\ &\leq \sqrt{(M_3 \rho_3^{k_2-k_1})^2 + (M_2 \rho_2^{k_2-k_1})^2 + [(i_2 - k_1 + 1) M_3 \rho_4^{k_2-k_1}]^2} \\ &\leq \sqrt{[M_3^2 + M_2^2 + (i_2 - k_1 + 1)^2 M_3^2] \rho^{2(k_2-k_1)}} \\ &= \rho^{k_2-k_1} \sqrt{M_1^2 + M_2^2 + (k_2 - k_1 + 1)^2 M_3^2}\ , \end{aligned} \tag{4.A.23}$$

where $\rho = \max\{\rho_2,\ \rho_3,\ \rho_4\}$.

In addition, iterative utilizations of Eqs. (4.65) and (4.66) show that

$$\begin{aligned} \mathbf{E}\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} = &\left(\prod_{i=k}^{0} \Theta(i, \varepsilon(i), \varepsilon(i+1)) \right) \mathbf{E}\begin{bmatrix} x(0) \\ \hat{x}(0) \end{bmatrix} \\ &+ \sum_{i=0}^{k} \left(\prod_{j=k}^{i+1} \Theta(j, \varepsilon(j), \varepsilon(j+1)) \right) \Gamma(i, \varepsilon(i), \varepsilon(i+1)) \begin{bmatrix} u(i) \\ u(i+1) \end{bmatrix}, \end{aligned} \tag{4.A.24}$$

$$\mathbf{Cov}\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} = \left(\prod_{i=k}^{0}\Theta(i,\varepsilon(i),\varepsilon(i+1))\right)\mathbf{Cov}\begin{bmatrix} x(0) \\ \hat{x}(0) \end{bmatrix}\left(\prod_{i=k}^{0}\Theta^T(i,\varepsilon(i),\varepsilon(i+1))\right)^T$$
$$+\sum_{i=0}^{k}\left(\prod_{j=k}^{i+1}\Theta(j,\varepsilon(j),\varepsilon(j+1))\right)\Phi(i,\varepsilon(i),\varepsilon(i+1))$$
$$\times\begin{bmatrix} Q(i) & 0 \\ 0 & R(i+1) \end{bmatrix}\Phi^T(i,\varepsilon(i),\varepsilon(i+1))$$
$$\times\left(\prod_{j=k}^{i+1}\Theta(j,\varepsilon(j),\varepsilon(j+1))\right)^T. \tag{4.A.25}$$

As every element of the system matrices is assumed to be differentiable with respect to each parametric modeling error and each parametric modeling error is assumed to be magnitude bounded, we can directly declare by the definitions of the matrices $\Gamma(k,\varepsilon(k),\varepsilon(k+1))$ and $\Phi(k,\varepsilon(k),\varepsilon(k+1))$ that their elements are also magnitude bounded at every sampled time instant k. The proof can now be completed by combining together Eqs. (4.A.23)–(4.A.25) and noting that the estimation error of the robust state estimator at the sampled time instant k is equal to $\hat{x}(k)-x(k)$. □

References

[1] T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[2] R.E. Kalman, A new approach to linear filtering and prediction problems, Transactions of the American Society of Mechanical Engineer – Journal of Basic Engineering (Series D) 82 (1960) 34–45.

[3] D.G. Luenberger, Observing the states of a linear system, IEEE Transactions on Military Electronics 8 (1964) 74–80.

[4] D.G. Luenberger, An introduction to observer, IEEE Transactions on Automatic Control 16 (1971) 596–602.

[5] D. Simon, Optimal State Prediction: Kalman, H_∞ and Nonlinear Approaches, Wiley-Interscience, John Wiley & Sons, Inc., Publication, Hoboken, New Jersey, USA, 2006.

[6] A.E. Bryson, Y.C. Ho, Applied Optimal Control: Optimization, Estimation and Control, Tarlor & Francis, New York, USA, 1975.

[7] F.Z. Zhang, Matrix Theory: Basic Results and Techniques, Springer, New York, 1999.

[8] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[9] P. Lancaster, M.T. Tismenetsky, The Theory of Matrices: With Applications, Academic, New York, 1985.

[10] A. Garulli, A. Vicino, G. Zappa, Conditional central algorithms for worst case set-membership identification and filtering, IEEE Transactions on Automatic Control 45 (2000) 14–23.

[11] T. Zhou, Sensitivity penalization based robust state estimation for uncertain linear systems, IEEE Transactions on Automatic Control 55 (2010) 1018–1024.

[12] T. Zhou, H.Y. Liang, On asymptotic behaviors of a sensitivity penalization based robust state estimator, Systems & Control Letters 60 (2011) 174–180.

[13] G.W. Stewart, Theory of the Combination of Observations Least Subject to Errors. (Translation of original works by C.F. Gauss: Theoria Combinationis Observationum Erroribus Minimis Obnoxiae), Classics in Applied Mathematics, SIAM, Philadelphia, USA, 1995.

[14] The National Academy of Engineering, U.S.A., https://www.nae.edu/Projects/Awards/DraperPrize.aspx, 2008. Unpublished manuscript.

[15] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M.I. Jordan, S.S. Sastry, Kalman filtering with intermittent observations, IEEE Transactions on Automatic Control 49 (2004) 1453–1461.

[16] T. Zhou, Robust recursive state estimation with random measurement droppings, IEEE Transactions on Automatic Control 61 (2016) 156–171.

[17] P. Bougerol, Kalman filtering with random coefficients and contractions, SIAM Journal on Control and Optimization 31 (1993) 942–959.

[18] A.H. Sayed, A framework for state-space estimation with uncertain models, IEEE Transactions on Automatic Control 46 (2001) 998–1013.

CHAPTER 5

State Estimation With Random Data Droppings

5.1 Introduction

The pioneering work [3] studies the optimal state estimation problem for a discrete-time linear stochastic system under the assumption that the raw measurements of the system are randomly dropped. The authors consider an estimation problem of a discrete-time linear stochastic system with random data droppings, which can be used to model either random communication link failures or sensor faults. A fundamental problem is how the random data droppings affect the optimal estimation performance. To formalize it, a binary random process is used to denote the data dropping process and derive an optimal recursive filter of the discrete-time linear stochastic system to minimize the mean square state estimation errors by using the Kalman filter technique. It turns out that the optimal filter resembles the standard Kalman filter of the same recursive structure and complexity and is named the intermittent Kalman filter (IKF) in the literature. Particularly, if the sensor measurement is available to the estimator, then the one-step state predictor is simply updated as in the Kalman filter. When the sensor measurement is dropped, that is, there is a measurement data dropping, the estimator becomes open-loop, and the predictor is not updated. Thus, the prediction error covariance matrix cannot be reduced, which is the key difference from the Kalman filter.

Clearly, the more data droppings, the more open loop involves, and the estimation error covariance matrix grows faster. A natural question is whether the optimal filter with data droppings will diverge or converge in an appropriate sense, which is also the focus of this chapter. By modeling the packet loss process as an independent and identically distributed (i.i.d.) Bernoulli process, there exists a critical packet loss rate above which the mean state estimation error covariance matrices will diverge [3]. However, they are unable to exactly quantify the critical loss rate for general systems except providing its lower and upper bounds, which are attainable under some special cases, for example, the lower bound is tight if the observation matrix is invertible. A less restrictive condition is provided in [17], where invertibility on the observable subspace is required. In [4] the loss rate for a wider class of systems is explicitly characterized, including second-order systems and the so-called nondegenerate higher-order systems. A remarkable discovery in [4] is that there are counterexamples of second-order systems for which the lower bound given by [3] is not tight.

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00005-1

In [4], plant models are assumed to be precisely known. In actual engineering applications, however, model errors are usually unavoidable, which may appreciably deteriorate estimation accuracies of an estimator [5–7,27]. In this chapter, we also discuss the intermittent Kalman filter under switching sensors [2].

The rest of the chapter is organized as follows. In Section 5.2, we discuss the basics of the Kalman filtering problem with intermittent observations, where the raw sensor measurements are directly sent to the estimator via the lossy channels. The optimal estimate is then given by the intermittent Kalman filter (IKF). The mean square stability of the IKF is also discussed. In Section 5.3, we consider the case of using switching sensors over the unreliable networks and show how the switchings affect the mean square stability of the IKF. In Section 5.4, we provide a linear coding approach to reduce the effects of packet droppings on the mean square stability of the IKF. Finally, we also study the parametric errors on the state estimation problems with packet droppings.

5.2 Intermittent Kalman Filtering (IKF)

Consider the discrete-time stochastic linear system

$$\left\{\begin{array}{rcl} x_{k+1} & = & Ax_k + w_k, \\ y_k & = & Cx_k + v_k, \end{array}\right. \tag{5.1}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^\ell$ are vector state and measurement, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^\ell$ are white Gaussian noises with zero means and covariance matrices $Q > 0$ and $R > 0$, respectively. C is of full row rank, that is, $rank(C) = l \leq n$. The initial state x_0 is assumed to be a random Gaussian vector with mean $\hat{x}_0$ and covariance matrix $P_0 > 0$. Moreover, $w_k,\ v_k$, and x_0 are mutually independent.

We focus on an estimation framework where the sensor measurements of the system are transmitted to an estimator via an unreliable communication channel. Due to random fading and/or congestion of the communication channel, packets may be lost while in transit through the channel. The data loss dropping is modeled by a binary stochastic process $\{\gamma_k\}_{k\geq 0}$. Furthermore, assume that $\{\gamma_k\}_{k\geq 0}$ contains no information of the system, and is independent of the system evolution. Let $\gamma_k = 1$ indicate that the packet containing the measurement information has been successfully delivered to the estimator, whereas $\gamma_k = 0$ corresponds to the measurement data dropping. We further always assume that γ_k is a binary independent and identically distributed process and $Pr\{\gamma_k = 1\} = p$ when studying the stability of the corresponding estimator.

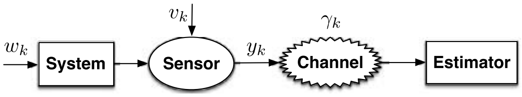

Figure 5.1: Network configuration.

5.2.1 The IKF Algorithm

We first discuss the case with raw measurement transmission; see Fig. 5.1. In this case, if there is a measurement loss $\gamma_k = 0$, then y_k is unavailable to the estimator. Otherwise, y_k is successfully transmitted to the estimator. The objective is to obtain an optimal estimator and characterize the effect of packet droppings on the estimator. To this purpose, let

$$\mathcal{F}_k \triangleq \{y_i\gamma_i, \gamma_i, i \leq k\}$$

be an increasing sequence of sets generated by the information received by the estimator up to time k, that is, all events that are generated by the random variables $\{y_i\gamma_i, \gamma_i, i \leq k\}$.

We are concerned with the minimum mean square error (MMSE) predictor and estimator with intermittent measurements given by the conditional expectations

$$\hat{x}_{k|k-1} = \mathbb{E}[x_k|\mathcal{F}_{k-1} \quad \text{and} \quad \hat{x}_{k|k} = \mathbb{E}[x_k|\mathcal{F}_k], \tag{5.2}$$

respectively. The associated estimation error covariance matrices are then defined as

$$P_{k|k} = \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T|\mathcal{F}_k] \tag{5.3}$$

and

$$P_{k+1|k} = \mathbb{E}[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^T|\mathcal{F}_k],$$

where A^T is the transpose of A.

The measurement noise v_k is defined in the following way:

$$p(v_k|\gamma_k) = \begin{cases} \mathcal{N}(0, R), & \gamma_k = 1, \\ \mathcal{N}(0, \sigma^2 I), & \gamma_k = 0, \end{cases} \tag{5.4}$$

for some σ^2. Therefore, the variance of the measurement at time k is R if $\gamma_k = 1$ and $\sigma^2 I$ otherwise. In the absence of measurement, it corresponds to the limiting case of $\sigma \to \infty$ [3].

Let $\hat{y}_{k|k-1} = \mathbb{E}[y_k|\mathcal{F}_{k-1}]$, $\tilde{y}_k = y_k - \hat{y}_{k|k-1}$, and $\tilde{x}_k = x_k - \hat{x}_{k|k-1}$, it is easy to see that

$$\mathbb{E}[\tilde{y}_k\tilde{x}_k^T|\mathcal{F}_{k-1}] = CP_{k|k-1}, \tag{5.5}$$

$$\mathbb{E}[\tilde{y}_k \tilde{y}_k^T | \mathcal{F}_{k-1}, \gamma_k] = C P_{k|k-1} C^T + \gamma_k R + (1-\gamma_k)\sigma^2 I. \tag{5.6}$$

This implies that

$$cov(x_k, y_k | \mathcal{F}_{k-1}, \gamma_k) = \begin{bmatrix} P_{k|k-1} & P_{k|k-1} C^T \\ C P_{k|k-1} & C P_{k|k-1} C^T + \gamma_k R + (1-\gamma_k)\sigma^2 I \end{bmatrix}. \tag{5.7}$$

Hence, the measurement update in the Kalman filter is modified as follows:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + P_{k|k-1} C^T \left(C P_{k|k-1} C^T + \gamma_k R + (1-\gamma_k)\sigma^2 I \right)^{-1} (y_k - C\hat{x}_{k|k-1}), \tag{5.8}$$

$$P_{k|k} = P_{k|k-1} - P_{k|k-1} C^T \left(C P_{k|k-1} C^T + \gamma_k R + (1-\gamma_k)\sigma^2 I \right)^{-1} P_{k|k-1} C. \tag{5.9}$$

Taking the limit as $\sigma \to \infty$, these equations can be further reduced to

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k (y_k - C\hat{x}_{k|k-1}), \tag{5.10}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C P_{k|k-1}, \tag{5.11}$$

where $K_k = P_{k|k-1} C^T (C P_{k|k-1} C^T + R)^{-1}$.

Clearly, the major difference from the Kalman filter only appears in the case of missing measurement, that is, $\gamma_k = 0$, which corresponds exactly to propagation of the previous prediction. This also results in that both the estimator $\hat{x}_{k|k}$ and the conditional error covariance matrix $P_{k|k}$ are random. To examine the asymptotic behavior of $P_{k|k}$, the standard approach to analyze the Riccati recursion in the Kalman filter is no longer feasible. The randomness of $P_{k|k}$ also brings significant challenge to study its asymptotic performance.

In addition, the time update equations in the KF continue to hold, that is,

$$\begin{aligned} \hat{x}_{k+1|k} &= A\hat{x}_{k|k}, \\ P_{k+1|k} &= A P_{k|k} A^T + Q, \end{aligned} \tag{5.12}$$

and $\hat{x}_{0|-1} = \bar{x}_0$, $P_{0|-1} = P_0$. Overall, the IKF is provided in detail in Algorithm 5.2.1.

Algorithm 5.2.1. Intermittent Kalman filter (IKF)

1. **Initialization**: $\hat{x}_{0|-1} = \bar{x}_0$, $P_{0|-1} = P_0$.
2. **Measurement update**: *at time k, the IKF updates as follows:*

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k (y_k - C\hat{x}_{k|k-1}), \tag{5.13}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C P_{k|k-1}, \tag{5.14}$$

where $K_k = P_{k|k-1} C^T (C P_{k|k-1} C^T + R)^{-1}$.

3. **Time update**: *the IKF updates in the same way as the KF via*

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_{k|k},\\ P_{k+1|k} &= AP_{k|k}A^T + Q.\end{aligned} \tag{5.15}$$

5.2.2 Mean Square Stability of the IKF

Letting $P_k := P_{k|k-1}$, it is easy to obtain the following switching Riccati recursion:

$$P_{k+1} = AP_kA^T + Q - \gamma_{k+1}AP_kC^T(CP_kC^T + R)^{-1}C^TP_kA^T. \tag{5.16}$$

Definition 5.1. *The estimator $\hat{x}_{k|k}$ is said to be mean square stable if for any initial estimate,*

$$\sup_k \mathbb{E}[\|\hat{x}_{k|k} - x_k\|^2] < \infty. \tag{5.17}$$

Clearly, the mean square stability of the IKF is equivalent to $\sup_k \mathbb{E}[\|P_k\|] < \infty$. Then, there exists a sharp transition for the mean square stability of the IKF.

Theorem 5.1. *([3]) If $(A, Q^{1/2})$ is controllable, (A, C) is detectable, and A is unstable, then there exists $p_c \in [0, 1)$ such that*

$$\begin{aligned}&\forall p \in [0, p_c], \exists P_0 > 0 \text{ such that } \lim_{k\to\infty} \mathbb{E}[P_k] = \infty, \text{ and}\\ &\forall p \in (p_c, 1], \forall P_0 > 0, \ \lim_{k\to\infty} \mathbb{E}[P_k] \le C_{P_0},\end{aligned} \tag{5.18}$$

where C_{P_0} depends on the initial condition $P_0 \geq 0$.

Proof. We adopt the proof in [3]. First, we note that the two cases expressed by the theorem are indeed possible. If $\lambda = 1$, then the modified Riccati difference equation reduces to the standard Riccati difference equation, which is known to converge to a fixed point under the theorem hypotheses. Hence, the covariance matrix is always bounded in this case for any initial condition $P_0 \geq 0$. If $\lambda = 0$, then we reduce to open-loop prediction, and if the matrix is unstable, then the covariance matrix diverges for some initial condition P_0. Next, we show the existence of a single point of transition between the two cases. Fix $0 < p_1 \leq 1$ such that $\mathbb{E}_{p_1}[P_k]$ is bounded for any initial condition $P_0 \geq 0$. Then, for any $p_2 \geq p_1$, we have that $\mathbb{E}_{p_2}[P_k]$ is also bounded for all $P_0 \geq 0$. In fact, we have

$$\begin{aligned}\mathbb{E}_{p_1}[P_{k+1}] &= \mathbb{E}[\mathbb{E}_{p_1}[AP_kA^T + Q - \gamma_{k+1}AP_kC^T(CP_kC^T + R)^{-1}C^TP_kA^T|P_k]]\\ &= \mathbb{E}[AP_kA^T + Q - p_1AP_kC^T(CP_kC^T + R)^{-1}C^TP_kA^T|P_k]\\ &\geq \mathbb{E}[AP_kA^T + Q - p_2AP_kC^T(CP_kC^T + R)^{-1}C^TP_kA^T|P_k]\\ &= \mathbb{E}_{p_2}[P_{k+1}].\end{aligned} \tag{5.19}$$

We can now choose

$$p_c = \inf\{\lambda^* | \lambda > \lambda^* \Rightarrow \mathbb{E}_\lambda[P_k] \text{ is bounded } \forall P_0 \geq 0\}, \tag{5.20}$$

completing the proof. □

Then, the problem of interest is how to evaluate the critical value p_c.

Lemma 5.1. *Consider the following recursion:*

$$M_{k+1} = \quad (1-p)AM_kA^T + Q, \tag{5.21}$$

where $M_0 = P_0$. Then

$$M_k \leq \mathbb{E}[P_k], \forall k.$$

Proof. Clearly, $M_0 \leq P_0$. Assume that $M_t \leq \mathbb{E}[P_t]$ for all $t \leq k$. As $Q > 0$, we can easily verify that $P_k > 0$ for all k. Then, by the matrix inversion lemma [8] it follows that

$$(P_k^{-1} + C^T R^{-1} C)^{-1} = P_k - P_k C^T (C P_k C^T + R)^{-1} C^T P_k \geq 0. \tag{5.22}$$

This implies that $P_k \geq P_k C^T (C P_k C^T + R)^{-1} C^T P_k$ and

$$P_{k+1} \geq A P_k A^T + Q - \gamma_{k+1} A P_k A^T = (1-\gamma_{k+1}) A P_k A^T + Q. \tag{5.23}$$

Taking expectation on both sides and using the fact $\mathbb{E}[P_k] \geq M_k$, the rest of proof is trivial by induction. □

Lemma 5.2. *Assume that γ_k is a binary independent and identically distributed process and $Pr\{\gamma_k = 1\} = p$. Consider the following recursion:*

$$V_{k+1} = A V_k A^T + Q - p \cdot A V_k C^T (C V_k C^T + R)^{-1} C^T V_k A^T, \tag{5.24}$$

where $V_0 = P_0$. Then

$$\mathbb{E}[P_k] \leq V_k, \forall k. \tag{5.25}$$

Proof. Note that (5.16) can be rewritten as follows:

$$P_{k+1} = (1-\gamma_{k+1}) A P_k A^T + Q - \gamma_{k+1} A (P_k^{-1} + C^T R^{-1} C)^{-1} A^T, \tag{5.26}$$

which is clearly convex in P_k. Then, taking the expectation of both sides of (5.16) and using the Jensen inequality [8], we obtain that

$$\mathbb{E}[P_{k+1}] \leq A\mathbb{E}[P_k]A^T + Q - p \cdot A\mathbb{E}[P_k]C^T (C\mathbb{E}[P_k]C^T + R)^{-1} C^T \mathbb{E}[P_k] A^T.$$

Together with mathematical arguments, it is easy to complete the proof. □

In view of the last two lemmas, we obtain the following result.

Theorem 5.2. *Let*

$$\begin{aligned} \underline{p} &= \arg\min\{p | \textit{the recursion in (5.21) is convergent } \forall P_0 \geq 0\} = 1 - 1/\rho^2(A) \quad (5.27) \\ \overline{p} &= \arg\min\{p | \textit{the recursion in (5.24) is convergent } \forall P_0 \geq 0\}, \quad (5.28) \end{aligned}$$

where $\rho(A)$ is the spectral radius of the matrix A. Then $\underline{p} \leq p_c \leq \overline{p}$.

To numerically obtain $\overline{p}$, it is required to use the bio-section method to solve LMIs. The details are provided in [3]. There are also some other ways to exactly evaluate the critical value p_c by analyzing the observability of the stochastically time-varying linear systems; see [9] and references therein.

5.2.3 Weak Convergence of the IKF

We also examine the weak convergence of the random covariance matrices P_t in (5.16).

Definition 5.2. *The sequence of random matrices $\{P_t\}$ is said to converge weakly or in distribution to a random matrix P_∞ if*

$$\lim_{t\to\infty} F_t(P) = F(P) \tag{5.29}$$

for every positive definite matrix P at which F is continuous. Here F_t and F are the cumulative distribution functions of the random matrices P_t and P_∞, respectively.

Then, we have the following results.

Theorem 5.3. *([10]) Suppose that $(A, Q^{1/2})$ is controllable, (A, C) is detectable, and A is unstable.*

(a) If $p > 0$, then $\{P_t\}$ weakly converges to a random matrix P_∞ for any initial P_0.
(b) If $p > p_c$, then the corresponding invariant distribution F has a finite mean, that is,

$$\int_{S_+^n} P dF(P) < \infty,$$

where S_+^n is the set of all positive definite matrices, and p_c is given in Theorem 5.1.

Theorem 5.3 states that for stabilizable and detectable systems, the sequence of random matrices $\{P_t\}$ in (5.16) converges in distribution to a unique invariant distribution F, irrespective of the initial condition. In particular, one may operate below p_c and still converge to a unique invariant distribution. However, operating at $p < p_c$ may not guarantee that the corresponding invariant distribution F has a finite mean.

Moreover, the invariant distribution F satisfies a moderate deviations principle with good rate function. To elaborate it, let $\mu_p(\cdot)$ be the probability measure induced by P_∞, where $p > 0$ is explicitly given to show its dependence on the packet receive rate. Then we have the following result.

Theorem 5.4. *([10]) The family of invariant distributions $\mu_p(\cdot)$ converges weakly to the Dirac probability measure δ_{P^*}. In particular, for each $\epsilon > 0$, we have the following convergence rate asymptotics:*

$$\lim_{p \to 1^-} \sup \; -\frac{\ln(\mu_p(B_\epsilon^C(P^*)))}{1-p} \leq -1,$$

where $B_\epsilon^C(P^*)$ is the closed ball centered at P^* with radius ϵ.

More details on the characterization of the invariant distribution F can be found in [10,11].

5.3 IKF With Switching Sensors

The striking difference of this section lies in the use of periodically switching sensors in the networked systems. Sensors of different nature, bandwidth, accuracies, and noise levels usually have different performances in specific operating and/or environmental conditions. Thus, the use of different sensors may provide rich information to increase the estimation/control performance. This is particularly important in the situation where a single sensor may not be able to provide sufficient information to estimate the state of a dynamical system.

Specifically, we consider an estimation framework of a stochastic system over a lossy network under two periodically switching sensors. See the networked system in Fig. 5.2 for an illustration. Here the periodically switching sensors are used to observe the system and result in a switching system. It is well known that the stability analysis of a switching system is usually more involved than that of a time-invariant system [12]. From this perspective, the problem of filter stability involving switching sensors for data transmission over a lossy network is expected to be more complicated than that of a single sensor.

Here there are two switching sensors to cooperatively monitor the system. At each time, one of them takes a noisy measurement from the system by

$$y_k = C_{\sigma_k} x_k + v_{\sigma_k}, \tag{5.30}$$

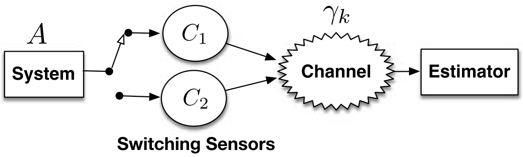

Figure 5.2: Networked systems over lossy channels under two periodically switching sensors.

where $\sigma_k \in \{1, 2\}$ represents the index of which sensor is activated to take measurement at time k, and v_{σ_k} is white Gaussian noise with zero mean and positive definite covariance matrix R_{σ_k}. Both C_1 and C_2 are of full row rank. The measurement matrix C_{σ_k} is now time-varying, which is used to alleviate the working load of one sensor for the purpose of prolonging the life time of the network or provide richer information for the estimator. As an initial attempt, we consider a periodically switching rule in this work. To be precise,

$$\sigma_k = \begin{cases} 1 & \text{if } k \text{ is odd,} \\ 2 & \text{if } k \text{ is even.} \end{cases} \tag{5.31}$$

Then, the associated measurement update is given by

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k (y_k - C_{\sigma_k} \hat{x}_{k|k-1}), \tag{5.32}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C_{\sigma_k} P_{k|k-1}, \tag{5.33}$$

where the Kalman gain $K_k = P_{k|k-1} C_{\sigma_k}^T (C_{\sigma_k} P_{k|k-1} C_{\sigma_k}^T + R_{\sigma_k})^{-1}$, and we obtain the following switching Riccati recursion:

$$\begin{aligned} P_{k+1} &= A P_k A^T + Q - \gamma_k A P_k C_{\sigma_k}^T (C_{\sigma_k} P_k C_{\sigma_k}^T + R_{\sigma_k})^{-1} C_{\sigma_k} P_k A^T \\ &:= g_k(P_k, R_{\sigma_k}). \end{aligned}$$

In this section, we consider the stability of the Kalman filter using two periodically switching sensors without packet losses. This allows us to focus on the sole effect of lossy channels on the stability of the Kalman filter in the next section. We recall the following result [13].

Lemma 5.3. *$g_k(\cdot, \cdot)$ is monotonically increasing in both arguments in the sense that*

$$g_k(P_1, R) \preceq g_k(P_2, R), \ \forall P_1 \preceq P_2; \tag{5.34}$$

$$g_k(P, R_1) \preceq g_k(P, R_2), \ \forall R_1 \preceq R_2. \tag{5.35}$$

For convenience, we show that the sensor noise levels do not affect the stability analysis of the Kalman filter under two switching sensors. To this purpose, denote $R_M = R_1 + R_2$ and $R_m = \min\{\lambda_{\min}(R_1), \lambda_{\min}(R_2)\} \cdot I$, where $\lambda_{\min}(R_i) > 0$ is the minimum eigenvalue of R_i. Then it follows from the monotonicity of $g_k(\cdot, \cdot)$ and from $R_m \preceq R_{\sigma_k} \preceq R_M$ that

$$g_k(P_k, R_m) \preceq g_k(P_k, R_{\sigma_k}) \preceq g_k(P_k, R_M), \forall k \in \mathbb{N}. \tag{5.36}$$

This essentially implies that the time-varying property of R_{σ_k} does not affect the stability analysis of P_k. It is also known from [4] that the stability conditions of both $P_{k+1} = g_k(P_k, R_m)$ and $P_{k+1} = g_k(P_k, R_M)$ are the same. Thus, there is no loss of generality to assume that $R_1 = R_2 = R$. This implies that the new challenge solely lies in time-varying nature of the observation matrix C_{σ_k}.

It should be noted that the stability analysis of a time-varying system is usually much more involved than that of a time-invariant system. Since the focus of this work is on quantifying the effect of the lossy network on the stability of the Kalman filter, we first derive the stability condition of the Kalman filter without packet losses, which corresponds to $\gamma_k = 1$ for all $k \in \mathbb{N}$.

By [14] a necessary and sufficient condition for the stability of the Kalman filter without packet losses is that (A, C_{σ_k}) is *uniformly detectable*. This requires the unstable modes of the system to be *uniformly observable* since all the state variables associated with the stable modes of the system will be exponentially stable in the mean square sense. For this purpose, we only need to focus on the state subspace corresponding to unstable modes. Hence, it is sensible to make the following assumption.

Assumption 5.1. *All the eigenvalues of A lie outside or on the unit circle.*

Then (A, C_{σ_k}) is required to be uniformly observable under Assumption 5.1 for the stability of the Kalman filter without packet losses, that is, there exist a positive integer h and positive numbers $\beta_0 > \alpha_0 > 0$ such that

$$\beta_0 I \succeq \sum_{i=k}^{k+h} (A^{i-k})^T C_{\sigma_i}^T C_{\sigma_i} A^{i-k} \succeq \alpha_0 I \succ 0, \forall k \in \mathbb{N}.$$

This uniform observability condition can be further simplified as stated in the following result.

Lemma 5.4. *The system (A, C_{σ_k}) with σ_k given in (5.31) is* uniformly observable *if and only if both*

$$\left(A^2, \begin{bmatrix} C_1 \\ C_2 A \end{bmatrix}\right) \text{ and } \left(A^2, \begin{bmatrix} C_1 A \\ C_2 \end{bmatrix}\right) \tag{5.37}$$

are observable.

Moreover, if A is nonsingular, then the observability properties of the systems

$$\left(A^2, \begin{bmatrix} C_1 \\ C_2 A \end{bmatrix}\right) \text{ and } \left(A^2, \begin{bmatrix} C_1 A \\ C_2 \end{bmatrix}\right) \tag{5.38}$$

are equivalent.

Proof. The first part directly follows from the definition of observability [15]. We only need to elaborate the second part. For notational simplicity, denote $\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ as $[C_1; C_2]$. Let the observability test matrices be $\mathcal{C}_1 = [C_1; C_2A; \ldots; C_1A^{2(n-1)}; C_2A^{2n-1}]$ and $\mathcal{C}_2 = [C_1A; C_2; \ldots; C_1A^{2n-1}; C_2A^{2(n-1)}]$. Considering $\mathcal{C}_1$ and $\mathcal{C}_2A$, it is clear that the rows of both matrices associated with C_2 are the same. By the Cayley–Hamilton theorem there exist $a_i \in \mathbb{R}$ such that $A^{2n} = a_0I + a_1A^2 + \cdots + a_{n-1}A^{2(n-1)}$. Premultiplying both sides of the equality by C_1, it follows that the last row of $\mathcal{C}_2A$ associated with C_1 can be linearly represented by the rows of $\mathcal{C}_1$. This further implies that each row of $\mathcal{C}_2A$ can be represented by the rows of $\mathcal{C}_1$. Hence, $rank(\mathcal{C}_2A) \leq rank(\mathcal{C}_1)$. Similarly, we can argue that $rank(\mathcal{C}_1A) \leq rank(\mathcal{C}_2)$. Since A is nonsingular, we obviously have that $rank(\mathcal{C}_1) = rank(\mathcal{C}_1A)$ and $rank(\mathcal{C}_2) = rank(\mathcal{C}_2A)$. Combing the preceding, we obtain that $rank(\mathcal{C}_1) = rank(\mathcal{C}_2)$, which completes the proof. □

Thus, the uniform observability property of a periodically switching system is converted into that of two time-invariant systems, each of which is observed by two sensors at each time.

In general, the nonsingularity assumption on A is mild, for example, it holds for all systems satisfying Assumption 5.1. By Lemma 5.4 we focus on the system with the following observability property in this paper since it is the basic requirement for the stability of the Kalman filter without packet losses under Assumption 5.1.

Assumption 5.2. *If $C = [C_1A; C_2]$, then the system (A^2, C) is observable.*

Remark 5.1. *By the PBH test [15] the observability of $\left(A^2, C\right)$ implies that (A, C) is observable, whereas the observability of (A, C) usually does not imply that $\left(A^2, C\right)$ is observable; take, for instance, $A = diag(1, -1)$ and $C = [1, 1]$. This, together with Lemma 5.4, essentially indicates that using two sensors to observe the same system at each time requires a weaker condition for the stability of the Kalman filter than that of a periodically switching sensor at each time, which certainly is consistent with our intuition as the former case supplies more information than the later one. We also mention that the observability of $(A^2, [C_1; C_2])$ does not imply that of (A^2, C); for example, take $A = diag(1, -1)$, $C_1 = [1\ 1]$ and $C_2 = [1\ -1]$. It should be noted that both (A, C_1) and (A, C_2) are observable.*

5.3.1 Mean Square Stability

Now, we establish the network condition on the packet loss process γ_k for the mean square stability of the intermittent Kalman filter under two periodically sensors.

Theorem 5.5. *Consider the networked system in Fig. 5.2. A necessary condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ *is that* $(\rho(A))^2(1-p) < 1$, *where* $\lambda_{\max}$ *is the maximum eigenvalue in magnitude of* A.

In fact, this necessary condition has been derived by many authors [3,9,16,17] under a single sensor case and shown to be sufficient as well for some special cases. It is interesting to investigate whether this condition is sufficient under the present framework. For a time-invariant observation matrix, that is, $C_1 = C_2$, it is shown that the condition in Theorem 5.5 is also sufficient if C_1 is invertible on the observable subspace [17] or (A, C_1) is a nondegenerate system [9]. Note that a periodic switching between two stable subsystems may lead to an unstable system due to the destabilizing effect of the switching. For example, we can verify that the system $x_{k+1} = A_k x_k$ is internally unstable if

$$A_k = \frac{1}{8} \cdot \begin{bmatrix} 0 & 9 + 7 \cdot (-1)^k \\ 9 - 7 \cdot (-1)^k & 0 \end{bmatrix}$$

although A_k has all eigenvalues inside the unit circle for each k. This intuitively implies that the derivation of a sufficient condition for the filter stability is more involved under the time-varying observation matrices.

In the previous section, the stability condition of the Kalman filter using two periodically switching sensors can be lifted into that of a time-invariant system with two measurement sensors if there is no packet loss (cf. Lemma 5.4). This motivates us to check whether under i.i.d. packet losses, the problem under consideration can be converted into the stability analysis of the Kalman filter for a time-invariant system using two measurement sensors simultaneously over two independent lossy channels, each of which is subject to an i.i.d. packet loss process. It turns out to be positive. To elaborate it, we recall a result in [9].

Lemma 5.5. *[9] Let* $\mathcal{O} = \sum_{i=1}^{\infty} \gamma_i (A^{-i})^T C_{\sigma_i}^T C_{\sigma_i} A^{-i}$. *Under Assumption 5.1, there exist two positive numbers* α *and* β *such that*

$$\beta \cdot \mathbb{E}[\mathcal{O}^{-1}] \succeq \sup_{k\in\mathbb{N}} \mathbb{E}[P_{k|k}] \succeq \alpha \cdot \mathbb{E}[\mathcal{O}^{-1}]. \tag{5.39}$$

Then we obtain an interesting result on the equivalent stability property of the networked systems.

Theorem 5.6. *Consider the networked systems in Fig. 5.2 and Fig. 5.3. If A is nonsingular, then necessary and sufficient conditions for the stability of the corresponding estimators are the same.*

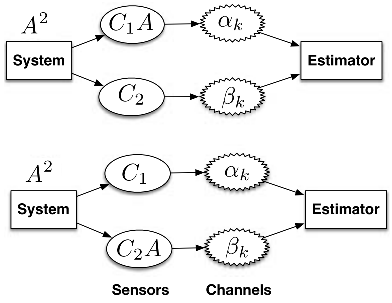

Figure 5.3: Networked systems over lossy channels. The open-loop system and the sensor measurement matrices are accordingly denoted above the blocks of systems and sensors. All the lossy channels are subject to the i.i.d. packet loss with the same statistical properties and mutually independent.

Proof. Noting that $P_{k|k-1} \succeq P_{k|k}$ and $P_{k+1|k} = AP_{k|k}A^T + Q$, it is obvious that $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ is equivalent to $\sup_{k\in\mathbb{N}} \mathbb{E}[P_{k|k}] \prec \infty$. By Lemma 5.5 the filter stability of the networked system in Fig. 5.2 is equivalent to

$$\mathbb{E}[\mathcal{O}^{-1}] \prec \infty. \tag{5.40}$$

Since γ_k is an i.i.d. process, $\mathcal{O}$ can be rewritten as follows:

$$\begin{aligned} \mathcal{O} &= \sum_{i=1}^{\infty} (A^{-2i})^T \begin{bmatrix} \gamma_{2i-1} A^T C_1^T & \gamma_{2i} C_2^T \end{bmatrix} \begin{bmatrix} C_1 A \\ C_2 \end{bmatrix} A^{-2i} \\ &\overset{d}{=} \sum_{i=1}^{\infty} (A^{-2i})^T \begin{bmatrix} \alpha_i A^T C_1^T & \beta_i C_2^T \end{bmatrix} \begin{bmatrix} C_1 A \\ C_2 \end{bmatrix} A^{-2i}, \end{aligned} \tag{5.41}$$

where $\overset{d}{=}$ means the equivalence in distribution of both sides, and α_i, β_i are two i.i.d. Bernoulli processes with the same statistics with γ_i, that is, $\mathbb{E}[\alpha_i] = \mathbb{E}[\beta_i] = p$. Thus, the filter stability of the networked systems in Fig. 5.2 is equivalent to that of the first networked system in Fig. 5.3. The rest of the proof is similarly established. □

This result can be immediately used to derive a sufficient condition for filter stability.

Theorem 5.7. *Consider the networked system in Fig. 5.2 satisfying Assumption 5.1. A sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ *is that*

$$\mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i (A^{-2i})^T \begin{bmatrix} A^T C_1^T & C_2^T \end{bmatrix} \begin{bmatrix} C_1 A \\ C_2 \end{bmatrix} A^{-2i} \right)^{-1} \prec \infty, \tag{5.42}$$

where ζ_i *is an i.i.d. process with* $\mathbb{P}\{\zeta_i = 1\} = p^2$.

Proof. By (5.41) define $\zeta_i = \min\{\alpha_i, \beta_i\}$, which is again an i.i.d. process with $\mathbb{P}\{\zeta_i = 1\} = \mathbb{P}\{\alpha_i = 1\}\mathbb{P}\{\beta_i = 1\} = p^2$. Then it follows that

$$\begin{bmatrix}\alpha_i A^T C_1^T & \beta_i C_2^T\end{bmatrix}\begin{bmatrix}C_1 A \\ C_2\end{bmatrix} \geq \zeta_i \cdot \begin{bmatrix}A^T C_1^T & C_2^T\end{bmatrix}\begin{bmatrix}C_1 A \\ C_2\end{bmatrix}. \tag{5.43}$$

Combing Lemma 5.5 and (5.41), we complete the proof. □

By Theorem 5.5 and 5.7 we obtain the simple sufficient condition for $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ for a certain class of systems.

Theorem 5.8. *Consider the networked system in Fig. 5.2 satisfying Assumption 5.1 and 5.2. If $C = [C_1; C_2]$ is of full row rank, then a sufficient condition for $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ is*

$$(\rho(A))^4 (1 - p^2) < 1. \tag{5.44}$$

Proof. Since $(\rho(A))^4 (1 - p) < 1$, there exists a sufficiently small $\epsilon > 0$ such that $(\rho(A) + \epsilon\|A\|)^4(1 - p^2) < 1$. Letting $\varrho = (\rho(A)) + \epsilon\|A\|$, it follows from Lemma 15 [9] that $\|A\|^k \leq M\rho^k$ for any $k \in \mathbb{N}$, where $M = \sqrt{n}(1 + 2/\epsilon)^{n-1}$.

If C is of full rank, then $C^T C \succ \lambda_{\min}(C^T C) \cdot I$, where $\lambda_{\min}(C^T C) > 0$ is the minimum eigenvalue of $C^T C$. This implies that

$$\mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i (A^{-2i})^T C^T C A^{-2i}\right)^{-1} \prec \frac{1}{\lambda_{\min}(C^T C)} \mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i (A^{-2i})^T A^{-2i}\right)^{-1}. \tag{5.45}$$

Note that $\mathbb{P}\{\zeta_1 = 0, \ldots, \zeta_k = 0, \ldots\} = \lim_{k\to\infty}(1-p^2)^k = 0$. Then, the sum $\sum_{i=0}^{\infty} \zeta_i (A^{-2i})^T A^{-2i}$ is positive definite with probability one.

Define the stopping time

$$\tau := \inf\{k \in \mathbb{N} | \zeta_k = 1\}, \tag{5.46}$$

whose probability mass distribution is given by $\mathbb{P}\{\tau = k + 1\} = p^2(1 - p^2)^k$. Hence,

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i (A^{-2i})^T A^{-2i}\right)^{-1} \leq \mathbb{E}[\zeta_\tau A^{2\tau}(A^{2\tau})^T] \\
&\leq (\mathbb{E}[\|A\|^{4\tau}])I \leq (M \cdot \mathbb{E}[\varrho^{4\tau}])I \\
&= M\varrho^4(1 - p^2)\sum_{k=0}^{\infty} \varrho^{4k}(1 - p^2)^k \cdot I,
\end{aligned}$$

which is finite since $\varrho^4(1 - p^2) < 1$. The rest of the proof follows from Theorem 5.7. □

Remark 5.2. *The main conservativeness of the sufficient condition lies in the use of Theorem 5.7. We use a simple example to illustrate the conservativeness when* $A = diag(\lambda_1, -\lambda_1)$ *and* $C_1 = C_2 = [1, 1]$. *By [9] a necessary and sufficient condition is that* $|\lambda_1|^2(1-p) < 1$, *which is still weaker than* $|\lambda_1|^4(1-p^2) < 1$. *Note that this approach does not fully exploit the system structure.*

Similarly, the following result is straightforward.

Theorem 5.9. *Consider the networked system in Fig. 5.2 satisfying Assumption 5.1 and 5.2. If either* C_1 *or* C_2 *is of full row rank, the a sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ *is*

$$(\rho(A))^4 (1-p) < 1. \tag{5.47}$$

Remark 5.3. *It should be noted that if* C_1 *is of full row rank and* C_2 *is a zero matrix, it follows from [9] that the condition in Theorem 5.9 is also sufficient.*

By Theorem 5.6 we obtain that $\mathbb{E}[\mathcal{O}^{-1}] \prec \infty$ is equivalent to the stability of the Kalman filter of the networked system

$$x_{k+1} = A^2 x_k + w_k, \tag{5.48}$$

which is observed by two sensors at each time with measurement equations

$$\begin{aligned} y_{k,1} &= C_1 A x_k + v_{k,1}, \\ y_{k,2} &= C_2 x_k + v_{k,2}, \end{aligned} \tag{5.49}$$

where $(A^2, [C_1 A; C_2])$ is observable, and $v_{k,1}$ and $v_{k,1}$ are two independent white Gaussian noises. The sensor measurements $y_{k,1}$ and $y_{k,2}$ are sent via two independent lossy channels to the estimator. See Fig. 5.3, where packet loss processes are modeled by two independent process α_k and β_k. Then, the corresponding Kalman filters of the networked systems in Fig. 5.3 require the same network condition for filter stability if A is nonsingular. Thus, it is sufficient to establish the network condition for the stability of the Kalman filter of the first networked system in Fig. 5.3.

In general, it is challenging to establish the necessary and sufficient condition for a general vector system. Nonetheless, the following procedures can help to reduce the complexity of the problem. Motivated by [9], we exploit the system structure under Assumption 5.2, which is classified as follows.

1. Both $(A^2, C_1 A)$ and (A^2, C_2) are observable.
2. Only one of $(A^2, C_1 A)$ and (A^2, C_2) is observable.
3. Neither $(A^2, C_1 A)$ nor (A^2, C_2) is observable, but $(A^2, [C_1 A; C_2])$ is observable.

In fact, we only need to consider Case 1 since the other two cases can be converted into the combination of Case 1 and that in [3,4]. We elaborate it in detail.

For Case 2, there is no loss in generality to assume that $(A^2, C_1 A)$ is observable but (A^2, C_2) is not observable. By the Kalman canonical decomposition [15] there exists a coordinate transformation such that (A^2, C) is transformed into the following structure:

$$A^2 = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix}, C_1 A = [C_{1,1}\ C_{1,2}], C_2 = [0\ C_{2,2}], \tag{5.50}$$

where $(A_{i,i}, C_{i,i})$ and $(A_{2,2}, C_{1,2})$ are observable. This means that the state variables corresponding to $A_{1,1}$ can only be observed by the sensor associated with the measurement matrix $C_1 A$. Then the filter stability analysis can be further reduced to the case of using only one sensor as in [3,4,9] and Case 1.

Theorem 5.10. *Under Case 2,* $\mathbb{E}[\mathcal{O}^{-1}] \prec \infty$ *if and only if* $\mathbb{E}[\mathcal{O}_1^{-1}] \prec \infty$ *and* $\mathbb{E}[\mathcal{O}_2^{-1}] \prec \infty$, *where*

$$\mathcal{O}_1 = \sum_{i=1}^{\infty} \alpha_i (A_{1,1}^{-i})^T C_{1,1}^T C_{1,1} A_{1,1}^{-i}$$

and

$$\mathcal{O}_2 = \sum_{i=1}^{\infty} (A_{2,2}^{-i})^T (\alpha_i C_{1,2}^T C_{1,2} + \beta_i C_{2,2}^T C_{2,2}) A_{2,2}^{-i}.$$

Proof. Since $\mathbb{E}[\mathcal{O}^{-1}] \prec \infty$, we can easily verify that $\mathbb{E}[\mathcal{O}_1^{-1}] \prec \infty$. Partition the state vector as $x_k = [x_{k,1}; x_{k,2}]$ in conformity with A^2. It follows that

$$\begin{aligned} x_{k+1,2} &= A_{2,2} x_{k,2} + w_{k,2}, \\ y_{k,1} &= C_{1,2} x_{k,2} + C_{1,1} x_{k,1} + v_{k,1}, \\ y_{k,2} &= C_{2,2} x_{k,2} + v_{k,2}. \end{aligned}$$

Since $\mathbb{E}[\mathcal{O}_1^{-1}] \prec \infty$, the estimation error covariance matrix corresponding to the state variables $x_{k,1}$ is stable. In particular, let $\tilde{x}_{k,i} = x_{k,i} - \hat{x}_{k,i}$; then $\sup_k \mathbb{E}[\tilde{x}_{k,1} \tilde{x}_{k,1}^T] \prec \infty$. Hence, we can use the following measurement to replace $y_{k,1}$:

$$y'_{k,1} = y_{k,1} - C_{1,1} \hat{x}_{k,1} = C_{1,2} x_{k,2} + v'_{k,1},$$

where $v'_{k,1} = C_{1,1} \tilde{x}_{k,1} + v_{k,1}$. Since $\mathbb{E}[\mathcal{O}_1^{-1}] \prec \infty$, it follows that $\sup_k \mathbb{E}[\tilde{x}_{k,2} \tilde{x}_{k,2}^T] \prec \infty$. Thus, the state vector of the following subsystems can be stably estimated:

$$x_{k+1,2} = A_{2,2} x_{k,2} + w_{k,2},$$

$$\begin{aligned} y'_{k,1} &= C_{1,2}x_{k,2} + v'_{k,1}, \\ y_{k,2} &= C_{2,2}x_{k,2} + v_{k,2}. \end{aligned}$$

By Lemma 5.5 we finally obtain that $\mathbb{E}[\mathcal{O}_2^{-1}] \prec \infty$.

The necessity can be similarly proved and is omitted. □

For Case 3, it follows from Proposition III.1 in [18] that there exists a coordinate transformation such that (A^2, C) has the structure either

$$A^2 = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix}, C_1 A = [0 \; C_{1,2}], C_2 = [C_{2,1} \; 0] \tag{5.51}$$

or

$$\begin{aligned} A^2 &= \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ 0 & A_{2,2} & A_{2,3} \\ 0 & 0 & A_{3,3} \end{bmatrix}, \\ C_1 A &= [0 \; C_{1,2} \; C_{1,3}], C_2 = [C_{2,1} \; 0 \; C_{2,3}]. \end{aligned} \tag{5.52}$$

The first structure indicates that the measurement matrix C_1 can only be used to observe the state subspace corresponding to $A_{2,2}$ and C_2 observes the complement state subspace, whereas in the second structure, both sensors can observe a common subspace corresponding to $A_{3,3}$. The decomposition in the first structure is very appealing as it helps us to convert the problems under consideration into the case with only an observation matrix, which has been considered in [3,4,9]. In the second structure, the common observable subspace associated with $A_{3,3}$ is observed by both sensors, which is the same as Case 1.

Hence, we only need to derive the network condition for stability of the Kalman filter over two independent lossy channels for the system satisfying that both $(A^2, C_1 A)$ and (A^2, C_2) are observable, which, jointly with Assumption 5.1, implies that (A^2, C_1) and $(A^2, C_2 A)$ are observable. To sum up, it is sufficient to focus on the systems satisfying the following:

Assumption 5.3. *Both (A^2, C_i) and $(A^2, C_i A)$ are observable for any $i \in \{1, 2\}$.*

5.3.2 Second-Order Systems

Together with [9], we are able to fully characterize the necessary and sufficient condition for the stability of the Kalman filter using two periodically switching sensors over a lossy network for the second-order system, that is, $A \in \mathbb{R}^{2\times 2}$.

Here we only focus on the second-order system satisfying the following condition.

Assumption 5.4. *$A = diag(\lambda_1, \lambda_2)$, where $\lambda_1 = \lambda_2 \exp(2\pi r I/d)$, $I^2 = -1$, and $d > r > 0$ are irreducible integers.*

Then a necessary and sufficient condition on the filter stability can be exactly given by single inequalities.

Theorem 5.11. *Consider the second-order networked system in Fig. 5.2 satisfying Assumptions 5.3–5.4. Then, a necessary and sufficient condition for $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ is*

$$|\lambda_1|^{2d/(d-c)}(1-p) < 1, \tag{5.53}$$

where c is determined by the number of invertible C_i and is given by

$$c = \begin{cases} 1 & \textit{if } \max\{rank(C_1), rank(C_2)\} = 1, \\ 0 & \textit{if } \min\{rank(C_1), rank(C_2)\} = 2, \\ 0.5 & \textit{otherwise.} \end{cases} \tag{5.54}$$

Sketch of Proof. (1) Case $c = 1$. If $C_1 = a \cdot C_2$, then consider the networked systems in Fig. 5.2, then it is equivalent to the system observed by one sensor. This is because the measurements from both sensors are the same except for a scaling by a, which is equivalent to the case without switching. Then, the rest of the proof follows from [9].

If $rank(C_1) = rank(C_2) = 1$, then consider the networked systems in Fig. 5.3. Let $\zeta_i = \max\{\alpha_i, \beta_i\}$ and define the stopping time

$$\tau_1 = \min\{k | \zeta_k = 1, k \geq 1\}.$$

Due to the independence of α_i and β_i, the probability mass distribution of τ is given by

$$\mathbb{P}\{\tau_1 = k\} = \begin{cases} 1-(1-p)^2 & \text{if } k = 1, \\ (1-p)^{2(k-1)}(1-(1-p)^2) & \text{if } k > 1. \end{cases} \tag{5.55}$$

By Assumption 5.3 it follows that $\lambda_1^2 \neq \lambda_2^2$. Together with Assumption 5.4, this implies that $2r/d$ is not an integer. Then, there exists a positive integer $r_1 < d$ such that $\lambda_1^2 = \lambda_2^2 \exp(2\pi r_1 I/d)$ and r_1, d are irreducible.

In view of the proof of Theorem 7 in [9], the necessary and sufficient condition becomes

$$\mathbb{E}[|\lambda_1|^{4\tau_1} 1_{\{\tau_1 \in \mathcal{S}_d\}}] < 1,$$

where $\mathcal{S}_d = \{kd | \forall k \in \mathbb{N}\}$, and 1_A is a standard indicator function for any set A. By (5.55) we can easily compute that

$$\mathbb{E}[|\lambda_1|^{4\tau_1} 1_{\{\tau_1 \in \mathcal{S}_d\}}] = \frac{1-(1-p)^2}{(1-p)^2} \frac{(|p_1|^2(1-p))^{2d}}{1-(|p_1|^2(1-p))^{2d}} < 1,$$

which is equivalent to $|\lambda_1|^{2d/(d-1)}(1-p) < 1$.

(2) Case $c = 0$. It is trivial, and the proof is omitted.

(3) Case $c = 0.5$. Without loss of generality, we assume that $rank(C_1) = 1$ and $rank(C_2) = 2$. Define the stopping time

$$\tau_2 = \min\{k|\alpha_k = 1, \beta_k = 0, \zeta_i = 0, \forall i \leq k-1\}. \tag{5.56}$$

By the independence of α_i and β_i the probability mass distribution of τ_2 is given by

$$\mathbb{P}\{\tau_2 = k\} = p(1-p)^{2k-1}. \tag{5.57}$$

Similarly, the necessary and sufficient condition becomes

$$\mathbb{E}[|\lambda_1|^{4\tau_2} 1_{\{\tau_2 \in \mathcal{S}_d\}}] < 1.$$

Then it follows that

$$\mathbb{E}[\lambda_1|^{4\tau_2} 1_{\{\tau_2 \in \mathcal{S}_d\}}] = \frac{p}{1-p} \frac{(|\lambda_1|^2(1-p))^{2d}}{1-(|\lambda_1|^2(1-p))^{2d}} < 1,$$

which is equivalent to $|\lambda_1|^{2d/(d-0.5)}(1-p) < 1$. □

Remark 5.4. *As remarked in [9], it is very difficult to establish a necessary and sufficient condition for the filter stability of the second-order system satisfying Assumption 5.4, whereas for the other cases, the condition becomes simple and is given by* $(\rho(A))^2\,(1-p) < 1$*, where* $\lambda_{\max}$ *is the largest open-loop pole in magnitude, the proof of which can be established by the same approach as in [9].*

5.3.3 Extension to Higher-Order Systems

The study of general vector systems is very challenging and left to our future work. However, if A is of a certain form, necessary and sufficient condition for the stability of the Kalman filter can be easily established.

Assumption 5.5. $A^{-1} = diag(J_1, \dots, J_m)$ *and* $rank(C_1) = rank(C_2) = 1$*, where* $J_i = \lambda_i^{-1} I_i + N_i \in \mathbb{R}^{n_i \times n_i}$ *and* $|\lambda_i| > |\lambda_{i+1}|$*;* I_i *is the identity matrix with compatible dimension, and the* (j,k)*th element of* N_i *is 1 if* $k = j+1$ *and 0 otherwise.*

Theorem 5.12. *Consider the networked system in Fig. 5.2 satisfying Assumptions 5.1–5.4. Then a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \prec \infty$ *is*

$$(\rho(A))^2\,(1-p) < 1. \tag{5.58}$$

Proof. It can be proved similarly as in Theorem 13 of [9]. □

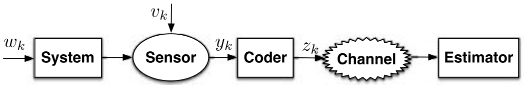

Figure 5.4: Network configuration with coded measurement transmission.

5.4 IKF With Coded Measurement Transmission

As opposite to the IKF, we study the scenario depicted in Fig. 5.4, where the signal z_k transmitted to the estimator is a coded version of y_k. We focus on a class of coders with finite storage memory, whose output can be computed recursively. More precisely, $z_k = \mathcal{E}_k(s_k, y_k)$, where the map $\mathcal{E}_k(\cdot, \cdot)$ denotes the coder at time k, and s_k denotes its internal state. Since z_k is to be transmitted to the estimator through an unreliable channel, the maximum information available to the estimator at time k is given by

$$\mathcal{F}_k = \{\gamma_i, z_i\gamma_i, i \le k\}.$$

In general, the higher the dimension of the coder output z_k, the larger the communication cost that will be incurred. Thus, the dimension of z_k should not be larger than that of y_k. Our goal is to design "good" coding methods to counteract the effect of random packet losses and to derive recursive formulas to compute the MMSE estimator.

5.4.1 Linear Temporal Coding

If the system output is directly transmitted, that is, $z_k = y_k$, then the MMSE estimator is computed by the IKF in Algorithm 5.2.1. Usually, the network requirement for stability of the IKF is strong. This motivates the idea of transmitting the output of the Kalman filter to the estimator [19], that is, $z_k = \mathbb{E}[x_k|y_1, \ldots, y_k]$.

The MMSE estimator is then given by

$$\hat{x}_{k|k} = \begin{cases} z_k & \text{if } \gamma_k = 1, \\ A\hat{x}_{k-1|k-1} & \text{if } \gamma_k = 0. \end{cases} \tag{5.59}$$

If (A, C) is observable, then a necessary and sufficient condition for the stability of the above filter is simply given by $(\rho(A))^2(1-p) < 1$. However, the dimension of the state estimate is generally much higher than that of y_k.

By the preceding there may exist a tradeoff between the effect of packet losses on the filter stability and communication resources. To this end, we study a linear temporal coding algorithm. Our proposed linear coding method is as follows. Take $\alpha_k^T = [\alpha_{k1}, \ldots, \alpha_{k(m-1)}, 1] \in$

$\mathbb{R}^{1\times m}$ (recall that $m = n - rank(C) + 1$). The coded output is given by

$$\begin{aligned} z_k &= y_k + \alpha_{k(m-1)} y_{k-1} + \cdots + \alpha_{k1} y_{k-m+1} \\ &= (\alpha_k^T \otimes I_q) col\{y_{k-m+1}, \ldots, y_k\} \in \mathbb{R}^q \end{aligned} \tag{5.60}$$

with the convention that $y_k = 0$ for $k < 0$, where $I_q \in \mathbb{R}^{q\times q}$ is the identity matrix, and $\otimes$ is the Kronecker product [8]. The design of $\{\alpha_k : k \in \mathbb{N}\}$ will be detailed later.

It is clear from (5.60) that the sequence $\{z_0, z_1, \ldots, z_k\}$ can be uniquely recovered from the sequence $\{y_0, y_1, \ldots, y_k\}$ for any $k \geq 0$. For this reason, the coded output is information preserving when there is no packet loss.

5.4.2 The MMSE Filter

It follows from (5.60) that the noise of z_k is correlated with $z_{k-1}, z_{k-2}, \ldots, z_{k-m+1}$. Hence, we cannot obtain an MMSE estimator by simply running a Kalman filter using z_k as the system output. To go around this, we define $\mu_k = col\{y_{k-m+1}, y_{k-m+2}, \ldots, y_{k-1}\}$ and

$$\mu_{k+1} = F\mu_k + Gy_k, \tag{5.61}$$

where $G = col\{0, 0, \ldots, 0, I_q\}$ and

$$F = \begin{bmatrix} 0 & I_{(m-1)q} \\ 0 & 0 \end{bmatrix}$$

with identity matrices $I_q \in \mathbb{R}^{q\times q}$ and $I_{(m-1)q} \in \mathbb{R}^{(m-1)q\times(m-1)q}$. Then, we can rewrite (5.1) and (5.60) as the augmented system

$$\begin{aligned} u_{k+1} &= \begin{bmatrix} A & 0 \\ GC & F \end{bmatrix} u_k + \begin{bmatrix} w_k \\ Gv_k \end{bmatrix}, \\ z_k &= H_k u_k + v_k. \end{aligned} \tag{5.62}$$

Clearly, the noise components in (5.62) are temporally independent. Hence, we can estimate an augmented state $u_{k+1} = [x_{k+1}^T, \mu_{k+1}^T]^T$ via a Kalman filter [20]. This leads to

$$\begin{aligned} \hat{u}_{k+1|k} &= \Phi\hat{u}_{k|k-1} + \gamma_k(\Phi\Sigma_{k|k-1}H_k^T + S_k)(H_k\Sigma_{k|k-1}H_k^T + R)^{-1}(z_k - H_k\hat{u}_{k|k-1}), \\ \Sigma_{k+1|k} &= \Phi\Sigma_{k|k-1}\Phi^T + \bar{Q} \\ &\quad - \gamma_k(\Phi\Sigma_{k|k-1}H_k^T + S_k)(H_k\Sigma_{k|k-1}H_k^T + R)^{-1}(\Phi\Sigma_{k|k-1}H_k^T + S_k)^T, \end{aligned} \tag{5.63}$$

where $\bar{Q} = \begin{bmatrix} Q & 0 \\ 0 & GRG^T \end{bmatrix}$, $\Phi = \begin{bmatrix} A & 0 \\ GC & F \end{bmatrix}$, and $S_k = \begin{bmatrix} 0 \\ GR \end{bmatrix}$.

If there is no packet loss, then our next result shows that the MMSE estimator (5.63) using $\{z_k\}$ and the IKF using $\{y_k\}$ are equivalent.

Theorem 5.13. *Suppose packet loss is not present. Then, for any coding vectors* $\{\alpha_k : k \in \mathbb{N}\}$*, we have* $\hat{x}_{k+1|k} = [I_n \;\; 0]\hat{u}_{k+1|k}$*, where* $\hat{u}_{k+1|k}$ *is given by (5.63).*

Proof. Note from (5.63) and [20] that $\hat{u}_{k+1|k} = \mathbb{E}[u_{k+1}|z_0, z_1, \ldots, z_k]$. Similarly, we get that $\hat{x}_{k+1|k} = \mathbb{E}[x_{k+1}|y_0, y_1, \ldots, y_k]$. Since the sequences $\{y_0, y_1, \ldots, y_k\}$ and $\{z_0, z_1, \ldots, z_k\}$ are equivalent (information preserving), we have

$$\mathbb{E}[u_{k+1}|z_0, z_1, \ldots, z_k] = \mathbb{E}[u_{k+1}|y_0, y_1, \ldots, y_k].$$

Multiplying its both sides by $[I_n \;\; 0]$, the right-hand side becomes

$$[I_n \;\; 0]\mathbb{E}[u_{k+1}|y_0, y_1, \ldots, y_k] = \mathbb{E}[x_{k+1}|y_0, y_1, \ldots, y_k].$$

It follows that $[I_n \;\; 0]\hat{u}_{k+1|k} = \hat{x}_{k+1|k}$. □

Although both estimators (5.63) and the IKF present the same estimate without packet loss, we will later show that, in the presence of packet loss, the coded output is information enhancing in the sense that (5.63) allows a larger critical packet loss rate for stability.

When there is no packet loss, our next result shows that the MMSE filter (5.63) using z_k and the IKF using y_k are equivalent. Let $\zeta_k := col\{z_1, z_2, \ldots, z_k\}$ and $\theta_k := col\{y_1, y_2, \ldots, y_k\}$.

Theorem 5.14. *Without packet losses and with* z_k *given by* (5.60)*, we have* $\mathbb{E}[x_k|\zeta_k] = \mathbb{E}[x_k|\theta_k]$.

Proof. By (5.60) the relationship between ζ_k and θ_k is $\zeta_k = V\theta_k$ with

$$V = \begin{bmatrix} I & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \alpha_{2,1}I & I & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \alpha_{3,1}I & \alpha_{3,2}I & I & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \alpha_{k,1}I & \cdots & \alpha_{k,m-1}I & I \end{bmatrix}.$$

Since V is nonsingular, the result follows from [20, p. 95]. □

However, a drawback of (5.63) lies in its increased complexity over its uncoded counterpart since the dimension of the estimated state u_k is $n + q(m - 1)$, whereas that of x_k is n. To avoid this extra numerical effort, a suboptimal scheme to directly estimate x_k is given in [1].

5.4.3 Mean Square Stability

We study the stability of the MMSE estimator (5.63) when the coded output is transmitted to the estimator over a lossy channel under the following trivial assumption.

Assumption 5.6. *A is unstable and invertible, and (A, C) is observable.*

To this end, we introduce the notion of *strong observability* and then show that the coded output possesses a strong observability, which allows us to work out the stability condition.

Consider the discrete-time system

$$\begin{aligned} x_{k+1} &= Ax_k + w_k, \\ y_k &= C_k x_k + v_k, \end{aligned} \tag{5.64}$$

which is the same as (5.1) except that C_k is allowed to be time varying, but $m = n - rank(C_k) + 1$ is constant.

Definition 5.3. *Given any $\tau \geq m$, system (5.64), or the pair $(A, \{C_k : k \in \mathbb{N}\})$, is said to be strongly observable with period τ if for any $1 \leq i_1 < i_2 < \ldots < i_{m-1} < \tau$ and $k \geq \tau - 1$, the regression matrix*

$$O(k, k - i_1, \ldots, k - i_{m-1}) = col\{C_k, C_{k-i_1} A^{-i_1}, \ldots, C_{k-i_{m-1}} A^{-i_{m-1}}\} \tag{5.65}$$

has full column rank.

Remark 5.5. *If a pair (A, C) is observable, then $\mathcal{C}$ is of full column rank. This implies that the pair (A, C) is strongly observable with period $\tau = m$. However, a pair (A, C) being observable does not imply that (A, C) is strongly observable with period $\tau > m$. An example is*

$$A = \begin{bmatrix} 2 & 0 \\ 2 & -2 \end{bmatrix}, \quad C = [1 \ \ 1], \tag{5.66}$$

for which the observability index $m = 2$, and (A, C) is observable, but (A, C) is not strongly observable with period $\tau > 2$ because $col\{C, CA^{-2}\} = col\{C, 0.25C\}$ is not of full column rank.

Denote $\mathcal{C} = \text{col}\{CA^{m-1}, CA^{m-2}, \ldots, C\}$ and $C_k = (\alpha_k^T \otimes I_q)\mathcal{C}$. We show that the periodic coded output makes an observable system strongly observable.

Lemma 5.6. *Consider system (5.1) under Assumption 5.6 and the coding scheme (5.60). If the coding vectors $\{\alpha_k : k \in \mathbb{N}\}$ are periodic with period $\tau \geq m$ (i.e., $\alpha_k = \alpha_{k+\tau}$) and*

$\alpha_0, \alpha_1, \ldots, \alpha_{\tau-1}$ *are randomly drawn from an absolutely continuous probability distribution,*[1] *then, with probability one,*[2] $(A, \{C_k : k \in \mathbb{N}\})$ *is strongly observable with period* τ.

Proof. For each $k \in \mathbb{N}$, define $\alpha_{km} = 1$, so that we can write $\alpha_k^T = [\alpha_{k1}, \cdots, \alpha_{km}]$. Let $1 \leq i_1 < \cdots < i_{m-1} < \tau$ and $O \triangleq O(k, k - i_1, \cdots, k - i_{m-1})$. Let also $A = MJM^{-1}$ be the Jordan decomposition of A with $J = J_1 \oplus \cdots \oplus J_B$, J_b, $b = 1, \cdots, B$ being the Jordan blocks in J, and j_b being the eigenvalue associated with J_b. Then,

$$\begin{aligned} C_k &= \sum_{l=1}^{m} \alpha_{kl} C A^{m-l} = CM \left(\sum_{l=1}^{m} \alpha_{kl} J^{m-l} \right) M^{-1} \\ &= CM \left(\bigoplus_{b=1}^{B} U_{k-i_1,b} \right) M^{-1} \end{aligned} \tag{5.67}$$

with $U_{k,b} = \sum_{l=1}^{m} \alpha_{k-i_1,l} J_b^{m-l}$. Now, for each $b \in \{1, \cdots, B\}$, all the entries on the main diagonal of $U_{k,b}$ have the same value, which we denote by $u_{k,b}$. Hence, $U_{k,b}$ is invertible if and only if $u_{k,b} \neq 0$. We have

$$u_{k,b} = \sum_{l=1}^{m} \alpha_{kl} j_b^{m-l}. \tag{5.68}$$

Since α_{kl}, $k \in \{1, \cdots, \tau\}$, $l \in \{1, \cdots, m\}$, are randomly drawn from absolutely continuous probability distribution, it follows that the measure of event that (5.68) happens is zero. This means that, with probability one, $u_{k,b} \neq 0$ for all $k \in \mathbb{N}$ and $b \in \{1, \cdots, B\}$, and therefore $U_{k,b}$ is invertible. Hence, so is $\bigoplus_{b=1}^{B} U_{k-i_1,b}$, and in view of (5.67), $\operatorname{rank}(C_k) = \operatorname{rank}(C)$.

The rest of the argument then follows by induction. For $0 \leq j < m - 1$ with $i_0 = 0$, let $O_j = \operatorname{col}\left\{C_k, C_{k-i_1} A^{-i_1}, \cdots, C_{k-i_j} A^{-i_j}\right\}$. Since (A, C) is observable and A is invertible, it follows that $\operatorname{rank}(\mathcal{C}A^{-i_j}) = n$ for all $j = 0, \cdots, m - 1$. Hence, if $\operatorname{rank}(O_j) < n$, then there exists at least one row of $\mathcal{C}A^{-i_{j+1}}$ that is not included in the linear span rowspan $\left\{O_j\right\}$ of the rows of O_j. Again, since $\alpha_{k-i_{j+1}}$ is randomly chosen, with probability one,

$$\begin{aligned} \operatorname{rowspan}\left\{C_{k-i_{j+1}} A^{-i_{j+1}}\right\} &= \operatorname{rowspan}\left\{\left(\alpha_{k-i_{j+1}}^T \otimes I_q\right) \mathcal{C}A^{-i_{j+1}}\right\} \\ &\subsetneq \operatorname{rowspan}\left\{O_j\right\}. \end{aligned}$$

Hence it follows that $\operatorname{rank}(O_{j+1}) > \operatorname{rank}(O_j)$ with probability one. This in turn implies $n \geq \operatorname{rank}(O) \geq \operatorname{rank}(C) + m - 1 = n$ and completes the proof. □

[1] Loosely speaking, this means that the probability density function contains no impulses.

[2] The measure of nonstrongly observable cases set is zero resulting from this distribution function.

We are ready to present one main result of this subsection.

Theorem 5.15. *Consider system (5.1) under Assumption 5.6 and the coding scheme (5.60). For any $\tau \geq m$, suppose that the coding vectors $\{\alpha_k : k \in \mathbb{N}\}$ are taken to be periodic with period τ and that $\alpha_0, \alpha_1, \ldots, \alpha_{\tau-1}$ are randomly drawn from an absolutely continuous probability distribution. Then, the MMSE estimator (5.63) is stable with probability one if*

$$|\rho(A)|^2(1-p)(P(\tau,m))^{1/\tau} < 1, \tag{5.69}$$

where $P(\tau,m) = \sum_{i=0}^{m-1}\binom{\tau}{i}(\frac{p}{1-p})^i \geq 1$, and $\binom{\tau}{i}$ denotes the number of combinations for choosing i from τ.

To prove the Theorem 5.15, we introduce the following lemma.

Lemma 5.7. *Suppose that $\{\alpha_k^T : k \in \mathbb{N}\}$ is periodic and $(A, \{\alpha_k^T \mathcal{C} : k \in \mathbb{N}\})$ is strongly observable with period τ. Under Assumption 5.6, if there are m packets received in time period $[(j-1)\tau, j\tau)$, $j \geq 1$, then there exists a positive value $\beta > 0$ (independent of P_0) such that $P_{j\tau|j\tau} < \beta I$.*

Proof. Suppose that $z_{t_k}, z_{t_{k-1}}, \ldots, z_{t_{k-m+1}}$ are received in time period $[(j-1)\tau, j\tau)$, $j \geq 1$, that is, $j\tau > t_k > t_{k-1} > \ldots > t_{k-m+1} \geq (j-1)\tau$. For ease of writing, here we write $O(t_k, t_{k-1}, \ldots, t_{k-m+1})$ as $O(k)$. Since $(A, \{\alpha_k^T \mathcal{C} : k \in \mathbb{N}\})$ is strongly observable with period τ, $O(k)$ is of full column rank. We can obtain a direct estimator of x_{t_k} by using $z_{t_k}, z_{t_{k-1}}, \ldots, z_{t_{k-m+1}}$:

$$\breve{x}_{t_k|t_k} = O^{\dagger}(k)\mathrm{col}\{z_{t_k}, z_{t_{k-1}}, \ldots, z_{t_{k-m+1}}\}, \tag{5.70}$$

where the superscript † denotes the Moore–Penrose pseudo-inverse [8]. Letting $z_k = C_k x_k + n_k$, since $x_{t-i} = A^{-i}x_t - \sum_{j=1}^{i} A^{-j} w_{t+j-i-1}$, $C_k = (\alpha_k^T \otimes I_q)\mathcal{C}$ and $n_k = \sum_{i=1}^{m} \alpha_{ki}(v_{k-i+1} - \sum_{j=i}^{m-1} CA^{-m+j} w_{k-j+i-1})$, we rewrite the estimator in (5.70) as

$$\begin{aligned}
\breve{x}_{t_k|t_k} \quad &= \quad O^{\dagger}(k)\mathrm{col}\Bigg\{C_{t_k}x_{t_k} + n_{t_k}, C_{t_{k-1}}A^{-t_k+t_{k-1}}\, x_{t_k} - C_{t_{k-1}} \sum_{j=1}^{t_k - t_{k-1}} A^{-t_k+t_{k-1}+j-1} w_{t_k-j} \\
&\qquad + n_{t_{k-1}}, \ldots, C_{t_{k-m+1}}A^{-t_k+t_{k-m+1}}x_{t_k} \\
&\qquad + C_{t_{k-m+1}} \sum_{j=1}^{t_k - t_{k-m+1}} A^{-t_k+t_{k-m+1}+j-1} w_{t_k-j} + n_{t_{k-m+1}}\Bigg\} \\
&= \quad O^{\dagger}(k)O(k)x_{t_k} + O^{\dagger}(k)\tilde{n}_{t_k},
\end{aligned} \tag{5.71}$$

where $\tilde{n}_{t_k}$ is a linear combination of the noises from time t_k to t_{k-m+1}.

Denote the estimation error covariance of $\breve{x}_{t_k|t_k}$ by $\breve{P}_{t_k|t_k}$. It follows that

$$\breve{P}_{t_k|t_k} \quad = \quad O^{\dagger}(k)\mathbb{E}[\tilde{n}_{t_k}\tilde{n}_{t_k}^T](O^{\dagger}(k))^T. \tag{5.72}$$

Since $t_k - t_{k-m+1}$ is finite, there exists a positive value $c > 0$ such that $\mathbb{E}[\tilde{n}_{t_k}\tilde{n}_{t_k}^T] < cI$, which results in

$$\breve{P}_{t_k|t_k} \quad < \quad c(O^T(k)O(k))^{\dagger}. \tag{5.73}$$

Since $O(k)$ is of full column rank, we have $O^T(k)O(k) > 0$. Since $\{\alpha_k^T : k \in \mathbb{N}\}$ is periodic and $t_k - t_{k-m+1} < \tau$, there exist a positive value $\kappa > 0$ such that

$$O^T(k)O(k) \quad > \quad \kappa I. \tag{5.74}$$

By substituting (5.74) into (5.73), $\breve{P}_{t_k|t_k} < c\kappa^{-1}I$. Since the estimation error covariance of the MMSE estimator is lower than that of $\breve{P}_{t_k|t_k}$, we have $P_{t_k|t_k} < c\kappa^{-1}I$. Based on the upper bounded divergence speed of the estimation error covariance (i.e., $|\lambda_{\max}|^2$), there exists $\varepsilon > 1$ such that $P_{j\tau|j\tau} \leq \varepsilon|\lambda_{\max}|^{2(j\tau-t_k)}P_{t_k|t_k}$. Since $j\tau - t_k < \tau$, the proof is completed by letting $\beta = \varepsilon|\lambda_{\max}|^{2\tau}c\kappa^{-1}$. □

Proof of Theorem 5.15. Firstly, we prove that $\sup_{k\in\mathbb{N}}\mathbb{E}[P_{k\tau|k\tau}] < \infty$. For any $j \in \{0, 1, \ldots, k\}$, denote the event that there are less than m packets received in each of $[(k-1)\tau, k\tau)$, $[(k-2)\tau, (k-1)\tau), \ldots, [j\tau, (j+1)\tau)$ but no less than m packets received in $[(j-1)\tau, j\tau)$ by $\Omega_{j,k}^m$. Let its probability be $p_{j,k}^m$. In particular, $\Omega_{0,k}^m$ means that there are less than m packets received in each of $[(k-1)\tau, k\tau), [(k-2)\tau, (k-1)\tau), \ldots, [0, \tau)$. Based on Lemma 5.6, with probability one, $(A, \alpha_k^T C : k \in \mathbb{N})$ is strongly observable with period τ, which satisfies the conditions in Lemma 5.7, and leads to $\mathbb{E}[P_{j\tau|j\tau}|\Omega_{j,k}^m] < \beta I$.

Then there exists $\varepsilon > 1$ such that

$$\begin{aligned}
\mathbb{E}[P_{k\tau|k\tau}] \quad &= \quad \sum_{j=0}^{k}\mathbb{E}[P_{k\tau|k\tau}|\Omega_{j,k}^m]p_{j,k}^m \\
&< \quad \varepsilon\sum_{j=0}^{k}|\lambda_{\max}|^{2(k-j-1)\tau}\mathbb{E}[P_{j\tau|j\tau}|\Omega_{j,k}^m]p_{j,k}^m \\
&< \quad \varepsilon\beta|\lambda_{\max}|^{-2\tau}\sum_{j=0}^{k}|\lambda_{\max}|^{2(k-j)\tau}p_{j,k}^m I.
\end{aligned} \tag{5.75}$$

Note that the probability of $\Omega_{j,k}^m$ is

$$p_{j,k}^m \quad = \quad (\sum_{i=0}^{m-1}\binom{\tau}{i}p^i(1-p)^{\tau-i})^{k-j}(1-\sum_{i=0}^{m-1}\binom{\tau}{i}p^i(1-p)^{\tau-i})$$

$$= \quad [(1-p)^{\tau} P(\tau,m)]^{k-j}(1-(1-p)^{\tau} P(\tau,m)). \tag{5.76}$$

Substituting this into (5.75) yields that

$$\mathbb{E}[P_{k\tau|k\tau}] < \varepsilon\beta|\lambda_{\max}|^{-2\tau}(1-(1-p)^{\tau}P(\tau,m))\sum_{j=0}^{k}(|\lambda_{\max}|^2(1-p)P(\tau,m)^{1/\tau})^{(k-j)\tau}. \tag{5.77}$$

From (5.77) it is clear that $|\lambda_{\max}|^2(1-p)P(\tau,m)^{1/\tau} < 1$ is a sufficient condition for $\sup_{k\in\mathbb{N}}\mathbb{E}[P_{k\tau|k\tau}] < \infty$. Since τ is finite, the proof is completed. □

Remark 5.6. *Notice that $(P(\tau,m))^{1/\tau} \to 1$ as $\tau \to \infty$. This implies that the sufficient condition (5.69) becomes necessary as $\tau \to \infty$.*

We use example (5.66) to illustrate the advantage of coding on stability.

Example 5.1. *In view of [9], a necessary and sufficient condition for the stability of IKF (i.e., without coding) of the second-order system* (5.66) *is given by*

$$|\rho(A)|^4(1-p) < 1. \tag{5.78}$$

Clearly, it follows from Theorem 5.15 that we can always choose a period τ such that the stability condition resulting from coding is strictly weaker than (5.78).

5.5 Robust State Estimation With Random Data Droppings

5.5.1 System With Parametric Errors

In this section, we consider the following linear stochastic systems with parametric errors:

$$\begin{cases} x_{k+1} &= A(\epsilon_k)x_k + B(\epsilon_k)w_k, \\ y_k &= C(\epsilon_k)x_k + v_k, \end{cases} \tag{5.79}$$

where ϵ_k characterizes the parametric errors of the plant state-space model. Here $A(\epsilon_k)$ and $C(\epsilon_k)$ stand respectively for the plant state transition and output matrices. Usually, they are given as $A(p_0,\epsilon_k)$ and $C(p_0,\epsilon_k)$ with p_0 representing the nominal value of a plant parameter vector. As this value is usually known in robust state estimation, its existence in system matrices is not explicitly revealed in the adopted model. This is only for avoiding long and complicated mathematical expressions.

Hence, when $\epsilon_k = 0$, these matrices are in fact the plant nominal system matrices. For ease of notation, let $A := A(0)$, $B := B(0)$, and $C := C(0)$. According to the adopted hypotheses, all these matrices are known. On the other hand, the vector ϵ_k is permitted to be time varying and generally unknown, which enables system (5.79) to describe a large class of plants and include many widely adopted models, such as those based on linear fractional transformations and so on. In model-based robust system design or state estimation, some upper magnitude bounds or stochastic properties are usually assumed available for this parametric error vector. This kind of information is further important in determining the design parameter.

5.5.2 Robust State Estimator

The main objectives of this section are to derive an estimate for the plant state vector x_k using the received plant output $\{y_i\}_{i=0}^{k}$ and information about the corresponding $\{\gamma_i\}_{i=0}^{k}$. When the linear system has a precise state space model, a widely adopted state estimation procedure is the Kalman filter. This estimation procedure, however, may sometimes not work very satisfactorily due to modeling errors. To overcome this disadvantage, various modifications have been suggested, such as in [21,22,28,31] and references therein. Among these modifications, one method is based on sensitivity penalization, in which a cost function is constructed on the basis of least squares/likelihood maximization interpretations for the Kalman filter and a penalization on the sensitivity of its innovation process to modeling errors. We further discuss it in detail. To this end, let $\hat{x}_{k|l}$ denote the optimal estimate of x_k based on observations $\{y_i\}_{i=0}^{l}$ and $P_{k|l}$ represent the pseudo-covariance matrix (PCM) of the corresponding state estimation errors. In view of [22,31], the robust estimate is given by

$$\hat{x}_{k+1|k+1} = A\hat{x}_{k|k+1} + B\hat{w}_{k|k+1}, \tag{5.80}$$

$$\begin{aligned}\begin{bmatrix}\hat{x}_{k|k+1}\\ \hat{w}_{k|k+1}\end{bmatrix} = {} & \arg\min_{x_k,w_k} \frac{1}{2}\Bigg\{\mu_k\Big[\|x_k-\hat{x}_{k|k}\|^2_{P^{-1}_{k|k}} + \|w_k\|^2_{Q^{-1}}\Big] \\ & +\gamma_{k+1}\Bigg[\mu_k\|e_k(0,0)\|^2_{R^{-1}} + (1-\mu_k)\Bigg(\Bigg\|\frac{\partial e_k(\epsilon_k,\epsilon_{k+1})}{\partial \epsilon_k}\Bigg\| \\ & + \Bigg\|\frac{\partial e_k(\epsilon_k,\epsilon_{k+1})}{\partial \epsilon_{k+1}}\Bigg\|\Bigg)_{\epsilon_k=\epsilon_{k+1}=0}\Bigg]\Bigg\},\end{aligned} \tag{5.81}$$

where the weighted norm $\|x\|_Q$ is defined as $\|x\|_Q = \sqrt{x'Qx}$ for any positive definition matrix Q, and the prediction error of the measurement is defined as

$$e_k(\epsilon_k,\epsilon_{k+1}) = y_{k+1} - C(\epsilon_{k+1})[A(\epsilon_k)x_k + B(\epsilon_k)w_k].$$

The scalar term μ_k is a design parameter reflecting a trade-off between nominal estimation accuracy and penalization on the first-order approximation of deviations of the innovation process, and it renders its optimizer become less sensitive to the system modeling errors. However, it is still not clear whether or not these minimizers share the global optimality property of the Kalman filter.

When there do not exist parametric uncertainties in the plant model, the matrix $P_{k|k}$ is in fact the covariance matrix of the estimation errors of the Kalman filter. This makes it possible to explain $\hat{x}_{k|k+1}$ and $\hat{w}_{k|k+1}$ respectively as the $y_i|_{i=0}^{k+1}$-based MLEs of x_k and w_k. However, when there exist modeling errors in system matrices, physical interpretations of the matrix $P_{k|k}$ need further clarifications. To avoid possible misunderstandings, it is called PCM in this chapter.

Here the same approach as in [22] is adopted to deal with state estimation for system (5.1) in which both parametric uncertainties and random measurement droppings exist. It is worth pointing out that although this extension has been attempted, the success is rather limited. One of the major restrictions is its implicit ergodic requirement on the received plant output measurements. Another is that information about the realization of the random process $\{\gamma_k\}$ has not been efficiently utilized. These disadvantages have been successfully overcome in this chapter through introducing another cost function.

By [29] the robust estimate and the corresponding PCM under packet droppings are given further.

Theorem 5.16. *Let $\lambda_k = (1-\mu_k)/\mu_k$ with μ_k given in (5.81). Assume that both $P_{k|k}$ and Q are nonsingular. Then, the robust estimate of the state system (5.79) based on $y_i|_{i=0}^{k+1}$ is recursively updated as*

$$\hat{x}_{k+1|k+1} = \begin{cases} A\hat{x}_{k|k} & \text{if } \gamma_{k+1} = 0, \\ \hat{A} + P_{k+1|k+1}C^T R^{-1}(y_{k+1} - C\hat{A}\hat{x}_{k|k}) & \text{if } \gamma_{k+1} = 1. \end{cases} \tag{5.82}$$

Moreover, the associated PCM $P_{k|k}$ is updated as

$$P_{k+1|k+1} = \begin{cases} AP_{k|k}A^T + BQB^T & \text{if } \gamma_{k+1} = 0, \\ \left\{\left[A\hat{P}_{k|k}A^T + \hat{B}\hat{Q}_k\hat{B}^T\right]^{-1} + C^T R^{-1} C\right\}^{-1} & \text{if } \gamma_{k+1} = 1, \end{cases} \tag{5.83}$$

where

$$\begin{aligned} \hat{P}_{k|k} &= \left(P_{k|k}^{-1} + \lambda_k S_k^T S_k\right)^{-1}, \\ \hat{Q}_k &= [Q^{-1} + \lambda_k T_k^T (I + \lambda_k S_k P_{k|k} S_k^T)^{-1} T_k]^{-1}, \\ \hat{B}_k &= B - \lambda_k A \hat{P}_{k|k} S_k^T T_k, \\ \hat{A} &= [A - \lambda_k \hat{B}_k \hat{Q}_k T_k^T S_k](I - \lambda_k \hat{P}_{k|k} S_k^T S_k), \\ S_k &= C(\epsilon_{k+1}) \frac{\partial A(\epsilon_k)}{\partial \epsilon_k} + \frac{\partial C(\epsilon_{k+1})}{\partial \epsilon_{k+1}} A(\epsilon_k) \Big|_{\epsilon_k = \epsilon_{k+1} = 0}, \\ T_k &= C(\epsilon_{k+1}) \frac{\partial B(\epsilon_k)}{\partial \epsilon_k} + \frac{\partial C(\epsilon_{k+1})}{\partial \epsilon_{k+1}} B(\epsilon_k) \Big|_{\epsilon_k = \epsilon_{k+1} = 0}. \end{aligned}$$

Note that when $\gamma_{k+1} = 0$, the above estimator is just a one-step state predictor using nominal system matrices. On the other hand, when $\gamma_{k+1} = 1$, the above estimator still has the same structure as that of the Kalman filter, except that the nominal system matrices should be adjusted. The adjustment method of these matrices is completely the same as that of the robust state estimator (RSE) developed in [31] and is no longer required if the design parameter μ_k is selected to be 1. This means that the above recursive estimation procedure is consistent with both RSE of [31] and IKF.

5.5.3 Convergence of the Robust State Estimator

In this subsection, the Riemannian distance for positive definite matrices defined in Chapter 2, which has already been utilized in analyzing asymptotic properties of IKF and Kalman filter with random coefficients [23,24], is adopted in the analysis of the robust state estimator given in the previous subsection.

Recall that, for two arbitrary positive definite matrices $P \in \mathcal{R}^{n\times n}$ and $Q \in \mathcal{R}^{n\times n}$, the Riemannian distance of Chapter 2 between them is defined as

$$\delta(P, Q) = \left(\sum_i \log^2(\lambda_i)\right)^{1/2},$$

where $\lambda_i(PQ^{-1})$ stands for the ith eigenvalue of the matrix PQ^{-1}. Obviously, this distance is invariant under a conjugacy transformation and the operation of matrix inversions, which is quite attractive in many situations of great engineering significance. It has also been proven that when equipped with this distance, the space of positive definite matrices is complete. This metric has been recognized to be very useful for many years in studying asymptotic properties of Kalman filtering with random system matrices [23]. Its effectiveness in studying asymptotic properties of recursive state estimations with random data dropping has also been discovered in [24] and [11].

The power of this Riemannian distance in theoretical studies for recursive computation has been enlarged by the so-called homographic transformation, whose definition is also given in Chapter 2. More precisely, for matrices P and Φ with appropriate dimensions, the homographic transformation $H_m(\Phi, P)$ is defined as

$$H_m(\Phi, P) = [\Phi_{11}P + \Phi_{12}][\Phi_{21}P + \Phi_{22}]^{-1}.$$

Here, the matrix Φ is divided as $\Phi = \left[\Phi_{ij}|_{i,j=1}^{2}\right]$ with its submatrices having compatible dimensions, whereas the matrix $\Phi_{21}P + \Phi_{22}$ is assumed to be square and of full rank. One of the most attractive properties of this transformation is its simplicity in representing cascade connections, which ensures that, for any two compatible matrices Φ_1 and Φ_2, it follows that

$$\mathbf{H}_m(\Phi_2, \mathbf{H}_m(\Phi_1, P)) = \mathbf{H}_m(\Phi_2\Phi_1, P).$$

Obviously, this relation is quite appreciative in analyzing properties of composite functions, which is often encountered in recursive estimations. As a matter of fact, this property has played important roles in analyzing the asymptotic properties of the covariance matrix of the Kalman filter and of the pseudo-covariance matrix $P_{k|k}$ and estimation errors of the aforementioned robust state estimator [11,23]. On the other hand, this property can be obtained through straightforward algebraic manipulations [11,25].

Particularly, define

$$\Phi_{k+1} = \begin{cases} \begin{bmatrix} A & BQB'A^{-T} \\ 0 & A^{-T} \end{bmatrix} & \text{if } \gamma_{k+1} = 0, \\ \begin{bmatrix} \tilde{A}_k & \tilde{Q}_k B' \tilde{A}_k^{-T} \\ \tilde{C}'_{k+1} \tilde{R}_k^{-1} \tilde{C}_{k+1} \tilde{A}_k & [I + \tilde{C}'_{k+1} \tilde{R}_k^{-1} \tilde{C}_{k+1} B \tilde{Q}_k B'] \tilde{A}_k^{-T} \end{bmatrix} & \text{if } \gamma_{k+1} = 1, \end{cases}$$

where

$$\begin{aligned} \check{Q}_k &= (Q + \lambda_k T_k' T_k)^{-1}, \\ \check{A}_k &= A - \lambda_k B \check{Q} T_k' S_k, \ \tilde{B}_k = \check{A}_k^{-1} B, \\ \tilde{A}_k &= \check{A}_k + B \check{Q}_k \tilde{B}_k' \tilde{S}_k' \tilde{S}_k, \\ \tilde{Q}_k &= \check{Q}_k + \check{Q}_k \tilde{B}_k' \tilde{S}_k' \tilde{S}_k \tilde{B}_k \check{Q}_k, \ \tilde{S}_k = \sqrt{\lambda_k} [I + \lambda_k T_k Q T_k']^{-1/2} S_k, \\ \tilde{C}_{k+1} &= \begin{bmatrix} \tilde{S}_k \check{A}_k^{-1} \\ C \end{bmatrix}, \ \tilde{R}_{k+1} = \begin{bmatrix} I + \tilde{S}_k \tilde{B}_k \check{Q}_k \tilde{B}_k' \tilde{S}_k' & 0 \\ 0 & R \end{bmatrix}. \end{aligned}$$

It has been proven in [29] that the pseudo-covariance matrix $P_{k+1|k+1}$ of the robust state estimator given by Eq. (5.83) can be rewritten in a more concise form as

$$P_{k+1|k+1} = \mathbf{H}_m \left(\Phi_{k+1}, P_{k|k} \right).$$

On the basis of this expression and the cascade property of the homographic transformation, the next relation immediately follows [26,29]:

$$P_{k|k} = \mathbf{H}_m \left(\prod_{t=k}^{1} \Phi_t, P_{0|0} \right). \tag{5.84}$$

This equality provides a quite compact expression of the relation between the pseudo-covariance matrix of the robust state estimator at the time instant k, that is, $P_{k|k}$ and its initial value $P_{0|0}$, which significantly reduces mathematical difficulties in analyzing its convergence properties, compared with the expression of Eq. (5.83) [29].

Combining the preceding two results, the following results on the convergence of the $P_{k|k}$ can be obtained.

Theorem 5.17. *([29]) Let $A^{[1]} = \tilde{A}_k$, $G^{[1]} = B\tilde{Q}_k^{1/2}$, $H^{[1]} = \tilde{R}_{k+1}^{-1/2}\tilde{C}_{k+1}$, and $A^{[2]} = A$, $G^{[2]} = BQ^{1/2}$. Assume that A, $\check{A}_k$, and $\tilde{A}_k$ are invertible and that there exist two positive integers m_1 and m_2 such that the matrix pair $(A^{[1]}(A^{[2]})^{m_1}, H^{[1]})$ is observable and one of the following three conditions are satisfied:*

- *the matrix pair $(A^{[1]}(A^{[2]})^{m_2}, G^{[1]})$ is controllable;*
- *the matrix pair $((A^{[2]})^{m_2}A^{[1]}, G^{[2]})$ is controllable;*
- *the matrix pair $(A^{[2]}, G^{[2]})$ is controllable.*

Then, the sequence of matrices $P_{k|k}$ converges to a stationary distribution with probability one, which is independent of its initial value $P_{0|0}$.

The proof of the theorem heavily relies on exploring the decrease of $P_{k|k}$ measured by the Riemannian distance and can be found in [29].

5.6 Asymptotic Properties of State Estimations With Random Data Dropping

In this section, we investigate the convergence rate and stationary distribution approximation in a unified framework for both the covariance matrix of the Kalman filter and the pseudo-covariance matrix of the robust recursive state estimator under the situation of random data loss. It is proven that when the measurement dropping process is described by a Markov chain and the associated plant is both controllable and observable if the dropping probability is less than 1, these two matrices converge exponentially to a stationary distribution independent of their initial values. In addition, if these covariance matrix and pseudo-covariance matrix are initialized with the stabilizing solution of the associated algebraic Riccati equation, then both are shown to converge to an ergodic process. Furthermore, two approximations are derived for their stationary probability distributions and to a bound of approximation errors. Compared with the delta function suggested in [10], a series of delta functions is shown to give a more accurate approximation, and replacement of the update gain matrix by a constant one usually deteriorates the steady estimation accuracy of these state estimators.

5.6.1 Unified Problem Description and Preliminaries

Investigations in the previous sections reveal that even when there are random data droppings in a networked system, the covariance matrix of the Kalman filter and the pseudo-covariance matrix of the sensitivity penalization-based robust state estimator still have a similar recursion formula. This suggests a unified study of their properties. Attention of this section is focused on their convergence characteristics that are important from both theoretical and application points of view [27–29]. To achieve this objective, let α and β be two constants belonging to $(0,\ 1)$, let $A^{[0]}$ and $A^{[1]}$ be two prescribed real square matrices with their dimensions equal to each other, and let $G^{[0]}$, $G^{[1]}$, and $H^{[1]}$ be other three known real matrices with compatible dimensions. Consider the following random matrix recursion:

$$P_{k+1|k+1} = \begin{cases} A^{[0]}P_{k|k}A^{[0]T} + G^{[0]}G^{[0]T}, & \gamma_{k+1} = 0, \\ \left\{\left(A^{[1]}P_{k|k}A^{[1]T} + G^{[1]}G^{[1]T}\right)^{-1} + H^{[1]T}H^{[1]}\right\}^{-1}, & \gamma_{k+1} = 1. \end{cases} \tag{5.85}$$

Assume that this recursion starts from an initial positive definite matrix $P_{0|0}$. Moreover, assume that $\gamma_k|_{k=0}^{\infty}$ is a Markov chain described by

$$\begin{bmatrix} \mathbf{P}_r(\gamma_k = 1) \\ \mathbf{P}_r(\gamma_k = 0) \end{bmatrix} = \begin{bmatrix} \alpha & 1-\beta \\ 1-\alpha & \beta \end{bmatrix} \begin{bmatrix} \mathbf{P}_r(\gamma_{k-1} = 1) \\ \mathbf{P}_r(\gamma_{k-1} = 0) \end{bmatrix}. \tag{5.86}$$

Here $\gamma_k = 1$ means that the plant output measurement at the kth time instant has been successfully transmitted, whereas $\gamma_k = 0$ represents a data transmission failure of the communication channel.

If $A^{[0]} = A^{[1]}$ and $G^{[0]} = G^{[1]}$, then when the matrices $A^{[1]}$, $G^{[1]}$, and $H^{[1]}$ are respectively the plant state transition matrix, input matrix, and output matrix, recursion (5.85) becomes that of the covariance matrix of estimation errors of the Kalman filter for a linear time-invariant plant with intermittent observations; that is, the matrix $P_{k|k}$ is in fact its CMEE. This recursion is originally derived in [3], and its properties are discussed extensively, for example, in [10,13,24,30].

When $A^{[0]} \neq A^{[1]}$ and/or $G^{[0]} \neq G^{[1]}$, the matrices $A^{[0]}$ and $G^{[0]}$ can be used to represent nominal values of the plant state transition matrix and input matrix, whereas $A^{[1]}$, $G^{[1]}$, and $H^{[1]}$ can be used to represent modifications of these matrices and the plant nominal output matrix. In these modifications, derivatives of plant parameters to modeling errors can be explicitly taken into account, as done in [29,31]. When these values are adopted, the aforementioned recursion becomes that of the pseudo-covariance matrix of the robust state estimator developed in [29] under the restrictions that both the plant nominal model and the derivatives of the plant parameters to modeling errors are time invariant. In other words, the matrix $P_{k|k}$ is now in fact the pseudo-covariance matrix of this robust state estimator.

In either of these two situations, as the matrices $A^{[*]}$ and $G^{[*]}$ with $* = 0, 1$ and $H^{[1]}$ are completely determined by plant nominal model parameters, all of them are known. Note that in these two state estimators, the matrix $P_{k|k}$ directly connects both their estimation accuracy and their update gain matrix. Its asymptotic characteristics are very important in understanding their properties [10,27–29].

In this section, the convergence rate of the random matrix recursion (5.85) is studied, as well as approximations to its stationary distribution. In this investigation, the Riemannian distance defined in Chapter 2 for positive definite matrices is once again utilized together with the homographic transformation defined in that chapter. In addition, some important properties of Hamiltonian matrices are also utilized, which are listed in Chapter 2.

In this analysis of the asymptotic properties of the aforementioned random matrix recursion, the following results on Markov processes are also utilized.

Lemma 5.8. *[32,33] Let $x_i|_{i=0}^{\infty}$ be a positive recurrent irreducible Markov chain defined on a probability space $(\Omega, \mathcal{F}, P)$ with a countable state space $\mathcal{I}$, and let $f(\cdot)$ be a real-valued function on $\mathcal{I}$. Denote the time of the αth entrance of the Markov chain into its jth state by $\tau_\alpha^{[j]}$, and $\sum_{k=\tau_\alpha^{[j]}}^{\tau_{\alpha+1}^{[j]}-1} f(x_k)$ by $f_\alpha^{[j]}$. If both $\mathbf{E}\left(|f_\alpha^{[j]}|^3\right)$ and $\mathbf{E}\left(|\tau_{\alpha+1}^{[j]} - \tau_\alpha^{[j]}|^3\right)$ are finite and $\sigma_j = \sqrt{\mathbf{V}_{ar}\{f_\alpha^{[j]} - s(f)(\tau_{\alpha+1}^{[j]} - \tau_\alpha^{[j]})\}}$ is greater than 0, then*

$$\sup_{t\in\mathcal{R}} \left| \mathbf{P}_r \left\{ \frac{1}{\sigma_j\sqrt{n\pi_j}} \left(\sum_{k=0}^{n} f(x_k) - (n+1)s(f) \right) < t \right\} - \phi(t) \right| = O\left(\left(\frac{ln(n)}{n} \right)^{1/4} \right), \quad (5.87)$$

where $s(f) = \sum_{i\in\mathcal{I}} \frac{f(i)}{\mu_i}$ with μ_i the mathematical expectation of the recurrence time of the ith state, and $\pi_i = \mu_i^{-1}$.

Lemma 5.9. *[34] Assume that a Markov process $x_i|_{i=0}^{\infty}$ has an unique stationary distribution μ. Then this process is ergodic when x_0 is any element of the support of the distribution μ.*

5.6.2 Asymptotic Properties of the Random Matrix Recursion

To analyze asymptotic properties of the aforementioned random process $P_{k|k}$, in this section, we adopt the following three conditions.

- Condition I: Both matrices $A^{[0]}$ and $A^{[1]}$ are invertible.
- Condition II: $(A^{[1]},\ G^{[1]})$ is controllable, and $(A^{[1]},\ H^{[1]})$ is observable.
- Condition III: $0 < \alpha,\ \beta < 1$.

Note that the measure of singular matrices is equal to zero [35]. It can therefore be declared that Condition I is usually satisfied. In addition, even if it is not satisfied, a small perturbation on the elements of the matrices $A^{[0]}$ and/or $A^{[1]}$ can make it satisfied. As the results of this section do not explicitly depend on either the inverse of $A^{[0]}$ or that of $A^{[1]}$, validity of the corresponding conclusions can be proven through taking their limit with reducing the perturbation magnitude to zero. On the other hand, recall that observability is necessary for plant state reconstruction from its input/output data, and controllability is necessary for locating plant poles into a desirable area that is important in stabilizing a state estimator [27]. It appears safe to claim that Condition II is not very restrictive for guaranteeing the convergence of the random matrix recursion of Eq. (5.85). Moreover, when $\alpha = 1$ and $\beta = 0$, this recursion reduces to that of state estimation without data loss, whereas $\alpha = 0$ and $\beta = 1$ mean that plant output measurement is never sent to the state estimator. The former has been well studied in traditional Kalman filtering and robust state estimations, whereas the latter has very limited engineering significance. If $\alpha = \beta = 1$, depending on the initial value of the Markov chain, the recursion of Eq. (5.85) becomes one of the aforementioned two scenarios.

When Condition I is satisfied, define the matrices $M^{[0]}$ and $M^{[1]}$ as

$$M^{[0]} = \begin{bmatrix} A^{[0]} & G^{[0]}G^{[0]T}A^{[0]-T} \\ 0 & \left(A^{[0]}\right)^{-T} \end{bmatrix},$$

$$M^{[1]} = \begin{bmatrix} A^{[1]} & G^{[1]}G^{[1]T}\left(A^{[1]}\right)^{-T} \\ H^{[1]T}H^{[1]}A^{[1]} & [I + H^{[1]T}H^{[1]}G^{[1]}G^{[1]T}]\left(A^{[1]}\right)^{-T} \end{bmatrix}.$$

Then, direct algebraic manipulations show that $P_{k+1|k+1}$ of Eq. (5.85) can be reexpressed as

$$P_{k+1|k+1} = \begin{cases} \mathbf{H}_m(M^{[0]},\ P_{k|k}), & \gamma_{k+1} = 0, \\ \mathbf{H}_m(M^{[1]},\ P_{k|k}), & \gamma_{k+1} = 1. \end{cases} \tag{5.88}$$

Recall that a matrix Φ is Hamiltonian if $\Phi^T J \Phi = J$, where $J = [0\ \ I; -I\ \ 0]$. By the definitions of the matrices $M^{[0]}$ and $M^{[1]}$ direct algebraic manipulations show that both they are Hamiltonian. It can therefore be declared from Lemma 2.5 that for any PDM X with a compatible dimension, both $\mathbf{H}_m(M^{[0]}, X)$ and $\mathbf{H}_m(M^{[1]}, X)$ are well defined and positive definite. Moreover, a repetitive utilization of Lemma 2.6 directly shows that if $P_{0|0}$ is positive definite, then for any positive integer j smaller than k,

$$\begin{aligned} P_{k|k} &= \mathbf{H}_m\left(M^{[\gamma_k]},\ \mathbf{H}_m\left(M^{[\gamma_{k-1}]}, \cdots, \mathbf{H}_m\left(M^{[\gamma_{j+1}]},\ P_{j|j}\right)\cdots\right)\right) \\ &= \mathbf{H}_m\left(\prod_{i=k}^{j+1} M^{[\gamma_i]},\ P_{j|j}\right). \end{aligned} \tag{5.89}$$

When $j=0$, this equation reveals a relation between $P_{k|k}$ and its initial value $P_{0|0}$. Furthermore, define α_{1h} and α_{0h} as

$$\alpha_{0h}=\sup_{X,Y>0,\ X\neq Y}\frac{\delta\left(\mathbf{H}_m(M^{[0]},X),\ \mathbf{H}_m(M^{[0]},Y)\right)}{\delta(X,Y)},$$
$$\alpha_{1h}=\sup_{X,Y>0,\ X\neq Y}\frac{\delta\left(\mathbf{H}_m(M^{[1]},X),\ \mathbf{H}_m(M^{[1]},Y)\right)}{\delta(X,Y)}.$$

It has been proved in [23,29] that $\alpha_{0h}\leq 1$ and $\alpha_{1h}<1$ whenever Conditions I and II are satisfied.[3]

Based on these properties and Lemma 5.8, we obtain the following theorem. Its proof is given in the appendix.

Theorem 5.18. *For a prescribed binary sequence $\gamma_i|_{i=1}^{k}$ with $\gamma_i\in\{0,\ 1\}$, define the functions $\Phi_k(\cdot)$ and $\rho_k(\cdot,\ \cdot)$ on the set of PDMs as $\Phi_k(X)=\mathbf{H}_m\left(M^{[\gamma_k]},\ \mathbf{H}_m\left(M^{[\gamma_{k-1}]},\ \cdots,\ \mathbf{H}_m\left(M^{[\gamma_1]},\ X\right)\cdots\right)\right)$ and $\rho_k(X,Y)=\delta(\Phi_k(X),\Phi_k(Y))$. Assume that Conditions I–III are satisfied. Then, for the random process $\gamma_i|_{i=1}^{\infty}$ and arbitrary PDMs X and Y,*

$$\lim_{k\to\infty}\rho_k(X,Y)=0\quad \textit{in probability.}\tag{5.90}$$

Moreover, the convergence rate is $O\left(e^{-\sqrt{k}}\right)$.

For a prescribed PDM X and a particular realization $\gamma_k|_{k=1}^{\infty}$ of the random measurement loss process, if $\lim_{k\to\infty}\Phi_k(X)$ exists, denote it by $P(\gamma_k|_{k=1}^{\infty})$. Using this notation, define the set

$$\mathcal{P}=\left\{P(\gamma_k|_{k=1}^{\infty})\ \middle|\ P(\gamma_k|_{k=1}^{\infty})=\lim_{n\to\infty}\mathbf{H}_m\left(\prod_{i=n}^{1}M^{[\gamma_i]},\ X\right),\qquad X>0,\ \gamma_i\in\{0,\ 1\}\right\}.\tag{5.91}$$

Then, Theorem 5.18 makes it clear that when Conditions I–III are satisfied, this matrix set is not empty and is independent of a particular PDM X. On the other hand, from its definition and Eq. (5.89), it is obvious that this matrix set consists of all the possible final values of the random process $P_{k|k}$. Moreover, as its element $P(\gamma_k|_{k=1}^{\infty})$ can be associated with the number $\sum_{i=1}^{\infty}\gamma_i 2^{i-1}$, the set $\mathcal{P}$ is apparently countable.

On the other hand, when Conditions I–III are satisfied, by Proposition 6 of [24] and Theorem 5 of [29] we can claim that the random process $P_{k|k}$ converges to a stationary distribution. Theorem 5.18 makes it clear that this stationary distribution is unique and the convergence

[3] More specifically, simultaneous satisfaction of these two conditions implies the existence of a finite positive integer m such that the mapping $\mathbf{H}_m(M^{[1]m},X)$ is strictly contractive.

rate is exponential. Moreover, the set $\mathcal{P}$ defined in Eq. (5.91) includes the support of this stationary distribution as a subset.

Under Condition II, a well-established conclusion in system theory is that the algebraic Riccati equation $P = \left[(A^{[1]}PA^{[1]T} + G^{[1]}G^{[1]T})^{-1} + H^{[1]T}H^{[1]}\right]^{-1}$ has a unique stabilizing solution. We denote it by $P^{\star}$ in the rest of this section. Moreover, a widely known result in Kalman filtering is that, under this condition, the Riccati recursion $P_{k+1|k+1} = \left[(A^{[1]}P_{k|k}A^{[1]T} + G^{[1]}G^{[1]T})^{-1} + H^{[1]T}H^{[1]}\right]^{-1}$ converges to $P^{\star}$ with the increment of the temporal variable k [27,28], which is in accordance with recursion (5.85) with $\alpha = 1$ and $\beta = 0$, that is, there is no data dropping.

On the basis of these results, from Lemma 5.9 we establish the ergodicity of the random process $P_{k|k}$.

Corollary 5.1. *In addition to Conditions I–III, if $P_{k|k}$ starts from $P^{\star}$, then, this random process is also ergodic.*

Proof. When Conditions I–III are satisfied, the random process $P_{k|k}$ exponentially converges to a unique stationary distribution. Assume that $\gamma_k \equiv 1$. Then, for an arbitrary PDM X,

$$P_{k|k} = \mathbf{H}_m\left(M^{[1]k},\ X\right), \qquad k = 1, 2, \cdots. \tag{5.92}$$

Under Condition II, from the convergence of the Kalman filter [27,28] we have that $\lim_{k\to\infty}\mathbf{H}_m\left(M^{[1]k}, X\right)$ exists and is equal to $P^{\star}$. Moreover, from the definition of the matrix $P^{\star}$ it is obvious that $\mathbf{H}_m\left(M^{[1]}, P^{\star}\right) = P^{\star}$. Therefore, $P^{\star}$ belongs to the support of the stationary distribution of the random process $P_{k|k}$.

We can therefore declare from Lemma 5.9 that the random process $P_{k|k}$ initialized with $P_{0|0} = P^{\star}$ is ergodic. This completes the proof. □

When both α and β belong to the open set $(0,\ 1)$, it can be directly proven that the Markov chain $\gamma_k|_{k=1}^{\infty}$ has a stationary distribution. Denote a random variable with this stationary distribution by γ_{∞}. Then, the probability that γ_{∞} takes the value of 1 or 0 does not depend on the temporal variable k, which can be respectively expressed as $\mathbf{P}_r(\gamma_{\infty} = 1) = \frac{1-\beta}{2-\alpha-\beta}$ and $\mathbf{P}_r(\gamma_{\infty} = 0) = \frac{1-\alpha}{2-\alpha-\beta}$.

From Corollary 5.1 it is clear that the stationary distribution of the random process $P_{k|k}$ can be well approximated by its time series samples. To clarify accuracy of this approximation, some properties of a Markov process given in Lemma 5.8 are utilized.

For a binary series $\gamma_i|_{i=0}^{-\infty}$ with $\gamma_i \in \{0,\ 1\}$, denote $\sum_{i=0}^{-\infty}\gamma_i 2^i$ by n and define $P^{[n]}$ as

$$P^{[n]} = \lim_{k\to\infty}\mathbf{H}_m\left(M^{[\gamma_0]}M^{[\gamma_{-1}]}\cdots M^{[\gamma_{-k}]},\ P^{\star}\right),$$

provided that the associated limit exists. Moreover, for a prescribed positive number ε, define the set $\mathcal{P}^{[n]}(\varepsilon)$ of PDMs as

$$\mathcal{P}^{[n]}(\varepsilon) = \left\{ P \;\middle|\; \delta(P^{[n]},\ P) \leq \varepsilon,\ \ P \geq 0 \right\}. \tag{5.93}$$

Then, by Theorem 5.18, for any n_1 and n_2 that can be expressed as $\sum_{i=0}^{-\infty} \gamma_i 2^i$ and the corresponding $P^{[n_1]}$ and $P^{[n_2]}$ exist, there is at least one finite length binary sequence $\gamma_i^{[n_1,n_2]}|_{i=1}^{N(n_1,n_2)}$ with $\gamma_i^{[n_1,n_2]} \in \{0,\ 1\}$ such that

$$\mathbf{H}_m\left(\prod_{i=N(n_1,n_2)}^{1} M^{[\gamma_i^{[n_1,n_2]}]},\ P^{[n_1]} \right) \in \mathcal{P}^{[n_2]}(\varepsilon) \text{ in probability.} \tag{5.94}$$

Note that when $\bar{\gamma}_i = \gamma_{i-k}$, we have that $M^{[\bar{\gamma}_i]} = M^{[\gamma_{i-k}]}$, $i = 0, 1, \cdots, k$. This means that

$$\mathbf{H}_m\left(M^{[\gamma_k]} M^{[\gamma_{k-1}]} \cdots M^{[\gamma_1]},\ P^{\star} \right) = \mathbf{H}_m\left(M^{[\bar{\gamma}_0]} M^{[\bar{\gamma}_{-1}]} \cdots M^{[\bar{\gamma}_{-k}]},\ P^{\star} \right), \tag{5.95}$$

and this relation is valid for all positive integers (including $+\infty$). In addition, it can be declared from Theorem 5.18 that, under Conditions I–III, $\lim_{k\to\infty} \mathbf{H}_m \left(M^{[\gamma_k]} M^{[\gamma_{k-1}]} \cdots M^{[\gamma_1]},\ P^{\star} \right)$ exists in probability. Therefore, the matrix set $\mathcal{P}$ defined in Eq. (5.91) can also be expressed as

$$\mathcal{P} = \left\{ P^{[n]} \;\middle|\; P^{[n]} = \lim_{k\to-\infty} \mathbf{H}_m\left(\prod_{i=0}^{k} M^{[\gamma_i]},\ \ P^{\star} \right),\ \ n = \sum_{i=0}^{-\infty} \gamma_i 2^i,\ \ \gamma_i \in \{0,\ 1\} \right\}. \tag{5.96}$$

On the other hand, Theorem 5.18 declares that when Conditions I–III are satisfied, $\lim_{k\to\infty} \rho_k(X, Y) = 0$ in probability for arbitrary PDMs X and Y. It can therefore be declared that, for arbitrary $P^{[p]}$ and $P^{[q]}$ belonging to the set $\mathcal{P}$, there exists a binary series $\gamma_j^{[pq]}|_{j=1}^{\infty}$ with $\gamma_j^{[pq]} \in \{\ 0,\ \ 1\ \}$ such that $P^{[p]} = \lim_{k\to\infty} \mathbf{H}_m \left(M^{[\gamma_k^{[pq]}]} M^{[\gamma_{k-1}^{[pq]}]} \cdots M^{[\gamma_1^{[pq]}]},\ P^{[q]} \right)$ in probability. In addition, it has been mentioned before that, for an arbitrary positive ε, only finitely many steps are required in probability to transform an element of $\mathcal{P}^{[p]}(\varepsilon)$ to an element of the set $\mathcal{P}^{[q]}(\varepsilon)$ by recursions (5.85). Note that $\mathcal{P}^{[p]}(\varepsilon)$ degenerates into $\{P^{[p]}\}$ as ε decreases to 0. This means that the Markov chain $P_{k|k}$ is approximately irreducible and positive recurrent.

Based on these observations, the following results are obtained, which give an approximation of the stationary distribution of the random process $P_{k|k}$ and its approximation accuracy. Their proof is deferred to the appendix of this chapter.

Theorem 5.19. *Let $F(x)$ denote the distribution function of $\delta(P_{\infty|\infty},\ P^{\star})$, let $P_{k|k}$ be the random matrix recursion of Eq. (5.85) with its initial value $P_{0|0} = P^{\star}$, and let $\gamma_k|_{k=0}^{\infty}$ be the corresponding Markov chain at its stationary state. For an arbitrary positive number d, define the set $\mathcal{B}_d = \{\ P \mid \delta(P,\ P^{\star}) \leq d\ \}$. Then, when Conditions I–III are satisfied,*

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n} I_{\mathcal{B}_d}(P_{k|k}) = F(d) \quad \textit{in probability,} \tag{5.97}$$

and the convergence rate is $O\left(\left(\frac{ln(n)}{n}\right)^{1/4}\right)$.

From Theorem 5.19 we can declare that when the stationary distribution of the random process $P_{k|k}$ is approximated by that of its samples, the approximation accuracy is of order $\left(\frac{ln(n)}{n}\right)^{1/4}$. Therefore, when a large number of samples of $P_{k|k}$ are available, the distribution function of the stationary process can be approximated with high accuracy.

5.6.3 Approximation of the Stationary Distribution

In the previous subsection, we have proved that when $P_{k|k}$ starts from $P^{\star}$ and the Markov chain γ_k is in its stationary state, the corresponding random sequence $P_{k|k}$ is ergodic. These results make it possible to approximate the stationary distribution of $P_{k|k}$ using its samples. In this section, some explicit formulas are given for approximations of this stationary distribution in which actual sampling on all $P_{k|k}$ is not required.

To investigate this approximation, the following results are first established, which makes it clear that in finite recursions, the homographic transformation of Eq. (5.88) generally cannot remove influences of its initial values.

Lemma 5.10. *Assume that Condition I is satisfied. Then, for arbitrary PDMs X and Y with compatible dimensions, $\mathbf{H}_m\left(M^{[*]}, X\right) = \mathbf{H}_m\left(M^{[*]}, Y\right)$ if and only if $X = Y$, no matter whether $* = 1$ or $* = 0$.*

Proof. By the definition of the homographic transformation direct algebraic manipulations show that when X and Y are positive definite and their dimensions are compatible,

$$\begin{aligned}\mathbf{H}_m\left(M^{[0]}, X\right) - \mathbf{H}_m\left(M^{[0]}, Y\right) &= \left[A^{[0]}XA^{[0]T} + G^{[0]}G^{[0]T}\right] - \left[A^{[0]}YA^{[0]T} + G^{[0]}G^{[0]T}\right] \\ &= A^{[0]}(X-Y)A^{[0]T}.\end{aligned} \tag{5.98}$$

Moreover,

$$\mathbf{H}_m\left(M^{[1]}, X\right) - \mathbf{H}_m\left(M^{[1]}, Y\right)$$

$$
\begin{aligned}
&= \left[(A^{[1]}XA^{[1]T} + G^{[1]}G^{[1]T})^{-1} + H^{[1]T}H^{[1]}\right]^{-1} \\
&\quad - \left[(A^{[1]}YA^{[1]T} + G^{[1]}G^{[1]T})^{-1} + H^{[1]T}H^{[1]}\right]^{-1} \\
&= \left[I + (A^{[1]}XA^{[1]T} + G^{[1]}G^{[1]T})H^{[1]T}H^{[1]}\right]^{-1} A^{[1]}(X-Y)A^{[1]T} \\
&\quad \times \left[I + H^{[1]T}H^{[1]}(A^{[1]}YA^{[1]T} + G^{[1]}G^{[1]T})\right]^{-1}. \qquad (5.99)
\end{aligned}
$$

The conclusions are immediate from these relations and the regularity of both $A^{[0]}$ and $A^{[1]}$. This completes the proof. □

Assume that the Markov chain γ_k is in its stationary state and $P_{k|k}$ starts from $P^\star$. Let $P_{0|0},\ P_{1|1},\ \cdots,\ P_{n|n}$ be its first $n+1$ samples, and consider all the possible values that these samples may take and the probability of their occurrence. Obviously by Lemma 5.10, when Condition I is satisfied, there are 2^k possible values that $P_{k|k}$ may take, which is in accordance with all the realizations of the Markov chain $\gamma_i|_{i=1}^{k}$ with $\gamma_i \in \{0,\ 1\}$. Recall that $\mathbf{H}_m\left(M^{[1]},\ P^\star\right) = P^\star$. It is clear that, for every positive integer k,

$$
\begin{aligned}
\mathbf{H}_m\left(M^{[1]k},\ P^\star\right) &= \mathbf{H}_m\left[M^{[1](k-1)}, \mathbf{H}_m\left(M^{[1]},\ P^\star\right)\right] \\
&= \mathbf{H}_m\left(M^{[1](k-1)},\ P^\star\right) = \cdots \\
&= P^\star. \qquad (5.100)
\end{aligned}
$$

On the other hand, let $P^{\S}$ denote the positive definite solution of the algebraic Lyapunov equation $P = A^{[0]}PA^{[0]T} + G^{[0]}G^{[0]T}$, provided that it does exist. This happens when the matrix $A^{[0]}$ is stable [27,28]. From Lemma 5.10 it is clear that if $P \notin \{P^\star,\ P^{\S}\}$, then $\mathbf{H}_m\left(M^{[*]},\ P\right) \neq P$, no matter whether $*$ is equal to 1 or 0.

From these arguments the following results can be obtained; their proofs are included in the appendix.

Lemma 5.11. *Let $\bar{\mathcal{P}}^{[n]}$ denote the set consisting of all possible values that $P_{k|k}|_{k=0}^{n}$ may take when it has its initial value $P^\star$ and recursively updates according to the stationary process of the Markov chain γ_k. Then the number of the elements in $\bar{\mathcal{P}}^{[n]}$ is equal to 2^n, and the set $\bar{\mathcal{P}}^{[n]}$ can be expressed as*

$$
\bar{\mathcal{P}}^{[n]} = \left\{P^\star,\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right\} \bigcup \left\{ P \left| \begin{array}{l} P = \mathbf{H}_m\left[M^{[\gamma_k]}M^{[\gamma_{k-1}]}\cdots M^{[\gamma_2]}, \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right] \\ \gamma_j \in \{0, 1\},\ j \in \{2, 3, \cdots, k\},\ k \in \{2, 3, \cdots, n\} \end{array} \right. \right\}. \qquad (5.101)
$$

For any sequence $\gamma_j|_{j=1}^{k}$ with $\gamma_j \in \{0,\ 1\}$ and $k \in \{1, 2, \cdots, n-1\}$, define $l(\gamma_j|_{j=1}^{k})$ and $\bar{P}^{[l(\gamma_j|_{j=1}^{k})]}$ respectively as $l(\gamma_j|_{j=1}^{k}) = 1 + 2^{k-1} + \sum_{j=1}^{k-1} \gamma_j 2^{j-1}$ and $\bar{P}^{[l(\gamma_j|_{j=1}^{k})]} = \mathbf{H}_m \left(M^{[\gamma_k]} M^{[\gamma_{k-1}]} \cdots M^{[\gamma_1]} M^{[0]},\ P^{\star} \right)$. Moreover, define $\bar{P}^{[1]} = P^{\star}$. Clearly, $l(\gamma_j|_{j=1}^{k})$ provides a natural order for each element of the set $\bar{\mathcal{P}}^{[n]}$. From the proof of Lemma 5.11 we see that $\mathcal{P}^{[n]} = \{\bar{P}^{[1]},\ \bar{P}^{[2]},\ \cdots,\ \bar{P}^{[2^n]}\}$.

The following theorem gives the convergence value of $\frac{1}{n+1} \sum_{k=0}^{n} I_{\mathcal{B}_d}(P_{k|k})$, which is helpful in deriving approximations for the stationary distribution of $P_{k|k}$. Its proof is given in the appendix.

Theorem 5.20. *For each prescribed positive d, define the set $\mathcal{N}_d = \left\{ j \mid \bar{P}^{[j]} \in \bar{\mathcal{P}}^{[n]} \bigcap \mathcal{B}_d \right\}$. Moreover, denote $\frac{1-\beta}{2-\alpha-\beta}$ by γ_{st}. Then*

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} I_{\mathcal{B}_d}(P_{k|k}) \\ = \quad & \lim_{n\to\infty} \sum_{j\in\mathcal{N}_d} \left(1 - \gamma_{st}^{n-\lceil log_2(j)\rceil}\right) \gamma_{st}^{\sum_{i=1}^{\lceil log_2(j)\rceil} \gamma_i^{[j]}} (1-\gamma_{st})^{\lceil log_2(j)\rceil - \sum_{i=1}^{\lceil log_2(j)\rceil} \gamma_i^{[j]}}, \end{aligned} \tag{5.102}$$

where $\gamma_i^{[j]}$ is the binary code for $j - 1 - 2^{\lceil log_2(j)\rceil - 1}$.

In this theorem, an explicit formula is given for the stationary distribution of the random process $P_{k|k}$. In principle, its value can be computed for each prescribed d, which means that this distribution function can be obtained with arbitrary accuracy, provided that a computer is available with sufficient computation speed and memory capacity. Note that the value of 2^n increases exponentially with the increment of the sample size n, and a large n is generally appreciated as it leads to a more accurate approximation. It appears reasonable to claim that, in general, conclusions of the theorem cannot be directly utilized in actual computations, and some other computationally more efficient approximations are still required.

From Eq. (5.102), however, it is obvious that when γ_{st} is approximately equal to 1, any element $P^{[j]}$ of the set $\bar{\mathcal{P}}^{[n]}$ with large $\lceil \log_2(j) \rceil - \sum_{i=1}^{\lceil \log_2(j)\rceil} \gamma_i^{[j]}$ has a small probability to occur. Note that this number is in fact equal to that of the zeros in the associated binary sequence. This result has some nice physical interpretations, as these samples are associated with many data transmission failures that can hardly happen if $\gamma_{st} \approx 1$. On the other hand, from the proof of Theorem 5.18 it is clear that with the increment of the length of the binary sequence, which is equivalent to the increment of the number j, the probability that it has a large number of zeros also increases. These mean that the contributions of an element $P^{[j]} \in \bar{\mathcal{P}}^{[n]}$ with large j to the stationary distribution of the random process $P_{k|k}$ are usually very small and therefore negligible.

On the other hand, note that $\mathbf{H}_m\left(M^{[0]}, X\right) = A^{[0]}XA^{[0]T} + G^{[0]}G^{[0]T}$. It is straightforward to prove from the definition of the Riemannian distance that, for an arbitrary PDM X, there exist finite positive numbers a and b that do not depend on the matrix X such that

$$\delta\left[\mathbf{H}_m\left(M^{[0]}, X\right),\ P^\star\right] \le a\delta(X, P^\star) + b. \tag{5.103}$$

On the basis of this inequality, we establish the following property for the matrix set $\bar{\mathcal{P}}^{[n]}$. Its proof is deferred to the appendix.

Theorem 5.21. *For a prescribed positive integer n (including $+\infty$), define the set $\tilde{\mathcal{P}}^{[n]} = \{\ P \mid P = \ \mathbf{H}_m\left(M^{[0]i},\ P^\star\right),\ i = 0, 1, \cdots, n\ \}$. Then, when Conditions I–III are satisfied,*

$$\lim_{\alpha_{1h}\to 0}\ \sup_{P\in\bar{\mathcal{P}}^{[n]}}\ \inf_{Q\in\tilde{\mathcal{P}}^{[n]}} \delta(P,\ Q) = 0. \tag{5.104}$$

In addition, it is obvious from the definition of the set $\tilde{\mathcal{P}}^{[n]}$ that if $P_{0|0} = P^\star$, then $P_{k|k} \notin \tilde{\mathcal{P}}^{[n]}$ only if $k \ge 2$ and there are more than two temporal instants at which the communication channel fails. Hence, we can directly prove from Eq. (5.A.40) that, for every integer n greater than 2,

$$\mathbf{P}_r\left\{\sup_{P\in\bar{\mathcal{P}}^{[n]}}\ \inf_{Q\in\tilde{\mathcal{P}}^{[n]}} \delta(P,\ Q) > 0\right\} \le 1 - \gamma_{st}^{n-1}(n + \gamma_{st} - n\gamma_{st}). \tag{5.105}$$

From this inequality and Theorems 5.20 and 5.21 we can declare that when α_{1h} is small and/or γ_{st} is close to 1, there certainly exits a finite integer n such that with a high probability, all the samples of the random process $P_{k|k}$ in its stationary state are concentrated around the elements of the set $\tilde{\mathcal{P}}^{[n]}$, and a matrix far away from this set usually has a negligible probability to occur. These concentrations become more dominating if both a and b are not very large and α_{1h} is significantly smaller than 1, which can be understood from Eqs. (5.A.45)–(5.A.49).

From these observations it seems reasonable to approximate the support of $P_{\infty|\infty}$ by the set $\tilde{\mathcal{P}}^{[\infty]}$. When this approximation is valid, a very simple explicit formula can be derived for the stationary distribution of the random process $P_{k|k}$, which is given in the next theorem. Its proof is deferred to the appendix.

Theorem 5.22. *Assume that α_{1h} is sufficiently small and/or γ_{st} is sufficiently close to 1 such that the set $\tilde{\mathcal{P}}^{[\infty]}$ is a good approximation for the support of the stationary process of $P_{k|k}$. Then*

$$\mathbf{P}_r\left\{P_{\infty|\infty} = \mathbf{H}_m\left(M^{[0]i}, P^\star\right)\right\} \approx \gamma_{st}(1 - \gamma_{st})^i,\ i = 0, 1, 2, \ldots. \tag{5.106}$$

Note that when $\gamma_{st} \approx 1$, $\gamma_{st}(1-\gamma_{st})^i$ decreases rapidly to 0 with the increment of the index i. This means that when the data arrival probability is high, only a few elements of the set $\mathcal{P}^{[\infty]}$, that is, $\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)$, are required in computing the approximation for the stationary distribution of the random process $P_{k|k}$. Another attractive characteristic of this approximation is that its accuracy does not depend on the length of time series samples and therefore can greatly reduce computation burdens.

Note also that the Kalman filter with data loss never outperforms that without data loss. This means that always $P_{\infty|\infty} - P^{\star} \geq 0$ [3,10,29,30]. As practical communication channels usually do not have a high data loss probability, we can declare from Theorem 5.22 that $\delta\left(\mathbf{H}_m\left(M^{[0]},\ P^{\star}\right),\ P^{\star}\right)$ is a good index for steady performance deterioration due to data loss. Moreover, to reduce the influence of data loss on state estimation accuracy and in addition to improve communication qualities, another possible method is to design a plant that makes this index as small as possible, which may be achieved through measurement variable selection and so on.

When the random data loss is described by a Bernoulli process and the data arrival probability is approximately equal to 1, it is suggested in [10] to approximate the stationary distribution of the CMEE of the Kalman filter with a delta function concentrated on $P^{\star}$. Although this approximation is intuitively understandable, its approximation accuracy is still not clear, as there are still no computationally efficient methods for determining parameters of the required rate function. On the other hand, the results of Theorems 5.21 and 5.22 suggest that rather than a delta function, it is more appropriate to approximate this stationary distribution with a series of delta functions. This difference has some significant influences on steady update gain matrix selection of the state estimator.

More precisely, a well-known fact about the Kalman filter is that to reduce computational complexity, its update gain matrix can be replaced by the steady value without any influence on its steady estimation accuracy. Note that in both the Kalman filter and the robust state estimator given in the previous section, the update gain matrix is equal to $P_{k+1|k+1}C_{k+1}^T(0)R_{k+1}^{-1}$ when a measurement arrives. Here $C_k(0)$ and R_k are respectively the nominal value of the plant output matrix and the covariance matrix of measurement errors at the time instant k. This means that when the nominal plant is time invariant, distribution of the update gain matrix of the estimator is completely determined by that of $P_{k+1|k+1}$. Therefore, the results of [10] suggest that when the measurement arrival probability is close to 1, the update gain matrix of the Kalman filter can be replaced by $P^{\star}C_{k+1}^T(0)R_{k+1}^{-1}$ without significant sacrifice of steady estimation accuracy, which is consistent with the aforementioned conclusions about conventional Kalman filtering. Theorem 5.22, however, suggests that this simple update gain matrix replacement generally deteriorates steady estimation accuracy, no matter whether the

Kalman filter or the robust state estimator of the previous section is utilized. This deterioration usually becomes more significant with the increment of the distance between $P^{\star}$ and $\mathbf{H}_m\left(M^{[0]},\ P^{\star}\right)$.

Although Theorem 5.22 provides a very simple approximation and numerical simulations reported in [11] show that this approximation is usually very accurate, it is still a challenging problem to derive a tighter bound on its approximation errors and more explicit conditions on system parameters under which this approximation is valid.

In this section, we investigate asymptotic properties for both the CMEE of the Kalman filter and the PCM of a robust recursive state estimator under the situation that the data loss process is described by a Markov chain. It appears that when the associated plant is both controllable and observable, this random matrix process converges exponentially to a stationary process that does not depend on its initial value. Moreover, when these state estimators start from the stabilizing solution of the algebraic Riccati equation defined by the system parameters of the associated plant, we show that this random matrix process becomes ergodic. An important observation is that when the data arrival probability is approximately equal to 1, the distribution of the corresponding stationary process can be well approximated by a set of delta functions. Based on these results, we derive two approximations for their stationary distributions together with an error bound for one of these two approximations. It has been made clear that replacement of the update gain of these estimators by a constant matrix usually sacrifices steady estimation accuracy even if the measurement arrival probability is close to 1. Numerical simulations show that these approximations usually have a high accuracy.

5.7 Bibliographic Notes

The network-theoretic approach has been widely used to model control/estimation over time-varying channels. In this case, the channel uncertainty is modeled as random packet losses. Packets are considered as single entities and can be lost stochastically with some probability. There are two typical statistical processes to model the packet loss process of the observed data that are used to estimate the state of a stochastic system. One is an independent and identically distributed (i.i.d.) binary process [3,17,24,36], and the other is a Markov process [9,13,16,37]. This chapter focused on the problem how the packet losses as an i.i.d. process affect the estimate of the system state due to the unreliability of the network channels.

Appendix 5.A

5.A.1 Proof of Theorem 5.18

Let $*$ be any element of the set $\{\ 0,\ \ 1\ \}$. We can declare from the definitions of α_{0h} and α_{1h} that, for every PDM pair X and Y, $\delta\left(\mathbf{H}_m(M^{[*]}, X),\ \mathbf{H}_m(M^{[*]}, Y)\right) \leq \alpha_{1h}^{*}\alpha_{0h}^{1-*}\delta(X, Y)$.

Hence, for arbitrary positive definite matrices X and Y,

$$\begin{aligned}
\rho_k(X,Y) &= \delta\left\{\mathbf{H}_m\left(\prod_{i=k}^{1} M^{[\gamma_i]}, X\right), \mathbf{H}_m\left(\prod_{i=k}^{1} M^{[\gamma_i]}, Y\right)\right\} \\
&= \delta\left\{\mathbf{H}_m\left[M^{[\gamma_k]},\ \mathbf{H}_m\left(\prod_{i=k-1}^{1} M^{[\gamma_i]},\ X\right)\right],\ \mathbf{H}_m\left[M^{[\gamma_k]},\ \mathbf{H}_m\left(\prod_{i=k-1}^{1} M^{[\gamma_i]},\ Y\right)\right]\right\} \\
&\le \alpha_{1h}^{\gamma_k}\alpha_{0h}^{1-\gamma_k}\delta\left\{\mathbf{H}_m\left(\prod_{i=k-1}^{1} M^{[\gamma_i]}, X\right), \mathbf{H}_m\left(\prod_{i=k-1}^{1} M^{[\gamma_i]}, Y\right)\right\} \\
&= \alpha_{1h}^{\gamma_k}\alpha_{0h}^{1-\gamma_k}\delta\left\{\mathbf{H}_m\left[M^{[\gamma_{k-1}]},\ \mathbf{H}_m\left(\prod_{i=k-2}^{1} M^{[\gamma_i]},\ X\right)\right],\right. \\
&\qquad \left.\mathbf{H}_m\left[M^{[\gamma_{k-1}]},\ \mathbf{H}_m\left(\prod_{i=k-2}^{1} M^{[\gamma_i]},\ Y\right)\right]\right\} \\
&\le \alpha_{1h}^{\gamma_k}\alpha_{0h}^{1-\gamma_k}\alpha_{1h}^{\gamma_{k-1}}\alpha_{0h}^{1-\gamma_{k-1}}\delta\left\{\mathbf{H}_m\left(\prod_{i=k-2}^{1} M^{[\gamma_i]},\ X\right),\ \mathbf{H}_m\left(\prod_{i=k-2}^{1} M^{[\gamma_i]},\ Y\right)\right\} \\
&= \cdots \\
&\le \left(\prod_{i=k}^{1}\alpha_{1h}^{\gamma_i}\alpha_{0h}^{1-\gamma_i}\right)\delta(X,Y) \\
&= \alpha_{1h}^{\sum_{i=1}^{k}\gamma_i}\alpha_{0h}^{k-\sum_{i=1}^{k}\gamma_i}\delta(X,Y) \\
&\le \alpha_{1h}^{\sum_{i=1}^{k}\gamma_i}\delta(X,Y). \qquad (5.A.1)
\end{aligned}$$

Define the function $f(\cdot)$ on the random process γ_k as $f(\gamma_k)=\gamma_k$. When both α and β belong to $(0,\ 1)$, it is obvious that the Markov chain γ_k is positive recurrent and only has two states, that is, $\gamma_k=1$ and $\gamma_k=0$. Using the symbols of Lemma 5.8, we obviously see that, for arbitrary $j\in\{0,\ 1\}$,

$$f_\alpha^{[j]}=\sum_{k=\tau_\alpha^{[j]}}^{\tau_{\alpha+1}^{[j]}-1} f(\gamma_k)\le(\tau_{\alpha+1}^{[j]}-1)-(\tau_\alpha^{[j]}-1)=\tau_{\alpha+1}^{[j]}-\tau_\alpha^{[j]}. \qquad (5.A.2)$$

From this relation and properties of Markov chains it is straightforward to prove that

$$s(f)=\frac{1}{\mu_1}>0,\ \mathbf{E}\left(|f_\alpha^{[j]}|^3\right)\le\mathbf{E}\left(|\tau_{\alpha+1}^{[j]}-\tau_\alpha^{[j]}|^3\right)<\infty, \qquad (5.A.3)$$

$$\mathbf{V}_{ar}\{f_{\alpha}^{[j]} - s(f)(\tau_{\alpha+1}^{[j]} - \tau_{\alpha}^{[j]})\} > 0. \tag{5.A.4}$$

Moreover, both μ_1 and π_1 are positive constants. Hence, by Lemma 5.8 we have that

$$\sup_{t \in \mathcal{R}} \left| \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < t \right\} - \phi(t) \right| = O\left[\left(\frac{\ln(k)}{k} \right)^{1/4} \right]. \tag{5.A.5}$$

From this equation we can declare that, for arbitrary positive ε_1, there exists a positive integer $N_1(\varepsilon_1)$ such that

$$\left| \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < t \right\} - \phi(t) \right| \leq \varepsilon_1 \tag{5.A.6}$$

for every real t, provided that $k \geq N_1(\varepsilon_1)$.

Therefore, when $k \geq N_1(\varepsilon_1)$, we have the following relations:

$$-\varepsilon_1 \leq \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < -t \right\} - \phi(-t) \leq \varepsilon_1, \tag{5.A.7}$$

$$-\varepsilon_1 \leq \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < t \right\} - \phi(t) \leq \varepsilon_1, \tag{5.A.8}$$

which further leads to the following inequality for all positive t:

$$\begin{aligned}
& \mathbf{P}_r \left\{ \left| \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) \right| < t \right\} \\
= \; & \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < t \right\} - \mathbf{P}_r \left\{ \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) < -t \right\} \\
\geq \; & [\phi(t) - \varepsilon_1] - [\phi(-t) + \varepsilon_1] \\
= \; & \phi(t) - \phi(-t) - 2\varepsilon_1.
\end{aligned} \tag{5.A.9}$$

On the other hand, the inequality $\left| \frac{\sqrt{\mu_1}}{\sigma_1 \sqrt{k}} \left(\sum_{i=0}^{k} \gamma_i - \frac{k+1}{\mu_1} \right) \right| < t$ is equivalent to

$$\left| \sum_{i=0}^{k} \gamma_i - (k+1)\pi_1 \right| < \sigma_1 t \sqrt{\pi_1} \sqrt{k}, \tag{5.A.10}$$

which implies that

$$\begin{aligned}\sum_{i=0}^{k} \gamma_i &> (k+1)\pi_1 - \sigma_1 t\sqrt{\pi_1}\sqrt{k} \\ &= \sqrt{k}\left(\frac{k+1}{\sqrt{k}}\sqrt{\pi_1} - \sigma_1 t\right)\sqrt{\pi_1}. \qquad (5.A.11)\end{aligned}$$

Note that $\pi_1 > 0$ and it is independent of k. It is obvious that $\frac{k+1}{\sqrt{k}}\sqrt{\pi_1} - \sigma_1 t$ is an increasing function of k. This means that, for an arbitrary positive t, there exists a positive integer $N_2(t)$ such that $\frac{k+1}{\sqrt{k}}\sqrt{\pi_1} - \sigma_1 t > 0$ for each $k \geq N_2(t)$.

Define $N_2(t)$ and $\xi(t)$ respectively as

$$\begin{aligned} N_2(t) &= \min\left\{k \,\middle|\, k \text{ is an integer}, \ \frac{k+1}{\sqrt{k}}\sqrt{\pi_1} - \sigma_1 t > 0\right\}, \\ \xi(t) &= \left(\frac{N_2(t)+2}{\sqrt{N_2(t)+1}}\sqrt{\pi_1} - \sigma_1 t\right)\sqrt{\pi_1}. \end{aligned}$$

Then it is obvious that, for arbitrary $k \geq N_2(t)+1$, $\left(\frac{k+1}{\sqrt{k}}\sqrt{\pi_1} - \sigma_1 t\right)\sqrt{\pi_1} \geq \xi(t) > 0$, which further leads to

$$\sum_{i=0}^{k} \gamma_i > \sqrt{k}\xi(t). \qquad (5.A.12)$$

In addition, by the definition of the function $\phi(t)$ or the properties of a normal distribution we can declare that, for arbitrary $\varepsilon_2 > 0$, there exists $t(\varepsilon_2) > 0$ such that

$$\phi[t(\varepsilon_2)] - \phi[-t(\varepsilon_2)] \geq 1 - \varepsilon_2. \qquad (5.A.13)$$

Now, for arbitrary $\varepsilon > 0$, let $\varepsilon_1 = \frac{\varepsilon}{4}$ and $\varepsilon_2 = \frac{\varepsilon}{2}$. Define $N(\varepsilon) = \max\{N_1(\varepsilon_1),\ N_2(t(\varepsilon_2)) + 1\}$. Then from Eqs. (5.A.9) and (5.A.13) we have that, for $k > N(\varepsilon)$,

$$\mathbf{P}_r\left\{\left|\frac{\sqrt{\mu_1}}{\sigma_1\sqrt{k}}\left(\sum_{i=0}^{k}\gamma_i - \frac{k+1}{\mu_1}\right)\right| < t(\varepsilon_2)\right\} \geq 1 - \frac{\varepsilon}{2} - 2 \times \frac{\varepsilon}{4} = 1 - \varepsilon. \qquad (5.A.14)$$

Based on this relation and Eq. (5.A.12), we can declare that

$$\begin{aligned}\mathbf{P}_r\left\{\sum_{i=0}^{k}\gamma_i > \sqrt{k}\xi\left[t\left(\frac{\varepsilon}{2}\right)\right]\right\} &\geq \mathbf{P}_r\left\{\left|\frac{\sqrt{\mu_1}}{\sigma_1\sqrt{k}}\left(\sum_{i=0}^{k}\gamma_i - \frac{k+1}{\mu_1}\right)\right| < t(\varepsilon_2)\right\} \\ &\geq 1 - \varepsilon. \qquad (5.A.15)\end{aligned}$$

A combination of this inequality and Eq. (5.A.1) makes it clear that if $k \geq N(\varepsilon)$, then, with a probability greater than $1 - \varepsilon$,

$$\rho_k(X, Y) \leq \alpha_{1h}^{\sqrt{k}\xi[t(\varepsilon/2)]}\delta(X, Y) = e^{-\sqrt{k}\xi[t(\varepsilon/2)]ln\frac{1}{\alpha_{1h}}}\delta(X, Y). \tag{5.A.16}$$

As $0 \leq \alpha_{1h} < 1$ and $\delta(X, Y)$ is a finite positive number when both X and Y are finite PDMs, we can declare that $\lim_{k\to\infty} \alpha_{1h}^{\sqrt{k}\xi[t(\varepsilon/2)]} = 0$. Since $\delta_k(X, Y)$ is nonnegative and ε is an arbitrary positive number, these relations mean that for any finite PDMs X and Y, $\lim_{k\to\infty} \rho_k(X, Y) = 0$ in probability. This completes the proof. □

5.A.2 Proof of Theorem 5.19

Assume that $P^{[j]} = \lim_{k\to\infty} \mathbf{H}_m \left(M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k}^{[j]}]},\ P^\star\right)$. Then, for $* = 0$ and $* = 1$, we have that

$$\begin{aligned}
\mathbf{H}_m\left(M^{[*]}, P^{[j]}\right) &= \mathbf{H}_m\left(M^{[*]}, \lim_{k\to\infty} \mathbf{H}_m\left(M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k}^{[j]}]},\ P^\star\right)\right) \\
&= \mathbf{H}_m\left(M^{[*]} \lim_{k\to\infty} M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k}^{[j]}]},\ P^\star\right) \\
&= \lim_{k\to\infty} \mathbf{H}_m\left(M^{[*]}M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k}^{[j]}]},\ P^\star\right) \\
&= \lim_{k\to\infty} \mathbf{H}_m\left(M^{[*]}M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k+1}^{[j]}]}, \mathbf{H}_m\left(M^{[\gamma_{-k}^{[j]}]}, P^\star\right)\right) \\
&= \lim_{k\to\infty} \mathbf{H}_m\left(M^{[*]}M^{[\gamma_0^{[j]}]}M^{[\gamma_{-1}^{[j]}]}\dots M^{[\gamma_{-k+1}^{[j]}]}, P^\star\right) \text{ (in prob.)} \\
&= P^{[\bar{j}]},
\end{aligned} \tag{5.A.17}$$

where $j = n(\gamma_i^{[j]}|_{i=0}^{-\infty}) = \sum_{i=0}^{-\infty} \gamma_i^{[j]}2^i$, $\bar{j} = n(\bar{\gamma}_i^{[j]}|_{i=0}^{-\infty}) = \sum_{i=0}^{-\infty} \bar{\gamma}_i^{[j]}2^i$, and $\bar{\gamma}_i^{[j]} = \gamma_{i+1}^{[j]}$ whenever $i \leq -1$, whereas $\bar{\gamma}_0^{[j]} = *$.

Define $n_{in}(*, \gamma_i^{[j]}|_{i=0}^{-\infty})$ as $n_{in}(*, \gamma_i^{[j]}|_{i=0}^{-\infty}) = \bar{j} - j$. Then

$$\begin{aligned}
n_{in}(*, \gamma_i^{[j]}|_{i=0}^{-\infty}) &= \sum_{i=0}^{-\infty} \bar{\gamma}_i^{[j]}2^i - \sum_{i=0}^{-\infty} \gamma_i^{[j]}2^i \\
&= (* - \gamma_0^{[j]}) + \sum_{i=-1}^{-\infty} (\gamma_{i+1}^{[j]} - \gamma_i^{[j]})2^i.
\end{aligned} \tag{5.A.18}$$

For a given sequence $\gamma_l^{[j]}|_{l=1}^{s}$ with $\gamma_l^{[j]} \in \{0,\ 1\}$, define $\gamma_{i,l}^{[j]}$ as $\gamma_{i,0}^{[j]} = \gamma_i^{[j]}$, $i = 0, -1, \ldots,$ and

$$\gamma_{i,l}^{[j]} = \begin{cases} \gamma_{i+1,l-1}^{[j]}, & i \leq -1, \\ \gamma_l^{[j]}, & i = 0, \end{cases} \qquad l = 1, 2, \ldots, s. \tag{5.A.19}$$

Denote $\mathbf{H}_m\left(M^{[\gamma_l^{[j]}]}M^{[\gamma_{l-1}^{[j]}]}\cdots M^{[\gamma_1^{[j]}]},\ P^{[j]}\right)$ by $P^{[jl]}$, $l = 1, 2, \ldots, s$. Then, a repetitive utilization of Eq. (5.A.17) leads to

$$\begin{aligned} j_s &= j_{s-1} + n_{in}(\gamma_s^{[j]}, \gamma_{i,s-1}^{[j]}|_{i=0}^{-\infty}) \\ &= j_{s-2} + n_{in}(\gamma_{s-1}^{[j]}, \gamma_{i,s-2}^{[j]}|_{i=0}^{-\infty}) + n_{in}(\gamma_s^{[j]}, \gamma_{i,s-1}^{[j]}|_{i=0}^{-\infty}) \\ &= \cdots \\ &= j_0 + \sum_{l=1}^{s} n_{in}(\gamma_l^{[j]}, \gamma_{i,l-1}^{[j]}|_{i=0}^{-\infty}), \end{aligned} \tag{5.A.20}$$

where $j_0 = j$. Therefore $j_s = j$ if and only if

$$\sum_{l=1}^{s} n_{in}(\gamma_l^{[j]}, \gamma_{i,l-1}^{[j]}|_{i=0}^{-\infty}) = 0. \tag{5.A.21}$$

On the other hand, from the definition of $n_{in}(*, \gamma_i^{[j]}|_{i=0}^{-\infty})$ straightforward algebraic manipulations show that

$$\begin{aligned} \sum_{l=1}^{s} n_{in}(\gamma_l^{[j]}, \gamma_{i,l-1}^{[j]}|_{i=0}^{-\infty}) &= \sum_{l=1}^{s}\left[(\gamma_l^{[j]} - \gamma_{0,l-1}^{[j]}) + \sum_{i=-1}^{-\infty}(\gamma_{i+1,l-1}^{[j]} - \gamma_{i,l-1}^{[j]})2^i\right] \\ &= \sum_{l=1}^{s}(\gamma_l^{[j]} - \gamma_{0,l-s}^{[j]})2^{l-s} + \sum_{i=0}^{-\infty}(\gamma_{0,i}^{[j]} - \gamma_{0,i-s}^{[j]})2^{i-s}. \end{aligned} \tag{5.A.22}$$

Therefore $j_s = j$ if and only if

$$\begin{aligned} \gamma_s^{[j]} &= \gamma_{0,0}^{[j]} - \sum_{l=1}^{s-1}(\gamma_l^{[j]} - \gamma_{0,l-s}^{[j]})2^{l-s} - \sum_{i=0}^{-\infty}(\gamma_{0,i}^{[j]} - \gamma_{0,i-s}^{[j]})2^{i-s} \\ &= \gamma_{0,0}^{[j]} + \sum_{l=1}^{s-1}(\gamma_{0,l-s}^{[j]} - \gamma_l^{[j]})2^{l-s} + \sum_{i=0}^{-\infty}(\gamma_{0,i-s}^{[j]} - \gamma_{0,i}^{[j]})2^{i-s} \\ &= \gamma_{0,0}^{[j]} + \sum_{i=-1}^{1-s}\gamma_{0,i}^{[j]}2^i + 2^{-s}\sum_{i=0}^{-\infty}(\gamma_{0,i-s}^{[j]} - \gamma_{0,i}^{[j]})2^i - \sum_{i=1}^{s-1}\gamma_{s-i}^{[j]}2^{-i}. \end{aligned} \tag{5.A.23}$$

Define the set

$$\mathcal{S}^{[j]} = \left\{ k \left| k = \min(s), \begin{array}{l} \gamma_{0,0}^{[j]} + \sum_{i=-1}^{1-s} \gamma_{0,i}^{[j]} 2^i + 2^{-s} \sum_{i=0}^{-\infty} (\gamma_{0,i-s}^{[j]} - \gamma_{0,i}^{[j]}) 2^i \\ \quad - \sum_{i=1}^{s-1} \gamma_{s-i}^{[j]} 2^{-i} \in \{0, 1\} \\ \gamma_i^{[j]} \in \{0,\ 1\},\ i = 1, 2, \cdots, s-1 \end{array} \right. \right\}.$$

Assume that this set is not empty for all possible j. Then, for any $s \in \mathcal{S}^{[j]}$, there exists a binary sequence $\gamma_i^{[j]}|_{i=1}^{s}$ with $\gamma_i^{[j]} \in \{0,\ 1\}$ such that

$$\mathbf{H}_m \left(M^{[\gamma_s^{[j]}]} M^{[\gamma_{s-1}^{[j]}]} \cdots M^{[\gamma_1^{[j]}]}, P^{[j]} \right) = P^{[j]}. \tag{5.A.24}$$

Assume that the Markov chain γ_k is in its stationary state in which both $\mathbf{P}_r(\gamma_k = 1)$ and $\mathbf{P}_r(\gamma_k = 0)$ take a constant value belonging to $(0,\ 1)$. Denote $\max\{\mathbf{P}_r(\gamma_k = 1),\ \mathbf{P}_r(\gamma_k = 0)\}$ and $\min\{\mathbf{P}_r(\gamma_k = 1),\ \mathbf{P}_r(\gamma_k = 0)\}$ respectively by p_{hs} and p_{ls}. Moreover, for a particular $s \in \mathcal{S}^{[j]}$, denote the corresponding $\gamma_i^{[j]}|_{i=1}^{s}$ by $\gamma_i^{[j,s]}|_{i=1}^{s}$. Then

$$\mathbf{P}_r(s) = \prod_{i=1}^{s} \left[\gamma_i^{[j,s]} \mathbf{P}_r(\gamma_i^{[j,s]} = 1) + (1 - \gamma_i^{[j,s]}) \mathbf{P}_r(\gamma_i^{[j,s]} = 0) \right]. \tag{5.A.25}$$

Therefore

$$\mathbf{P}_r(s) \geq \prod_{i=1}^{s} \min\{\mathbf{P}_r(\gamma_k = 1),\ \mathbf{P}_r(\gamma_k = 0)\} = p_{ls}^s, \tag{5.A.26}$$

$$\mathbf{P}_r(s) \leq \prod_{i=1}^{s} \max\{\mathbf{P}_r(\gamma_k = 1),\ \mathbf{P}_r(\gamma_k = 0)\} = p_{hs}^s. \tag{5.A.27}$$

Hence, when an integer s belonging to the set $\mathcal{S}^{[j]}$ takes a finite value, its occurrence probability is certainly greater than 0.

As in Lemma 5.8, let $\tau_v^{[j]}$ denote the vth time instant such that $j_s = j_0$, and let $f_v^{[j]}(P_{k|k})$ denote the random variable $\sum_{k=\tau_v^{[j]}}^{\tau_{v+1}^{[j]}-1} I_{\mathcal{B}_d}(P_{k|k})$. Then

$$\begin{aligned} \left| f_v^{[j]}(P_{k|k}) \right| &= \left| \sum_{k=\tau_v^{[j]}}^{\tau_{v+1}^{[j]}-1} I_{\mathcal{B}_d}(P_{k|k}) \right| \leq (\tau_{v+1}^{[j]} - 1) - (\tau_v^{[j]} - 1) \\ &= \tau_{v+1}^{[j]} - \tau_v^{[j]} \in \mathcal{S}^{[j]}. \end{aligned} \tag{5.A.28}$$

Hence

$$\begin{aligned}\mathbf{E}\left\{\left|f_v^{[j]}(P_{k|k})\right|^3\right\} &\leq \mathbf{E}\left\{(\tau_{v+1}^{[j]}-\tau_v^{[j]})^3\right\} = \sum_{k\in\mathcal{S}^{[j]}} k^3\mathbf{P}_r(k) \\ &\leq \sum_{k=1}^{\infty} k^3 p_{hs}^k. \end{aligned} \tag{5.A.29}$$

Note that $k^3 = (k+1)k(k-1)+k$. We can directly prove that

$$\sum_{k=1}^{\infty} k^3 p_{hs}^k = \frac{(1+p_{hs})^2+2p_{hs}}{(1-p_{hs})^4} p_{hs}. \tag{5.A.30}$$

Therefore, when p_{hs} belongs to $(0,\ 1)$, both $\mathbf{E}\left\{\left|f_v^{[j]}(P_{k|k})\right|^3\right\}$ and $\mathbf{E}\left\{(\tau_{v+1}^{[j]}-\tau_v^{[j]})^3\right\}$ are finite.

Note also that $\mathbf{H}_m\left(M^{[1]}, P^\star\right) = P^\star$ and $\lim_{k\to\infty}\mathbf{H}_m\left[M^{[1]k}, \mathbf{H}_m\left(M^{[0]}, P^\star\right)\right] = P^\star$. It is obvious that when $j = \sum_{i=0}^{-\infty} 2^i$, the set $\mathcal{S}^{[j]}$ has at least two finite integers with occurrence probability greater than 0. Therefore, when the random process $P_{k|k}$ is started from $P^\star$, the corresponding $f_\alpha^{[j]} - s(f)(\tau_{\alpha+1}^{[j]} - \tau_\alpha^{[j]})$ has a variance greater than 0.

Denote $\sum_{k\in\mathcal{S}^{[j]}} k\mathbf{P}_r(k)$ by $\mu^{[j]}$. Then we directly prove that

$$F(d) = \sum_{j\in\mathcal{I}} \frac{1}{\mu^{[j]}} I_{\mathcal{B}_d}(P^{[j]}). \tag{5.A.31}$$

On the other hand, by Lemma 5.8 we have that

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n} I_{\mathcal{B}_d}(P_{k|k}) = \sum_{j\in\mathcal{I}}\frac{1}{\mu^{[j]}} I_{\mathcal{B}_d}(P^{[j]}) \tag{5.A.32}$$

with the convergence rate of order $\left(\frac{\ln(n)}{n}\right)^{1/4}$. Combining the last two equations together, we can now complete the proof for the case in which $\mathcal{S}^{[j]} \neq \emptyset$ for every possible j.

If there exists j such that the set $\mathcal{S}^{[j]}$ is empty, then the conclusions can still be established by modifying $P^{[j]}$ to $\mathcal{P}^{[j]}(\varepsilon)$ in the arguments, where ε is a prescribed positive number. More precisely, by Theorem 5.18, for arbitrary j_1 and j_2, there always exists a finite step transformation from an element of $\mathcal{P}^{[j_1]}(\varepsilon)$ to the set $\mathcal{P}^{[j_2]}(\varepsilon)$. Therefore, the corresponding set $\mathcal{S}^{[j]}$ is certainly not empty. The results can then be established by decreasing ε to 0. □

5.A.3 Proof of Lemma 5.11

From $P_{0|0} = P^\star$ and $\mathbf{H}_m\left(M^{[1]},\ P^\star\right) = P^\star$ it is clear that $P_{1|1}$ has only one additional possible value, that is, $\mathbf{H}_m\left(M^{[0]},\ P^\star\right)$. Hence, the number of elements in $\bar{\mathcal{P}}^{[1]}$ is 2, and $\bar{\mathcal{P}}^{[1]} = \{\ P^\star,\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\}$.

Assume that the conclusions are valid with $n = l$, that is, $\#\left(\bar{\mathcal{P}}^{[l]}\right) = 2^l$, and

$$\bar{\mathcal{P}}^{[l]} = \left\{P^\star,\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right\}\bigcup\left\{P\ \middle|\ \begin{array}{l} P = \mathbf{H}_m\left[M^{[\gamma_k]}M^{[\gamma_{k-1}]}\cdots M^{[\gamma_2]}, \mathbf{H}_m\left(M^{[0]}, P^\star\right)\right] \\ \gamma_j \in \{0, 1\},\ j \in \{2, 3, \cdots, k\},\ k \in \{2, 3, \cdots, l\} \end{array}\right\}.$$

Then, for $n = l + 1$, we have that

$$\begin{aligned} \bar{\mathcal{P}}^{[l+1]} &= \bar{\mathcal{P}}^{[l]}\bigcup\left\{\ P_{l+1}\ \middle|\ P_{l+1} = \mathbf{H}_m\left(M^{[\gamma_{l+1}]},\ P\right),\ P \in \bar{\mathcal{P}}^{[l]}\backslash\bar{\mathcal{P}}^{[l-1]},\ \gamma_{l+1} \in \{0,\ 1\}\ \right\} \\ &= \bar{\mathcal{P}}^{[l]}\bigcup\left\{P_{l+1}\ \middle|\ \begin{array}{l} P_{l+1} = \mathbf{H}_m\left\{M^{[\gamma_{l+1}]},\ \mathbf{H}_m\left[M^{[\gamma_l]}M^{[\gamma_{l-1}]}\cdots M^{[\gamma_2]}, \mathbf{H}_m\left(M^{[0]}, P^\star\right)\right]\right\} \\ \gamma_j \in \{0,\ 1\},\ j \in \{2, 3, \cdots, l+1\} \end{array}\right\} \\ &= \bar{\mathcal{P}}^{[l]}\bigcup\left\{\ P_{l+1}\ \middle|\ \begin{array}{l} P_{l+1} = \mathbf{H}_m\left[M^{[\gamma_{l+1}]}M^{[\gamma_l]}\cdots M^{[\gamma_2]},\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right] \\ \gamma_j \in \{0,\ 1\},\ j \in \{2, 3, \cdots, l+1\} \end{array}\ \right\}. \end{aligned} \qquad (5.A.33)$$

From the regularity of the matrices $A^{[0]}$ and $A^{[1]}$, by Lemma 2.6, we can directly prove that

$$\bar{\mathcal{P}}^{[l]}\bigcap\left\{P_{l+1}\ \middle|\ \begin{array}{l} P_{l+1} = \mathbf{H}_m\left[M^{[\gamma_{l+1}]}M^{[\gamma_l]}\cdots M^{[\gamma_2]},\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right] \\ \gamma_j \in \{0,\ 1\},\ j \in \{2, 3, \cdots, l+1\} \end{array}\ \right\} = \emptyset. \qquad (5.A.34)$$

Therefore, $\#(\bar{\mathcal{P}}^{[l+1]}) = \#(\bar{\mathcal{P}}^{[l]}) + 2^{l-1} = 2^l$, and

$$\bar{\mathcal{P}}^{[l+1]} = \left\{P^\star,\ \mathbf{H}_m\left(M^{[0]},\ P^\star\right)\right\}\bigcup\left\{P\ \middle|\ \begin{array}{l} P = \mathbf{H}_m\left[M^{[\gamma_k]}M^{[\gamma_{k-1}]}\cdots M^{[\gamma_2]}, \mathbf{H}_m\left(M^{[0]}, P^\star\right)\right] \\ \gamma_j \in \{0, 1\},\ j \in \{2, 3, \cdots, k\},\ k \in \{2, 3, \cdots, l+1\} \end{array}\right\}. \qquad (5.A.35)$$

This completes the proof. □

5.A.4 Proof of Theorem 5.20

First, probabilities are investigated for the occurrence of $P_{k|k} = \bar{P}^{[j]}$ with $j = 1 + 2^{s-1} + \sum_{l=1}^{s-1} \gamma_l^{[j]} 2^{l-1}$. From the definition of $\bar{P}^{[j]}$ it is obvious that $P_{k|k} = \bar{P}^{[j]}$ if and only if

$k \geq s+1$, $\gamma_l = 1$ when $l \in \{1, 2, \cdots, k-s-1\}$, $\gamma_{k-s} = 0$, and $\gamma_{i+k-s} = \gamma_i^{[j]}$ when $i \in \{1, 2, \cdots, s\}$. Hence

$$\begin{aligned}\mathbf{P}_r\left(P_{k|k} = \bar{P}^{[j]}\right) &= \mathbf{P}_r\left(\gamma_1 = 1, \cdots, \gamma_{k-s-1} = 1, \gamma_{k-s} = 0, \gamma_{k-s+1} = \gamma_1^{[j]}, \cdots, \gamma_k = \gamma_s^{[j]}\right) \\ &= \prod_{l=1}^{k-s-1} \mathbf{P}_r\left(\gamma_l = 1\right) \times \mathbf{P}_r\left(\gamma_{k-s} = 0\right) \prod_{i=1}^{s} \mathbf{P}_r\left(\gamma_{i+k-s} = \gamma_i^{[j]}\right) \\ &= \gamma_{st}^{k-s-1}(1-\gamma_{st})p_j, \end{aligned} \tag{5.A.36}$$

where $p_j = \prod_{i=1}^{s} \mathbf{P}_r\left(\gamma_{i+k-s} = \gamma_i^{[j]}\right)$.

Therefore, the occurrence of $\bar{P}^{[j]}$ in the samples $P_{0|0}$, $P_{1|1}$, $\cdots$, $P_{n|n}$ has the probability

$$\begin{aligned}\bar{p}_j &= \sum_{k=s+1}^{n} \mathbf{P}_r\left(P_{k|k} = \bar{P}^{[j]}\right) \\ &= \sum_{k=s+1}^{n} \gamma_{st}^{k-s-1}(1-\gamma_{st})p_j = (1-\gamma_{st}^{n-s})p_j. \end{aligned} \tag{5.A.37}$$

Note that when $\gamma_l^{[j]} \in \{0,\ 1\}$, $l = 1, 2, \cdots, s$, it is certain that $0 \leq \sum_{l=1}^{s-1} \gamma_l^{[j]} 2^{l-1} \leq 2^{s-1} - 1$. We therefore have that

$$1 + 2^{s-1} \leq j \leq 2^s, \tag{5.A.38}$$

which is equivalent to $1 + \log_2(j-1) \geq s \geq \log_2(j)$. As s is a positive integer, it is obvious that

$$s = \lceil \log_2(j) \rceil. \tag{5.A.39}$$

Therefore, $\gamma_l^{[j]}$ with $l \in \{1, 2, \cdots, \lceil \log_2(j) \rceil\}$ is the binary code of $j - 1 - 2^{\lceil \log_2(j) \rceil - 1}$. We can therefore declare that, for any given j belonging to $\{1, 2, \cdots, 2^n\}$, both s and $\gamma_l^{[j]}|_{l=1}^{s}$ are uniquely determined through the requirement that $j = 1 + 2^{s-1} + \sum_{l=1}^{s-1} \gamma_l^{[j]} 2^{l-1}$.

On the other hand, let $N_0(j)$ denote the number of zeros in the sequence $\gamma_i^{[j]}|_{i=1}^{s}$. Then

$$\begin{aligned} p_j &= \prod_{i=1}^{s} \mathbf{P}_r\left(\gamma_{i+k-s} = \gamma_i^{[j]}\right) \\ &= \prod_{i=1}^{s} \mathbf{P}_r^{\gamma_i^{[j]}}(\gamma_i^{[j]} = 1)\mathbf{P}_r^{(1-\gamma_i^{[j]})}(\gamma_i^{[j]} = 0) \end{aligned}$$

$$\begin{aligned} &= \gamma_{st}^{\sum_{i=1}^{s}\gamma_i^{[j]}}(1-\gamma_{st})^{s-\sum_{i=1}^{s}\gamma_i^{[j]}} \\ &= (1-\gamma_{st})^{N_0(j)}\gamma_{st}^{\lceil\log_2(j)\rceil-N_0(j)}. \end{aligned} \tag{5.A.40}$$

Summarizing Eqs. (5.A.37), (5.A.39), and (5.A.40), the following formula is obtained for $\bar{p}_j$:

$$\bar{p}_j = (1-\gamma_{st}^{n-\lceil\log_2(j)\rceil})\gamma_{st}^{\lceil\log_2(j)\rceil-N_0(j)}(1-\gamma_{st})^{N_0(j)}. \tag{5.A.41}$$

Note that $N_0(j) = s - \sum_{i=1}^{s}\gamma_i^{[j]}$. Hence, by Eq. (5.A.39), the ergodicity of the random process $P_{k|k}$ established in Corollary 5.1, and the Bernoulli law of large numbers [38] we can claim that

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n} I_{\mathcal{B}_d}(P_{k|k}) \\ = \;&\lim_{n\to\infty}\sum_{j\in\mathcal{N}_d}\bar{p}_j \\ = \;&\lim_{n\to\infty}\sum_{j\in\mathcal{N}_d}\left(1-\gamma_{st}^{n-\lceil log_2(j)\rceil}\right)\gamma_{st}^{\sum_{i=1}^{\lceil log_2(j)\rceil}\gamma_i^{[j]}}(1-\gamma_{st})^{\lceil log_2(j)\rceil-\sum_{i=1}^{\lceil log_2(j)\rceil}\gamma_i^{[j]}} \end{aligned} \tag{5.A.42}$$

with exponential convergence rate. This completes the proof. □

5.A.5 Proof of Theorem 5.21

Recall that for an arbitrary positive integer m, $\mathbf{H}_m\left(M^{[1]m},\ P^\star\right) = P^\star$. Therefore, investigating properties of samples of the random process $P_{k|k}$ starting from $P_{0|0} = P^\star$, we can assume, without any loss of generality that, that the first n elements of any realization of the random process $\gamma_k|_{k=0}^{\infty}$ is a binary sequence satisfying

$$\begin{aligned} &\gamma_{t_j} = \gamma_{t_j+1} = \cdots = \gamma_{t_j+m_j} = 0,\ \ m_j \geq 0,\ \ j = 1, 2, \cdots, p, \\ &\gamma_{t_j+m_j+1} = \gamma_{t_j+m_j+2} = \cdots = \gamma_{t_{j+1}-1} = 1,\ \ 1 = t_1 < t_2 < \cdots < t_p \leq n. \end{aligned}$$

Denote by $\hat{\mathcal{P}}^{[n]}$ the set consisting of the first n $P_{k|k}$s generated from this particular realization of the data loss process.

Concerning the Riccati recursion of Eq. (5.88), the following inequality is obtained from Eq. (5.103) for every positive integer m and every positive definite X:

$$\begin{aligned} \delta\left[\mathbf{H}_m\left(M^{[0]m}, X\right),\ P^\star\right] &= \delta\left\{\mathbf{H}_m\left[M^{[0]}, \mathbf{H}_m\left(M^{[0](m-1)}, X\right)\right],\ P^\star\right\} \\ &\leq a\delta\left[\mathbf{H}_m\left(M^{[0](m-1)}, X\right),\ P^\star\right] + b \end{aligned}$$

$$
\begin{aligned}
&= \cdots \\
&\leq a^m \delta(X, P^\star) + a^{m-1}b + a^{m-2}b + \cdots + b \\
&= a^m \delta(X, P^\star) + \frac{1-a^m}{1-a} b.
\end{aligned}
\tag{5.A.43}
$$

Moreover,

$$
\begin{aligned}
\delta\left[\mathbf{H}_m\left(M^{[1]m}, X\right),\ P^\star\right] &= \delta\left[\mathbf{H}_m\left(M^{[1]m}, X\right),\ \mathbf{H}_m\left(M^{[1]m}, P^\star\right)\right] \\
&\leq \alpha_{1h}^m \delta(X, P^\star).
\end{aligned}
\tag{5.A.44}
$$

On the basis of these two inequalities, we establish the following inequality for each $j \geq 1$ and every k belonging to $\{t_j + m_j + 1,\ t_j + m_j + 2,\ \cdots,\ t_{j+1} - 1\}$:

$$
\begin{aligned}
\delta(P_{k|k}, P^\star) &= \delta\left[\mathbf{H}_m\left(\prod_{i=k}^{1} M^{[\gamma_i]}, P^\star\right),\ P^\star\right] \\
&= \delta\left[\mathbf{H}_m\left(M^{[1](k-t_j-m_j)} M^{[0](m_j+1)} M^{[1](t_j-t_{j-1}-m_{j-1})} \cdots M^{[0](m_1+1)},\ P^\star\right),\ P^\star\right] \\
&\leq \alpha_{1h}^{k-t_j-m_j} \delta\left[\mathbf{H}_m\left(M^{[0](m_j+1)} M^{[1](t_j-t_{j-1}-m_{j-1})} \cdots M^{[0](m_1+1)},\ P^\star\right),\ P^\star\right] \\
&\leq \alpha_{1h}^{k-t_j-m_j} \left\{a^{m_j+1} \delta\left[\mathbf{H}_m\left(M^{[1](t_j-t_{j-1}-m_{j-1})} M^{[0](m_{j-1}+1)} \cdots M^{[0](m_1+1)}, P^\star\right),\right.\right. \\
&\qquad \left.\left. P^\star\right] + \frac{1-a^{m_j+1}}{1-a} b\right\} \\
&\leq \cdots \\
&\leq \alpha_{1h}^{k-t_j-m_j} \frac{1-a^{m_j+1}}{1-a} b + \alpha_{1h}^{k-t_{j-1}-(m_j+m_{j-1})} \frac{1-a^{m_{j-1}+1}}{1-a} b + \cdots \\
&\quad + \alpha_{1h}^{k-t_1-\sum_{i=1}^{j} m_i} \frac{1-a^{m_1+1}}{1-a} b.
\end{aligned}
\tag{5.A.45}
$$

Recall that $0 \leq \alpha_{1h} < 1$. We therefore have that if $0 \leq a < 1$, then

$$
\begin{aligned}
\delta(P_{k|k}, P^\star) &\leq \alpha_{1h}^{k-t_j-m_j} \frac{b}{1-a} + \alpha_{1h}^{k-t_{j-1}-(m_j+m_{j-1})} \frac{b}{1-a} + \cdots + \alpha_{1h}^{k-t_1-\sum_{i=1}^{j} m_i} \frac{b}{1-a} \\
&= \alpha_{1h}^{k-t_j-m_j} \frac{b}{1-a} \left[1 + \alpha_{1h}^{t_j-t_{j-1}-m_{j-1}} + \cdots + \alpha_{1h}^{t_j-t_1-\sum_{i=1}^{j} m_i}\right] \\
&= \alpha_{1h}^{k-t_j-m_j} \frac{jb}{1-a}.
\end{aligned}
\tag{5.A.46}
$$

Moreover, if $a = 1$, then

$$\begin{aligned}
&\delta(P_{k|k},\, P^{\star}) \\
\leq\ & \alpha_{1h}^{k-t_j-m_j}(m_j+1)b+\alpha_{1h}^{k-t_{j-1}-(m_j+m_{j-1})}(m_{j-1}+1)b+\cdots+\alpha_{1h}^{k-t_1-\sum_{i=1}^{j} m_i}(m_1+1)b \\
\leq\ & \alpha_{1h}^{k-t_j-m_j}b\left\{(m_j+1)+\alpha_{1h}^{t_j-t_{j-1}-m_{j-1}}(m_{j-1}+1)+cdots+\alpha_{1h}^{t_j-t_1-\sum_{i=1}^{j-1} m_i}(m_1+1)\right\} \\
\leq\ & \alpha_{1h}^{k-t_j-m_j}jb\max_{1\leq i\leq j}(m_i+1).
\end{aligned} \tag{5.A.47}$$

Furthermore, if $a>1$, then

$$\begin{aligned}
\delta(P_{k|k},\, P^{\star}) &\leq \alpha_{1h}^{k-t_j-m_j}\frac{a^{m_j+1}}{a-1}b+\alpha_{1h}^{k-t_{j-1}-(m_j+m_{j-1})}\frac{a^{m_{j-1}+1}}{a-1}b+\cdots+\alpha_{1h}^{k-t_1-\sum_{i=1}^{j} m_i}\frac{a^{m_1+1}}{a-1}b \\
&\leq \alpha_{1h}^{k-t_j-m_j}\frac{ab}{a-1}\left\{a^{m_j}+\alpha_{1h}^{t_j-t_{j-1}-m_{j-1}}a^{m_{j-1}}+\cdots+\alpha_{1h}^{t_j-t_1-\sum_{i=1}^{j-1} m_i}a^{m_1}\right\} \\
&\leq \alpha_{1h}^{k-t_j-m_j}\frac{jab}{a-1}\max_{1\leq i\leq j}a^{m_i}.
\end{aligned} \tag{5.A.48}$$

On the other hand, for any $k\in\{t_j,\ t_j+1,\ \cdots,\ t_j+m_j\}$ with $j=1,2,\cdots,p$, from Lemma 2.6 and Eq. (5.89), since $0\leq\alpha_{0h}\leq 1$, we have that

$$\begin{aligned}
&\delta\left[P_{k|k},\mathbf{H}_m\left(M^{[0](k-t_j+1)},\ P^{\star}\right)\right] \\
\leq\ & \delta\left[\mathbf{H}_m\left(\prod_{i=k}^{t_j}M^{[\gamma_i]},\, P_{t_j-1|t_j-1}\right),\mathbf{H}_m\left(M^{[0](k-t_j+1)},\, P^{\star}\right)\right] \\
=\ & \delta\left[\mathbf{H}_m\left(M^{[0](k-t_j+1)},\, P_{t_j-1|t_j-1}\right),\mathbf{H}_m\left(M^{[0](k-t_j+1)},\, P^{\star}\right)\right] \\
\leq\ & \alpha_{0h}^{k-t_j+1}\delta(P_{t_j-1|t_j-1},\ P^{\star}) \\
\leq\ & \delta(P_{t_j-1|t_j-1},\ P^{\star}).
\end{aligned} \tag{5.A.49}$$

Note that $\mathbf{H}_m\,(I,\ P^{\star})=P^{\star}$. We can therefore declare from Eqs. (5.A.45)–(5.A.49) that

$$\lim_{\alpha_{1h}\to 0}\ \sup_{P\in\hat{\mathcal{P}}^{[n]}}\ \inf_{Q\in\tilde{\mathcal{P}}^{[n]}}\ \delta(P,\ Q)=0. \tag{5.A.50}$$

The proof can now be completed by noting that this conclusion is valid for all possible $(t_j,m_j)|_{j=1}^{p}$ and p. □

5.A.6 Proof of Theorem 5.22

When the assumption is satisfied, assume that $\mathbf{P}_r\left\{P_{\infty|\infty}=\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right\}=a_i$, $i=0,1,\ldots$. Then, from the definition of probabilities we have that

$$\sum_{i=0}^{\infty} a_i=1. \tag{5.A.51}$$

On the other hand, note that $\mathbf{H}_m\left[M^{[0]},\ \mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right]=\mathbf{H}_m\left(M^{[0](i+1)},\ P^{\star}\right)$. Moreover, when the Markov chain achieves its stationary state, $\mathbf{P}_r(\gamma_k=1)=\gamma_{st}$. We can therefore declare that when the random process $P_{k|k}$ reaches its stationary state,

$$\mathbf{P}_r\left\{P_{n+1|n+1}=\mathbf{H}_m\left(M^{[0](i+1)},\ P^{\star}\right)\right\}=(1-\gamma_{st})\mathbf{P}_r\left\{P_{n|n}=\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right\}. \tag{5.A.52}$$

Moreover, to guarantee the stationarity of the random process, it is necessary that

$$\lim_{n\to\infty}\mathbf{P}_r\left\{P_{n+1|n+1}=\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right\}=\lim_{n\to\infty}\mathbf{P}_r\left\{P_{n|n}=\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right\}. \tag{5.A.53}$$

Therefore

$$a_{i+1}=(1-\gamma_{st})a_{i-1},\quad i=1,2,\ldots. \tag{5.A.54}$$

Substituting this relation into Eq. (5.A.51), we obtain the following equation:

$$a_0+(1-\gamma_{st})a_0+(1-\gamma_{st})^2a_0+\cdots=1. \tag{5.A.55}$$

Hence

$$a_0=\frac{1}{\sum_{i=0}^{\infty}(1-\gamma_{st})^i}=\gamma, \tag{5.A.56}$$

which further leads to

$$\mathbf{P}_r\left\{P_{\infty|\infty}=\mathbf{H}_m\left(M^{[0]i},\ P^{\star}\right)\right\}=a_0(1-\gamma_{st})^i=\gamma_{st}(1-\gamma_{st})^i. \tag{5.A.57}$$

This completes the proof. □

References

[1] T. Sui, K. You, M. Fu, D. Marelli, Stability of MMSE state estimators over lossy networks using linear coding, Automatica 51 (2015) 167–174.

[2] K. You, T. Sui, M. Fu, Kalman filtering over lossy networks under switching sensors, Asian Journal of Control 17 (2015) 45–54.

[3] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M.I. Jordan, S.S. Sastry, Kalman filtering with intermittent observations, IEEE Transactions on Automatic Control 49 (2004) 1453–1464.

[4] Y. Mo, B. Sinopoli, Towards finding the critical value for Kalman filtering with intermittent observations, http://arxiv.org/abs/1005.2442, 2010.

[5] N. De Freitas, C. Andrieu, P. Højen-Sørensen, M. Niranjan, A. Gee, Sequential Monte Carlo methods for neural networks, in: Sequential Monte Carlo Methods in Practice, Springer, 2001, pp. 359–379.

[6] A. Garulli, A. Vicino, G. Zappa, Conditional central algorithms for worst case set-membership identification and filtering, IEEE Transactions on Automatic Control 45 (2000) 14–23.

[7] J. George, Robust Kalman–Bucy filter, IEEE Transactions on Automatic Control 58 (2013) 174–180.

[8] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2012.

[9] K. You, M. Fu, L. Xie, Mean square stability for Kalman filtering with Markovian packet losses, Automatica 47 (2011) 2647–2657.

[10] S. Kar, B. Sinopoli, J.M. Moura, Kalman filtering with intermittent observations: weak convergence to a stationary distribution, IEEE Transactions on Automatic Control 57 (2012) 405–420.

[11] T. Zhou, Asymptotic behavior of recursive state estimations with intermittent measurements, IEEE Transactions on Automatic Control 61 (2016) 400–415.

[12] H. Lin, P.J. Antsaklis, Stability and stabilizability of switched linear systems: a survey of recent results, IEEE Transactions on Automatic Control 54 (2009) 308–322.

[13] L. Shi, M. Epstein, R. Murray, Kalman filtering over a packet-dropping network: a probabilistic perspective, IEEE Transactions on Automatic Control 55 (2010) 594–604.

[14] B. Anderson, J. Moore, Detectability and stabilizability of time-varying discrete-time linear systems, SIAM Journal on Control and Optimization 19 (1981) 20–32.

[15] C. Chen, Linear System: Theory and Design, Saunders College Publishing, Philadelphia, PA, USA, 1984.

[16] L. Xie, L. Xie, Stability analysis of networked sampled-data linear systems with Markovian packet losses, IEEE Transactions on Automatic Control 54 (2009) 1368–1374.

[17] K. Plarre, F. Bullo, On Kalman filtering for detectable systems with intermittent observations, IEEE Transactions on Automatic Control 54 (2009) 386–390.

[18] V. Gupta, N. Martins, J. Baras, Optimal output feedback control using two remote sensors over erasure channels, IEEE Transactions on Automatic Control 54 (2009) 1463–1476.

[19] L. Schenato, Optimal estimation in networked control systems subject to random delay and packet drop, IEEE Transactions on Automatic Control 53 (2008) 1311–1317.

[20] B.D. Anderson, J.B. Moore, Optimal filtering, Englewood Cliffs 21 (1979) 22–95.

[21] S.M. Mohamed, S. Nahavandi, Robust finite-horizon Kalman filtering for uncertain discrete-time systems, IEEE Transactions on Automatic Control 57 (2012) 1548–1552.

[22] T. Zhou, H.Y. Liang, On asymptotic behaviors of a sensitivity penalization based robust state estimator, Systems & Control Letters 60 (2011) 174–180.

[23] P. Bougerol, Kalman filtering with random coefficients and contractions, SIAM Journal on Control and Optimization 31 (1993) 942–959.

[24] A. Censi, Kalman filtering with intermittent observations: convergence for semi-Markov chains and an intrinsic performance measure, IEEE Transactions on Automatic Control 56 (2011) 376–381.

[25] H. Kimura, Chain-Scattering Approach to H^{∞}-Control, Birkhauser, Boston, USA, 1997.

[26] H. Kimura, Chain-Scattering Approach to H^{∞}-Control, Springer Science & Business Media, 1996.

[27] T. Kailath, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[28] D. Simon, Optimal State Prediction: Kalman, H_∞ and Nonlinear Approaches, Wiley-Interscience, A John Wiley & Sons, Inc., Publication, Hoboken, New Jersey, USA, 2006.

[29] T. Zhou, Robust recursive state estimation with random measurement droppings, IEEE Transactions on Automatic Control 61 (2016) 156–171.

[30] Y.L. Mo, B. Sinopoli, Kalman filtering with intermittent observations: tail distribution and critical value, IEEE Transactions on Automatic Control 57 (2012) 677–689.

[31] T. Zhou, Sensitivity penalization based robust state estimation for uncertain linear systems, IEEE Transactions on Automatic Control 55 (2010) 1018–1024.

[32] D. Landers, L. Rogge, On the rate of convergence in the central limit theorem for Markov-chains, Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete 35 (1976) 57–63.

[33] O. Stenflo, A survey of average contractive iterated function systems, Journal of Differential Equations and Applications 18 (2012) 1355–1380.

[34] J.H. Elton, An ergodic theorem for iterated maps, Ergodic Theory and Dynamical Systems 7 (1987) 481–488.

[35] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[36] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks, Proceedings of the IEEE 95 (2007) 163–187.

[37] M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses, Automatica 43 (2007) 598–607.

[38] Y.S. Chow, H. Teicher, Probability Theory: Independence, Interchangeability, Martingales, 3rd edition, Springer-Verlag, New York, USA, 1997.

CHAPTER 6

Distributed State Estimation in an LSS

6.1 Introduction

In various engineering applications, it is usually appreciative from a practical point of view to estimate the states of a subsystem using only information of the subsystems that are directly connected to it. Examples include systems that are constituted from a large number of subsystems, a system that consists of several subsystems that are far away from each other geometrically, and so on. Major reasons behind this approach are that lumped estimations are usually computationally infeasible, and/or economically expensive, and/or not very suitable from the viewpoint of maintenance, and so on.

Several methods have been developed to achieve this objective. For example, in [1], the well-known Kalman filter is extended to a multidimensional Roesser model utilizing the concept of wave advance process. In [2], the Jacobi over-relaxation method is combined with dynamic average consensus algorithms under the framework of Baysian estimations. In [3], a method is suggested for iterative estimating states and internal outputs of subsystems of a one-dimensional spatially distributed dynamic systems.

Motivated by the design procedure of the Luenberger observer and the Kalman filter, a method is suggested in [4] to design a one-step ahead state predictor that has the same structure as that of the plant, using the model of Section 3.3 for a networked dynamic system. Afterward, these results have been extended to distributed state filter designs [5]. In this chapter, we summarize major results of these investigations and discusses their extensions to robust distributed state estimations. Asymptotic properties of the distributed predictor have also been investigated, and conditions are established for the equivalence in steady estimation accuracy between the distributed predictor and the lumped Kalman filter.

More precisely, consider a linear time varying dynamic system $\mathbf{\Sigma}$ consisting of N subsystems. Assume that the dynamics of its ith subsystem, denote it by $\mathbf{\Sigma}_i$, is described by the following discrete state-space model,

$$\begin{bmatrix} x(k+1,i) \\ z(k,i) \\ y(k,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k,i) & A_{\mathbf{xv}}(k,i) & B_{\mathbf{x}}(k,i) & 0 \\ A_{\mathbf{zx}}(k,i) & A_{\mathbf{zv}}(k,i) & 0 & 0 \\ C_{\mathbf{x}}(k,i) & C_{\mathbf{v}}(k,i) & 0 & D(k,i) \end{bmatrix} \begin{bmatrix} x(k,i) \\ v(k,i) \\ d(k,i) \\ w(k,i) \end{bmatrix} \tag{6.1}$$

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00006-3

Similar to those in Eq. (3.25), $k = 0, 1, \cdots$, and $i = 1, 2, \cdots, N$, stand here again respectively for the temporal variable and the index number of a discrete linear time varying subsystem, $x(k, i)$ represents the state vector of the ith subsystem $\mathbf{\Sigma}_i$ at time k, $z(k, i)$ and $v(k, i)$ respectively its internal output vector and internal input vector, while $y(k, i)$ its external output vector. In addition, $d(k, i)$ and $w(k, i)$ are adopted in the above model to represent respectively a process disturbance vector and a measurement error vector. Through this chapter, the random processes associated respectively with these random disturbances are assumed to be white, and their mathematical expectations and covariance matrices are assumed to be respectively equal to zero and an identity matrix. Moreover, it is assumed that these two random processes are independent of each other, and random processes associated with different subsystems are also independent.

Let $z(k)$ and $v(k)$ represent the vectors constituted respectively from all the subsystem internal output vectors and all the subsystem internal input vectors. That is, $z(k) = \mathbf{col}\left\{z(k, i)|_{i=1}^{N}\right\}$ and $v(k) = \mathbf{col}\left\{v(k, i)|_{i=1}^{N}\right\}$. Interactions among plant subsystems are described by

$$v(k) = \Phi(k)z(k) \tag{6.2}$$

which reflects the engineering fact that each internal input signal of a subsystem is actually an internal output signal of some other subsystems. This is completely the same as that of Eq. (3.26).

In the above system model, direct influences are not existent from the process disturbance vector $d(k, i)$ to the external subsystem output vector $y(k, i)$. Moreover, both the process disturbance vector $d(k, i)$ and the measurement error vector $w(k, i)$ do not directly affect the internal subsystem output vector $z(k, i)$. Furthermore, the measurement error vector $w(k, i)$ does not directly influence the subsystem state vector $x(k, i)$. These assumptions make the associated system model different a little from that described by Eqs. (3.25) and (3.25), which are adopted mainly for simplifying mathematical expressions in the derivations and presentations of the distributed state estimators.

6.2 Predictor Design With Local Measurements

Similarly to the Luenberger observer, to predict the state of the dynamical system $\mathbf{\Sigma}$ described by Eqs. (6.1) and (6.2), an estimator is constructed that has the same structure as that of the system $\mathbf{\Sigma}$ itself, and differences between the outputs of the estimator and those of the system $\mathbf{\Sigma}$ are used to adjust the states of the estimator. More precisely, the estimator also consists of N subsystems, and the dynamics of its ith subsystem, denoted by $\hat{\mathbf{\Sigma}}_i$, is described by the following state space model:

$$\begin{bmatrix} \hat{x}(k+1,i) \\ \hat{z}(k,i) \\ \hat{y}(k,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k,i) & A_{\mathbf{xv}}(k,i) & K_{\mathbf{x}}(k,i) \\ A_{\mathbf{zx}}(k,i) & A_{\mathbf{zv}}(k,i) & 0 \\ C_{\mathbf{x}}(k,i) & C_{\mathbf{v}}(k,i) & 0 \end{bmatrix} \begin{bmatrix} \hat{x}(k,i) \\ \hat{v}(k,i) \\ y(k,i)-\hat{y}(k,i) \end{bmatrix}. \tag{6.3}$$

In addition, signal transmissions among these subsystems, that is, $\hat{\mathbf{\Sigma}}_i|_{i=1}^N$, are completely the same as those of the dynamical system $\mathbf{\Sigma}$. More specifically,

$$\mathbf{col}\left\{\hat{v}(k,i)|_{i=1}^N\right\} = \Phi(k)\mathbf{col}\left\{\hat{z}(k,i)|_{i=1}^N\right\}. \tag{6.4}$$

To simplify notations, we further abbreviate the vectors $\mathbf{col}\left\{\hat{z}(k,i)|_{i=1}^N\right\}$ and $\mathbf{col}\left\{\hat{v}(k,i)|_{i=1}^N\right\}$ as $\hat{z}(k)$ and $\hat{v}(k)$.

Differently from the Luenberger observer described in Chapter 3, each subsystem in the above estimator only receives information about the external output vector of its associated subsystem in the dynamical system $\mathbf{\Sigma}$. This associated subsystem has a state space model similar to that of itself. This makes the estimator realizable in a distributed way. On the other hand, as the subsystems of the estimator are connected through Eq. (6.4), information exchange exists among the estimator subsystems. Existence of these connections distinguish the estimator from those that *independently* predict the state vector of each subsystem using its local output measurements. Fig. 6.1 gives a schematic diagram of this one-step state predictor (OSSP) for a networked system and of the networked system itself.

In estimations, a basic requirement is that an estimate must be unbiased, that is, the mathematical expectation of estimation errors must be equal to zero. In addition, it is usually preferable that the estimation error of an estimator has a minimal covariance matrix. In this section, we discuss how to find the optimal gain matrix $K_{\mathbf{x}}(k,i)$, $k=0,1,2,\cdots$, $i=1,2,\cdots,N$, such that for each subsystem of the dynamical system $\mathbf{\Sigma}$, the estimate of the state predictor is unbiased and the covariance matrix of its prediction errors is minimized. In other words, we need to search the gain matrix $K_{\mathbf{x}}(k,i)$ for each subsystem $\hat{\mathbf{\Sigma}}_i$ such that, at each time instant $i=0,1,2,\cdots$, the following two requirements are satisfied for every $i=1,2,\cdots,N$:

$$\begin{aligned} &\mathbf{E}(x(k,i)-\hat{x}(k,i))=0, \\ &\mathbf{E}\{(x(k,i)-\hat{x}(k,i)(x(k,i)-\hat{x}(k,i))^T\} \text{ is minimized.} \end{aligned}$$

6.2.1 Derivation of the Optimal Gain Matrix

As in Chapter 3, we define the following matrices to simplify mathematical expressions: $K_{\mathbf{x}}(k)=\mathbf{diag}\left\{K_{\mathbf{x}}(k,i)|_{i=1}^N\right\}$, $A_{*\#}(k)=\mathbf{diag}\left\{A_{*\#}(k,i)|_{i=1}^N\right\}$, $B_{\mathbf{x}}(k)=\mathbf{diag}\left\{B_{\mathbf{x}}(k,i)|_{i=1}^N\right\}$,

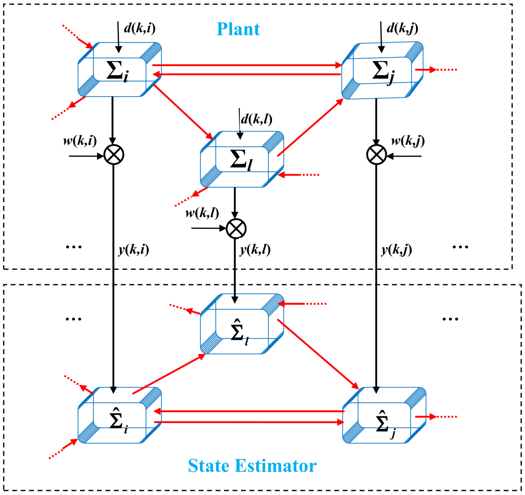

Figure 6.1: A distributed one-step ahead state predictor.

$C_{\mathbf{x}}(k) = \mathbf{diag}\left\{C_{\mathbf{x}}(k,i)|_{i=1}^{N}\right\}$, $C_{\mathbf{v}}(k) = \mathbf{diag}\left\{C_{\mathbf{v}}(k,i)|_{i=1}^{N}\right\}$, and $D(k) = \mathbf{diag}\left\{D(k,i)|_{i=1}^{N}\right\}$, in which $*, \# = \mathbf{x}$ or $\mathbf{z}$ or $\mathbf{v}$. In addition, the vectors $\mathbf{col}\left\{d(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{w(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{y(k,i)|_{i=1}^{N}\right\}$, and $\mathbf{col}\left\{\hat{y}(k,i)|_{i=1}^{N}\right\}$ are respectively denoted by $d(k)$, $w(k)$, $y(k)$, and $\hat{y}(k)$. Using these symbols, we can show through straightforward algebraic manipulations that the input–output relations of the dynamic system $\mathbf{\Sigma}$ of Eq. (6.1) and its DOSSP of Eq. (6.3) can be equivalently described by the following lumped state space models:

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \left\{ \begin{bmatrix} A_{\mathbf{xx}}(k) & B_{\mathbf{x}}(k) & 0 \\ C_{\mathbf{x}}(k) & 0 & D(k) \end{bmatrix} + \begin{bmatrix} A_{\mathbf{xv}}(k) \\ C_{\mathbf{v}}(k) \end{bmatrix} \Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1}\left[A_{\mathbf{zx}}(k) \ \ 0 \ \ 0\right] \right\} \begin{bmatrix} x(k) \\ d(k) \\ w(k) \end{bmatrix}, \tag{6.5}$$

$$\begin{bmatrix} \hat{x}(k+1) \\ \hat{y}(k) \end{bmatrix} = \left\{ \begin{bmatrix} A_{\mathbf{xx}}(k) & K_{\mathbf{x}}(k) \\ C_{\mathbf{x}}(k) & 0 \end{bmatrix} + \begin{bmatrix} A_{\mathbf{xv}}(t) \\ C_{\mathbf{v}}(t) \end{bmatrix} \Phi(k)\left[I - A_{\mathbf{zv}}(t)\Phi(k)\ \right]^{-1}\left[A_{\mathbf{zx}}(t) \ \ 0\right] \right\} \begin{bmatrix} \hat{x}(k) \\ y(k) - \hat{y}(k) \end{bmatrix}. \tag{6.6}$$

On the basis of these relations, some further matrix operations show that

$$\tilde{x}(k+1) = [A(k) - K_{\mathbf{x}}(k)C(k)]\tilde{x}(k) + K_{\mathbf{x}}(k)D(k)w(k) - B_{\mathbf{x}}(k)d(k), \tag{6.7}$$

where $\tilde{x}(k) = \mathbf{col}\left\{\tilde{x}(k,i)|_{i=1}^{N}\right\}$ with $\tilde{x}(k,i) = \hat{x}(k,i) - x(k,i)$, that is, $\tilde{x}(k)$ is the predictor estimation error vector at the time instant k; moreover, $A(k) = A_{\mathbf{xx}}(k) + A_{\mathbf{xv}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} A_{\mathbf{zx}}(k)$ and $C(k) = C_{\mathbf{x}}(k) + C_{\mathbf{v}}(k)\Phi(k)\left[I - A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1} A_{\mathbf{zx}}(k)$.

Using this recursive expression for estimation errors, straightforward algebraic manipulations show that when the random processes $d(k,i)$ and $w(k,i)$ are white and independent of each other, as assumed in the problem description, we have

$$\mathbf{E}\{\tilde{x}(k+1)\} = [A(k) - K_{\mathbf{x}}(k)C(k)]\mathbf{E}\{\tilde{x}(k)\}, \tag{6.8}$$

$$\begin{aligned}\mathbf{E}\{\tilde{x}(k+1)\tilde{x}^T(k+1)\} &= [A(k) - K_{\mathbf{x}}(k)C(k)]\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}[A(k) - K_{\mathbf{x}}(k)C(k)]^T \\ &\quad + K_{\mathbf{x}}(k)D(k)D^T(k)K_{\mathbf{x}}^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k).\end{aligned} \tag{6.9}$$

Eq. (6.8) makes it clear that if at the time instant k, each element of the matrix $A(k) - K_{\mathbf{x}}(k)C(k)$ is of a finite magnitude and the DOSSP is unbiased, then it is also unbiased at the next time instant $k+1$. Note that when the matrix $I - A_{\mathbf{zv}}(k)\Phi(k)$ is invertible and each element of the gain matrix $K_{\mathbf{x}}(k)$ is finite in magnitude, it is certain that every element of the matrix $A(k) - K_{\mathbf{x}}(k)C(k)$ has only a finite magnitude. In addition, the invertibility of the matrix $I - A_{\mathbf{zv}}(k)\Phi(k)$ is guaranteed by the well-posedness of the dynamic system $\mathbf{\Sigma}$ [4,6–9], whereas the latter is an essential requirement in constructing a system that can work properly. Moreover, the finiteness of the gain matrix $K_{\mathbf{x}}(k)$ is essential in actual realizations of a designed predictor. These arguments mean that these two requirements must usually be satisfied in actual engineering problems. Hence, for a well-designed predictor, if the estimate is unbiased at its starting time instant, say $k=0$, then its unbiasedness is kept afterward.

On the other hand, note that both matrices $\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}$ and $D(k)D^T(k)$ are at least positive semidefinite. From these properties and Eq. (6.9) we obtain an important characteristic of the correlation matrix $\mathbf{E}\{\tilde{x}(k+1)\tilde{x}^T(k+1)\}$, which is helpful in establishing the optimality of the derived update gain matrix. To clarify dependence of the correlation matrix of prediction errors on the gain matrix of the predictor, in the following lemma, we denote by $P(k+1, K_{\mathbf{x}}(k))$ the correlation matrix $\mathbf{E}\{\tilde{x}(k+1)\tilde{x}^T(k+1)\}$ at the $(k+1)$th time constant corresponding to the gain matrix $K_{\mathbf{x}}(k)$.

Lemma 6.1. *Assume that λ is an arbitrary number belonging to the interval* $[0, 1]$. *Then, the following inequality is valid for each pair of matrices $K_{\mathbf{x}1}(k)$ and $K_{\mathbf{x}2}(k)$ with consistent dimensions:*

$$P\{k+1, \lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)\} \le \lambda P(k+1, K_{\mathbf{x}1}(k)) + (1-\lambda)P(k+1, K_{\mathbf{x}2}(k)). \tag{6.10}$$

Moreover, if the matrix $D(k)D^T(k)$ is regular, then the "$\le$" symbol in this equation can be replaced by the "$=$" symbol if and only if λ belongs to the set $\{0,\ 1\}$ *or $K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)$ is not of FRR.*

Proof. To shorten mathematical expressions and present essential ideas behind the proof more clearly, symbols $\Theta(k)$, $\Psi(k)$, and $\Xi(k)$ are adopted to denote respectively the matrices $D(k)D^T(k) + C(k)P(k)C^T(k)$, $A(k)P(k)C^T(k)$, and $A(k)P(k)A^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k)$. From Eq. (6.9) we have that

$$\begin{aligned}
&P(k+1, K_{\mathbf{x}}(k)) \\
&= [A(k) - K_{\mathbf{x}}(k)C(k)]P(k)[A(k) - K_{\mathbf{x}}(k)C(k)]^T + K_{\mathbf{x}}(k)D(k)D^T(k)K_{\mathbf{x}}^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k) \\
&= K_{\mathbf{x}}(k)\left\{D(k)D^T(k) + C(k)P(k)C^T(k)\right\}K_{\mathbf{x}}^T(k) - A(k)P(k)C^T(k)K_{\mathbf{x}}^T(k) \\
&\quad - K_{\mathbf{x}}(k)C(k)P(k)A^T(k) + A(k)P(k)A^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k) \\
&= K_{\mathbf{x}}(k)\Theta(k)K_{\mathbf{x}}^T(k) - \Psi(k)K_{\mathbf{x}}^T(k) - K_{\mathbf{x}}(k)\Psi^T(k) + \Xi(k).
\end{aligned} \tag{6.11}$$

Note that the matrix $P(k)$ is at least positive semidefinite, which is guaranteed by its definition. We can therefore declare that the matrix $D(k)D^T(k) + C(k)P(k)C^T(k)$ also is at least positive semidefinite. Hence, for arbitrary $\lambda \in [0,\ 1]$,

$$\begin{aligned}
&P(k+1, \lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)) \\
&= [\lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)]\,\Theta(k)\,[\lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)]^T \\
&\quad - \Psi(k)\,[\lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)]^T - [\lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)]\,\Psi^T(k) + \Xi(k) \\
&= \lambda^2 K_{\mathbf{x}1}(k)\Theta(k)K_{\mathbf{x}1}^T(k) + \lambda(1-\lambda)K_{\mathbf{x}1}(k)\Theta(k)K_{\mathbf{x}2}^T(k) + \lambda(1-\lambda)K_{\mathbf{x}2}(k)\Theta(k)K_{\mathbf{x}1}^T(k) \\
&\quad + (1-\lambda)^2 K_{\mathbf{x}2}(k)\Theta(k)K_{\mathbf{x}2}^T(k) + \lambda\left\{-\Psi(k)K_{\mathbf{x}1}^T(k) - K_{\mathbf{x}1}(k)\Psi^T(k) + \Xi(k)\right\} + (1-\lambda) \\
&\quad \times \left\{-\Psi(k)K_{\mathbf{x}2}^T(k) - K_{\mathbf{x}2}(k)\Psi^T(k) + \Xi(k)\right\} \\
&= \lambda\left\{K_{\mathbf{x}1}(k)\Theta(k)K_{\mathbf{x}1}^T(k) - \Psi(k)K_{\mathbf{x}1}^T(k) - K_{\mathbf{x}1}(k)\Psi^T(k) + \Xi(k)\right\} \\
&\quad + (1-\lambda)\left\{K_{\mathbf{x}2}(k)\Theta(k)K_{\mathbf{x}2}^T(k) - \Psi(k)K_{\mathbf{x}2}^T(k) - K_{\mathbf{x}2}(k)\Psi^T(k) + \Xi(k)\right\} \\
&\quad - \lambda(1-\lambda)\,\{K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)\}\,\Theta(k)\,\{K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)\}^T \\
&\le \lambda P(k+1, K_{\mathbf{x}1}(k)) + (1-\lambda)P(k+1, K_{\mathbf{x}2}(k)).
\end{aligned} \tag{6.12}$$

When the matrix $D(k)D^T(k)$ is positive definite, the matrix $\Theta(k)$ is certainly positive definite, as $\Theta(k) = D(k)D^T(k) + C(k)P(k)C^T(k)$ by its definition. Therefore, when $\lambda \notin \{0\}\bigcup\{1\}$ and $K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)$ is of full row rank, it is certain that

$$\lambda(1-\lambda)\,\{K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)\}\,\Theta(k)\,\{K_{\mathbf{x}1}(k) - K_{\mathbf{x}2}(k)\}^T > 0. \tag{6.13}$$

The proof can now be completed by combining Eqs. (6.12) and (6.13). □

Denote the set consisting of all permissible gain matrices $K_{\mathbf{x}}(k)$ in the above predictor design by $\mathcal{K}_{\mathbf{x}}$. Clearly, if $K_{\mathbf{x}1}(k)$, $K_{\mathbf{x}2}(k) \in \mathcal{K}_{\mathbf{x}}$, then $\lambda K_{\mathbf{x}1}(k) + (1-\lambda)K_{\mathbf{x}2}(k)$ also belongs to $\mathcal{K}_{\mathbf{x}}$ for an arbitrary real number λ. We can therefore declare that the set $\mathcal{K}_{\mathbf{x}}$ is convex. From Lemma 6.1 and properties of convex functions [10] we have that if for each $k = 1, 2, \cdots, N$, the gain matrix $K_{\mathbf{x}}(k)$ minimizes the correlation matrix $\mathbf{E}\{\tilde{x}(t+1)\tilde{x}^T(t+1)\}$, then the gain matrix $\mathbf{diag}\{K_{\mathbf{x}}(k)|_{k=1}^{N}\}$ is certainly a global minimizer.

Various efficient algorithms have been developed for convex optimizations, such as the cutting plane method, the interior point method, and so on [11]. It is worth pointing out, however, that these algorithms are not very suitable for obtaining the optimal gain matrix $K_{\mathbf{x}}(k)$. Essentially, it is because that the available convex optimization methods cannot be very easily implemented online. Moreover, for a large-scale networked system, which usually includes a great amount of subsystems and whose state vector $x(k)$ usually has a high dimension, even off-line optimizations are often computationally prohibitive.

Denote $\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}$ and $\mathbf{E}\{\tilde{x}(k,i)\tilde{x}^T(k,j)\}$ respectively by $P(k)$ and $P_{ij}(k)$, $i, j = 1, 2, \cdots, N$. Assume that the dimensions of $x(k,i)$ and $v(k,i)$ are respectively $m_{\mathbf{x}i}$ and $m_{\mathbf{v}i}$. Define the integers $M_{\mathbf{x}i}$, $M_{\mathbf{v}i}$, $M_{\mathbf{x}}$, and $M_{\mathbf{v}}$ respectively as $M_{\mathbf{x}} = \sum_{k=1}^{N} m_{\mathbf{x}k}$, $M_{\mathbf{v}} = \sum_{k=1}^{N} m_{\mathbf{S}k}$, and $M_{\mathbf{x}i} = M_{\mathbf{S}i} = 0$ when $i = 1$ and $M_{\mathbf{x}i} = \sum_{k=1}^{i-1} m_{\mathbf{x}k}$ and $M_{\mathbf{v}i} = \sum_{k=1}^{i-1} m_{\mathbf{v}k}$ when $2 \le i \le N$. Moreover, define the matrices

$$J_{\mathbf{x}i} = \mathbf{col}\left\{0_{M_{\mathbf{x}i}\times m_{\mathbf{x}i}}, I_{m_{\mathbf{x}i}}, 0_{(M_{\mathbf{x}}-M_{\mathbf{x},i+1})\times m_{\mathbf{x}i}}\right\} \quad \text{and}$$
$$J_{\mathbf{v}i} = \mathbf{col}\left\{0_{M_{\mathbf{v}i}\times m_{\mathbf{v}i}}, I_{m_{\mathbf{v}i}}, 0_{(M_{\mathbf{v}}-M_{\mathbf{v},i+1})\times m_{\mathbf{S}i}}\right\}.$$

Then, from the definitions of the matrices $P(k)$ and $P_{ij}(k)$ we can straightforwardly prove that

$$P_{ij}(k) = J_{\mathbf{x}i}^T P(k) J_{\mathbf{x}j}, \quad \forall\, i, j = 1, 2, \cdots, N. \tag{6.14}$$

From these relations we obtain the optimal gain matrix $K_{\mathbf{x}}(k,i)$ that minimizes $P_{ii}(t+1)$. Its mathematical derivations are deferred to the appendix of this chapter.

Theorem 6.1. *Denote the matrices* $[A_{\mathbf{xx}}(k,i) \;\; A_{\mathbf{xv}}(k,i)]$ *and* $[C_{\mathbf{x}}(k,i) \;\; C_{\mathbf{v}}(k,i)]$ *respectively by* $A_{\mathbf{x}}(k,i)$ *and* $C(k,i)$. *Define the matrix*

$$W(k,i) = \begin{bmatrix} J_{\mathbf{x}i}^T \\ J_{\mathbf{v}i}^T \Phi(k)\left[\, I - A_{\mathbf{zv}}(k)\Phi(k)\,\right]^{-1} A_{\mathbf{vx}}(k) \end{bmatrix}. \tag{6.15}$$

Assume that the matrix $D(k,i)$ *is of full row rank. For every subsystem* $\mathbf{\Sigma}_i$ *and every time instant* k, *denote by* $K_{\mathbf{x}}^{\text{opt}}(k,i)$ *the optimal gain matrix* $K_{\mathbf{x}}(k,i)$ *that minimizes the covariance matrix* $P_{ii}(k+1)$. *Then this optimal gain matrix can be expressed as*

$$K_{\mathbf{x}}^{\text{opt}}(k,i) = A_{\mathbf{x}}(k,i)\left\{I + W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1}$$

$$\times W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}. \tag{6.16}$$

The above analysis shows that if $K_{\mathbf{x}}(k)$ minimizes $P(t+1)$, then it is necessary that its diagonal block matrix $K_{\mathbf{x}}(k,i)$ minimizes $P_{ii}(t+1)$, $i = 1, 2, \cdots, N$. On the other hand, Theorem 6.1 declares that when the matrix $D(k,i)$ is of full row rank, there is only one $K_{\mathbf{x}}(k,i)$ that minimizes $P_{ii}(t+1)$. In addition to these, the results of Lemma 6.1 make it clear that if $D(k)D^T(k)$ is invertible, which is equivalent to that for every $i \in \{1,\ 2,\ \cdots,\ N\}$, the matrix $D(k,i)$ is of full row rank, then there is only one $K_{\mathbf{x}}(k)$ that minimizes $P(k+1)$. These imply that $\mathbf{diag}\left\{K_{\mathbf{x}}^{\text{opt}}(k,i)|_{i=1}^{N}\right\}$ is usually the optimal gain matrix for the DOSSP.

Using the optimal gain matrix $K_{\mathbf{x}}^{\text{opt}}(k,i)$, $i = 1, 2, \cdots, N$, we derive an explicit expression for the covariance matrix $\mathbf{E}\{\tilde{x}(t+1)\tilde{x}^T(t+1)\}$, which is given by the following theorem.

Theorem 6.2. *Concerning the distributed one-step ahead state predictor, assume that the gain matrix for its ith subsystem $\hat{\mathbf{\Sigma}}_i$ is given by Eq. (6.16), $i = 1, 2, \cdots, N$. Partition the covariance matrix $P(k+1)$ of its prediction errors consistently with the dimensions of the state vectors of its subsystems. Then, for every $i, j = 1, 2, \cdots, N$, the ith row jth column block matrix of the covariance matrix $P(k+1)$ can be recursively expressed as follows:*

$$
\begin{aligned}
&P_{ij}(k+1)\\
&= \begin{cases}
A_{\mathbf{x}}(k,i)\left\{I + W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1} & \\
\quad \times W(k,i)P(k)W^T(k,i)A_{\mathbf{x}}^T(k,i) + B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i), & i = j, \\
A_{\mathbf{x}}(k,i)\left\{I + W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1} & \\
\quad \times W(k,i)P(k)W^T(k,j)\left\{I + C^T(k,j)\left[D(k,j)D^T(k,j)\right]^{-1}\right. & \\
\quad\quad \left. \times C(k,j)W(t,j)P(k)W^T(t,j)\right\}^{-1} A_{\mathbf{x}}^T(k,j), & i \neq j.
\end{cases}
\end{aligned}
\tag{6.17}
$$

The proof of Theorem 6.2 is given in the appendix attached to the end of this chapter.

When there exists $i \in \{1, 2, \cdots, N\}$ such that the matrix $D(k,i)D^T(k,i)$ is only positive semidefinite, that is, the matrix $D(k,i)$ is row rank deficient, similar results can also be obtained [4].

Theorems 6.1 and 6.2 make it clear that for the suggested DOSSP, both its optimal gain matrix and its covariance matrix of prediction errors can be computed in a recursive way. This is very similar to that of the widely known Kalman filter. However, from viewpoint of both realizations and computations, there are significant differences between the state predictor discussed here and the lumped Kalman filter. More specifically, in obtaining an estimate about the state vector of the plant $\boldsymbol{\Sigma}$, the DOSSP utilizes only local plant output measurements obtained from an associated subsystem of the plant. This is helpful in reducing costs of communications and so on and in improving realizability of the estimator when the plant has a large number of subsystems. On the other hand, in actual computations of the estimates of the DOSSP, inversions are required only for matrices with dimensions $(m_{\mathbf{x}i} + m_{\mathbf{v}i}) \times (m_{\mathbf{x}i} + m_{\mathbf{v}i})$, $i = 1, 2, \cdots, N$, which is independent of the subsystem number N. This implies that this state prediction method is scalable to plants with great amount of subsystems, which is significantly different from that of the lumped Kalman filter. As a matter of fact, in the lumped Kalman filter, the inversion of a $(\sum_{i=1}^{N} m_{\mathbf{x}i} \times \sum_{i=1}^{N} m_{\mathbf{x}i})$-dimensional matrix should usually be calculated. Note that $m_{\mathbf{x}i}$ increases linearly with the increment of the subsystem number N. Moreover, matrix inversions are generally computationally expensive. Furthermore, computations of the inverse of a large-dimensional matrix has a very high probability to meet numerical stability problems [12,13]. It is clear that for a networked system with a large number of subsystems, the above state prediction procedure is more computationally attractive from the viewpoint of both computational costs and numerical stability.

To clarify these points, we give a brief comparison of the computation complexities of these two state prediction methods. When the lumped Kalman filter is used in predicting the state vector of system $\boldsymbol{\Sigma}$, the corresponding computational complexity in each iteration of state prediction is of order $\left(\sum_{i=1}^{N} m_{\mathbf{x}i}\right)^3$. On the other hand, when the DOSSP given by Theorems 6.1 and 6.2 is utilized, it is of order $(N + 1) \sum_{i=1}^{N} (m_{\mathbf{x}i} + m_{\mathbf{v}i})^3$. Note that $m_{\mathbf{v}i}$ is in general appreciably smaller than $m_{\mathbf{x}i}$ in an engineering system [6,14–17]. These imply that for a plant with a large number of subsystems, that is, a plant with large N, the computational complexity of the DOSSP is usually far less than that of the lumped Kalman filter. To make this point clearer, we consider here a simple situation in which the dimension of the state vector of each plant subsystem is equivalent. Denote this dimension by $m_{\mathbf{x}}$. Moreover, assume that $m_{\mathbf{v}i} \leq m_{\mathbf{x}}$. Then straightforward algebraic operations show that the ratio between $(N + 1) \sum_{i=1}^{N} (m_{\mathbf{x}i} + m_{\mathbf{v}i})^3$ and $\left(\sum_{i=1}^{N} m_{\mathbf{x}i}\right)^3$ is not greater than $\frac{8}{N} + \frac{8}{N^2}$. This means that when the number of plant subsystems is regarded to be a variable, the computational complexity of the DOSSP increases one order slower than that of the lumped Kalman filter.

On the other hand, when the Kalman filter is independently applied to each individual subsystem without considering subsystem interactions, the computational complexity in every state prediction iteration is of order $\sum_{i=1}^{N} m_{\mathbf{x}i}^3$, which increases *only linearly* with the subsystem number N. Computationally, this estimator is more attractive. However, its complete ignorance of subsystem interactions may often cause divergent estimates, which is illustrated by a numerical example in [4].

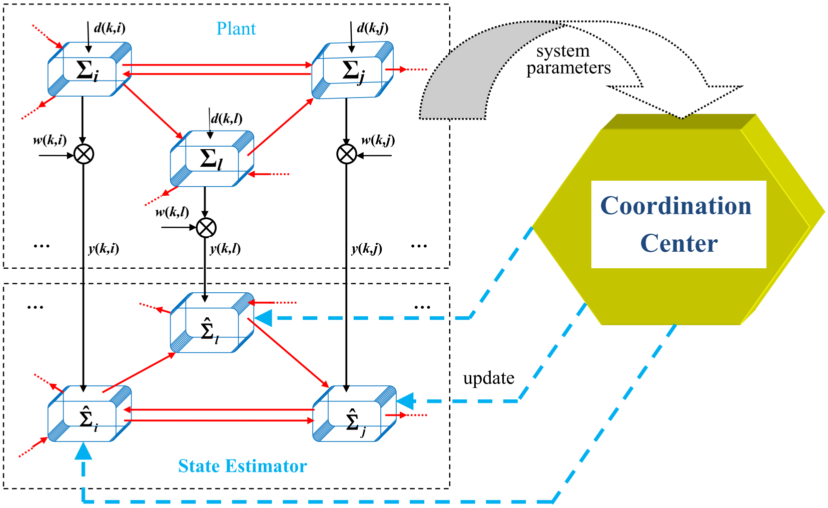

Figure 6.2: Implementation of the optimal distributed one-step ahead state predictor.

Note that the Kalman filter usually has a full gain matrix, which implies that the measured outputs of all plant subsystems are used in estimating the states of one subsystem. In contrast to this, in the aforementioned estimation procedure, the gain matrix has a block diagonal structure, which makes the associated optimal predictor more suitable to be implemented in a distributed way than the Kalman filter. However, it is worth emphasizing that it is still not able to realize the derived DOSSP in a completely distributed way. The reasons are that in computing the gain matrix $K_{\mathbf{x}}^{\text{opt}}(k,i)$ for the subsystem $\hat{\mathbf{\Sigma}}_i$, information about the parameters of the whole system and the covariance matrix $P(k)$ is required, that is, some coordinations are still necessary among subsystems in this DOSSP. Fig. 6.2 provides a schematic illustration for the implementation of the derived DOSSP for a networked system. This implementation procedure can be briefly described as follows.

At every prediction iteration, each plant subsystem transmits its system parameters to a unit, which is called the coordination center, that has stored the covariance matrix of prediction errors at the previous time instant. In this collaboration unit, the optimal gain matrices $K_{\mathbf{x}}^{\text{opt}}(k,i)|_{i=1}^{N}$ are computed, and the covariance matrix $P(k)$ is updated according respectively to Eqs. (6.16) and (6.17). After this update, the computed optimal gain matrices are separately delivered to each subsystem $\hat{\mathbf{\Sigma}}_i$ of the state predictor described by Eqs. (6.3) and (6.4), and the latter renews its state vector $\hat{x}(k,i)$ according to Eq. (6.3) on the basis of the local plant output measurement $y(k,i)$ and $\hat{v}(k,i)$ received from all subsystems that have direct influences on its state vector and output vector.

The existence of the coordination center makes the aforementioned predictor not completely distributed. To emphasize this character of the state predictor, it is called a coordinated distributed one step ahead state predictor (CDOSSP). The remaining theoretical challenging issue is to investigate possibilities or conditions for releasing these coordinations and making the estimator completely distributed. For plants having time invariant dynamics, if the covariance matrix of state prediction errors converges with the increment of the temporal variable k, then from Theorem 6.2 it is clear that the optimal gain matrix of each subsystem in the CDOSSP, that is, $K_{\mathbf{x}}^{\mathrm{opt}}(k,i)$, also converges. Under such a situation, it is possible to replace $K_{\mathbf{x}}^{\mathrm{opt}}(k,i)$ with $\lim_{t\to\infty} K_{\mathbf{x}}^{\mathrm{opt}}(k,i)$ for every subsystem $\hat{\mathbf{\Sigma}}_i$ in the state predictor. This replacement makes the collaboration unit no longer necessary and the state predictor completely distributed, at the cost of sacrificing the optimality of the state predictor in its transient period. This approach has also been widely adopted in lumped state estimations to reduce their computational costs, which leads to a suboptimal estimator [18–20]. However, further efforts are required to establish convergence conditions for the CDOSSP and explicit expressions for the associated steady-state covariance matrix and the optimal gain matrix.

Now, we consider storage requirements in the realization of the aforementioned CDOSSP. Note that to predict plant states using the lumped Kalman filter, in addition to the current system parameters and the previous covariance matrix of prediction errors, the predicted state vector of the whole plant at the previous time instant is also required. These mean that when storage requirement has to be taken into account, the numbers of data to be stored in each iteration by the lumped Kalman filter and the CDOSSP are respectively $\left(\sum_{i=1}^{N} m_{\mathbf{x}i}+1\right)\sum_{i=1}^{N} m_{\mathbf{x}i}$ and $\sum_{i=1}^{N} m_{\mathbf{x}i}\times\sum_{i=1}^{N} m_{\mathbf{x}i}$. Obviously, although less storage units are required by the CDOSSP, the difference between these two state predictors is not very significant, especially when $\sum_{i=1}^{N} m_{\mathbf{x}i}$ is large.

In addition to these, in the implementation of the lumped Kalman filter, besides the system parameters, all measured plant outputs should also be brought together. Let $m_{\mathbf{y}i}$, $m_{\mathbf{z}i}$, $m_{\mathbf{d}i}$, and $m_{\mathbf{w}i}$ denote respectively the dimensions of the vectors $y(k,i)$, $z(k,i)$, $d(k,i)$, and $w(k,i)$. Then, if each subsystem asks for an estimate of the plant state vector, the number of data that should be transferred in each state prediction between the plant and the Kalman filter, is clearly equal to $\sum_{i=1}^{N}(m_{\mathbf{x}i}+m_{\mathbf{z}i}+m_{\mathbf{y}i})(m_{\mathbf{x}i}+m_{\mathbf{v}i}+m_{\mathbf{d}i}+m_{\mathbf{w}i})+\sum_{i=1}^{N} m_{\mathbf{v}i}\sum_{i=1}^{N} m_{\mathbf{z}i}+\sum_{i=1}^{N} m_{\mathbf{y}i}+N\sum_{i=1}^{N} m_{\mathbf{x}i}$. On the other hand, when the aforementioned CDOSSP is used, data transmissions are required only for the system parameters of the plant, the gain matrix of each subsystem in the state predictor, and the internal plant outputs. The number of the elements belonging to these vectors or matrices amounts to $\sum_{i=1}^{N}(m_{\mathbf{x}i}+m_{\mathbf{z}i}+m_{\mathbf{y}i})(m_{\mathbf{x}i}+m_{\mathbf{v}i}+m_{\mathbf{d}i}+m_{\mathbf{w}i})+\sum_{i=1}^{N} m_{\mathbf{v}i}\sum_{i=1}^{N} m_{\mathbf{z}i}+\sum_{i=1}^{N} m_{\mathbf{x}i}m_{yi}+\sum_{i=1}^{N} m_{\mathbf{v}i}$. Obviously, when the plant has a large number of subsystems, communication is generally a demanding requirement for both the lumped Kalman filter and the CDOSSP.

On the other hand, note that in engineering problems, $m_{\mathbf{v}i}$ is usually not greater than $m_{\mathbf{y}i}$, and subsystems having direct interactions are rarely far away from each other geometrically [6,9,16]. Therefore, if N is much greater than $\max_i\{m_{\mathbf{y}i}\}$, then some communication cost reduction can be expected through adopting the CDOSSP. In addition, if plant parameters are known prior to estimations, then they can be stored in a state predictor, and their on line transmissions are no longer required. Under this situation, the numbers of data required to be transmitted for the lumped Kalman filter and the CDOSSP are respectively $\sum_{i=1}^{N} m_{\mathbf{y}i} + N\sum_{i=1}^{N} m_{\mathbf{x}i}$ and $\sum_{i=1}^{N} m_{\mathbf{x}i}m_{\mathbf{y}i} + \sum_{i=1}^{N} m_{\mathbf{v}i}$, and a significant communication cost reduction can be anticipated by the CDOSSP.

Compared with other existing distributed state estimation methods, such as those reported in [1–3], the plant model adopted here is more general and more appropriate in describing engineering systems. This means that the associated estimation procedure is applicable to a much wider class of actual systems. Moreover, the aforementioned analysis also reveals that this CDOSSP can be more easily scaled to a plant with a large amount of subsystems and arbitrary subsystem interactions.

6.2.2 Relations With the Kalman Filter

In the previous subsection, an optimal CDOSSP has been derived for a time-varying linear networked system, which can be recursively realized. Some of its important properties have also been discussed there. In this subsection, it will be made clear that the gain matrix in each subsystem of the optimal CDOSSP is equal to that of the well-known Kalman filter when only the output measurements of one plant subsystem are used to estimate the plant state vector. This relation may be helpful in reducing the implementation costs of the optimal CDOSSP and analyzing its convergence properties. In the next subsection, we discuss its applications to the robustification of the distributed state prediction procedure.

First, we introduce the following dynamic system $\bar{\mathbf{\Sigma}}_i$, and its one-step state predictions are performed with the Kalman filter:

$$x(k+1) = A(k)x(k) + B_{\mathbf{T}}(k)d(k), \qquad y(k,i) = \bar{C}_i(k)x(k) + D(k,i)w(k,i) \tag{6.18}$$

where $\bar{C}_i(k) = J_{\mathbf{y}i}^T C(k)$. Denote by $P_i^{[\text{kal}]}(k)$ the covariance matrix of the one-step prediction error of the Kalman filter at the $(k+1)$th sampling time. Moreover, assume that the value of this matrix has been calculated and its determinant is not equal to zero. Then, by means of the predictor update form of the Kalman filter [19,20], it can be straightforwardly shown that the Kalman filter of this dynamic system $\bar{\mathbf{\Sigma}}_i$ can be expressed as

$$\bar{\hat{x}}(k+1) = A(k)\bar{\hat{x}}(k) + K_i^{[\text{kal}]}(k)[y(k,i) - \bar{C}_i(k)\bar{\hat{x}}(k)], \tag{6.19}$$

where

$$K_i^{[\mathrm{kal}]}(k) = A(k)\left[(P_i^{[\mathrm{kal}]}(k))^{-1} + C_i^T(k)C_i(k)\right]^{-1} C_i^T(k)[D(k,i)D^T(k,i)]^{-1/2}. \qquad (6.20)$$

Moreover, the covariance matrix of prediction errors at the time instant $k+1$, that is, $P_i^{[\mathrm{kal}]}(k+1)$, has the recursive representation

$$P_i^{[\mathrm{kal}]}(k+1) = A(k)\left[(P_i^{[\mathrm{kal}]}(k))^{-1} + C_i^T(k)C_i(k)\right]^{-1} A^T(k) + B_{\mathbf{T}}(k)B_{\mathbf{T}}^T(k). \qquad (6.21)$$

To establish relations between the Kalman filter and the CDOSSP given in the aforementioned subsection, we derive the following expression for the optimal gain matrix of the CDOSSP and the associated covariance matrix of estimation errors.

Theorem 6.3. *Assume that the matrix $D(k)$ is of full row rank. Define the matrices $C_i(k) = J_{\mathbf{y}i}^T[D(k)D^T(k)]^{-1/2}C(k)$ and $A_i(k) = J_{\mathbf{x}i}^T A(k)$, $i = 1, 2, \cdots, N$. Then, for each subsystem $\mathbf{\Sigma}_i$, the gain matrix $K_{\mathbf{T}}^{\mathrm{opt}}(k,i)$ of the optimal CDOSSP can be rewritten as*

$$K_{\mathbf{T}}^{\mathrm{opt}}(k,i) = A_i(k)\left[P^{-1}(k) + C_i^T(k)C_i(k)\right]^{-1} C_i^T(k)[D(k,i)D^T(k,i)]^{-1/2}. \qquad (6.22)$$

In addition, for each $i, j = 1, 2, \cdots, N$, the covariance matrix $P_{ij}(k+1)$ has the following equivalent expression:

$$P_{ij}(k+1) = \begin{cases} A_i(k)\left[P^{-1}(k) + C_i^T(k)C_i(k)\right]^{-1} A_i^T(k) + B_{\mathbf{T}}(k,i)B_{\mathbf{T}}^T(k,i), & i = j, \\ A_i(k)\left[P^{-1}(k) + C_i^T(k)C_i(k)\right]^{-1} P^{-1}(k)\left[P^{-1}(k) + C_j^T(k)C_j(k)\right]^{-1} A_j^T(k), & i \neq j. \end{cases} \qquad (6.23)$$

A proof of the theorem is given in the appendix of this chapter.

Recall that $A_i(k) = J_{\mathbf{x}i}^T A(k)$. Straightforward comparisons of Eqs. (6.22) and (6.20) and of Eqs. (6.23) and (6.21) show that if $P_i^{[\mathrm{kal}]}(k) = P(k)$, then, for every $i = 1, 2, \cdots, N$,

$$K_{\mathbf{T}}^{\mathrm{opt}}(k,i) = J_{\mathbf{x}i}^T K_i^{[\mathrm{kal}]}(k), \quad P_{ii}(k+1) = J_{\mathbf{x}i}^T P_i^{[\mathrm{kal}]}(k+1) J_{\mathbf{x}i}. \qquad (6.24)$$

It is interesting to note that when the plant has only one subsystem, that is, the plant itself is lumped, the matrix $J_{\mathbf{x}i}$ vanishes to the identity matrix. The relations of Eq. (6.24) make it clear that for a lumped plant, the CDOSSP derived in the previous subsection reduces to the Kalman filter. This should not be an astonishment, noting that the objectives of the CDOSSP are to give an unbiased estimate for the state vector of each subsystem with minimal covariance matrix of estimation errors, which are consistent with those of the Kalman filter. On the other hand, when the number of plant subsystems is greater than one, the above relations reveal that, associated with the ith plant subsystem $\mathbf{\Sigma}_i$, the gain matrix $K_{\mathbf{T}}^{\mathrm{opt}}(k,i)$ of the CDOSSP is equivalent to the ith block of the gain matrix of the Kalman filter, in which the state vector of the whole system is estimated using *only* the measured output vector of the ith subsystem. Moreover, the covariance matrix for prediction errors with this subsystem by the CDOSSP is equivalent to the ith diagonal block of the covariance matrix of the associated Kalman filter. This is a quite surprising property, noting that although in both the Kalman filter and the CDOSSP, the unbiasedness and covariance matrices are adopted to measure prediction performances, the Kalman filter of the dynamic system $\bar{\mathbf{\Sigma}}_i$ uses global performance indices of the dynamic system $\mathbf{\Sigma}$ but its local measurement, whereas in deriving the CDOSSP, both the adopted prediction performances and the adopted measurements are local.

Based on Theorem 6.3, prediction accuracies can also be compared between the predictor given in the previous subsection and the lumped Kalman filter, which uses simultaneously all the plant output measurements.

Corollary 6.1. *Assume that the lumped Kalman filter uses simultaneously all the measured outputs of each plant subsystem. Moreover, assume that the covariance matrix of its prediction errors at the time instant k equals $\bar{P}^{[\mathrm{kal}]}(k)$. Denote its ith diagonal block matrix by $\bar{P}_{ii}^{[\mathrm{kal}]}(k)$. Furthermore, assume that $\bar{P}^{[\mathrm{kal}]}(k)-P(k)$ is negative semidefinite and $\bar{P}^{[\mathrm{kal}]}(k)$ is invertible. Then, for every $i=1,2,\ldots,N$, we certainly have that*

$$P_{ii}(k+1)\geq \bar{P}_{ii}^{[\mathrm{kal}]}(k+1). \tag{6.25}$$

Proof. When the lumped Kalman filter is used in the prediction of the plant state vector, direct applications of its predictor update form [19,20] show that under the condition that the covariance matrix $\bar{P}^{[\mathrm{kal}]}(k)$ is invertible, we have that

$$\bar{P}^{[\mathrm{kal}]}(k+1)=A(k)\left[(\bar{P}^{[\mathrm{kal}]}(k))^{-1}+C^T(k)\left[D(k)D^T(k)\right]^{-1}C(k)\right]^{-1}A^T(k)+B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k). \tag{6.26}$$

On the other hand, from the definitions of the matrices $J_{\mathbf{y}j}$ and C_j we can straightforwardly prove that, for each $i\in\{1,2,\cdots,N\}$,

$$\begin{aligned}&C^T(k)\left[D(k)D^T(k)\right]^{-1}C(k)\\ =\;&\sum_{j=1}^{N}\left\{J_{\mathbf{y}j}^T\left[D(k)D^T(k)\right]^{-1/2}C(k)\right\}^T\left\{J_{\mathbf{y}j}^T\left[D(k)D^T(k)\right]^{-1/2}C(k)\right\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} C_j^T(k) C_j(k) \\
&\geq C_i^T(k) C_i(k).
\end{aligned} \tag{6.27}
$$

Moreover, note that whenever the matrix $\bar{P}^{[\mathrm{kal}]}(k)$ is invertible, the seminegative definiteness of the matrix $\bar{P}^{[\mathrm{kal}]}(k) - P(k)$ is equivalent to $(\bar{P}^{[\mathrm{kal}]}(k))^{-1} \geq P^{-1}(k)$. We can therefore declare that when $\bar{P}^{[\mathrm{kal}]}(k) - P(k)$ is negative semidefinite and $\bar{P}^{[\mathrm{kal}]}(k)$ is regular, it is certain that

$$
(\bar{P}^{[\mathrm{kal}]}(k))^{-1} + C^T(k)\left[D(k)D^T(k)\right]^{-1} C(k) \geq P^{-1}(k) + C_i^T(k)C_i(k), \qquad i = 1, 2, \cdots, N. \tag{6.28}
$$

Note also that both matrices $C^T(k)\left[D(k)D^T(k)\right]^{-1} C(k)$ and $C_i^T(k)C_i(k)$ are at least positive semidefinite. The above inequalities are equivalent to that, for each $i \in \{1,\ 2,\ \cdots,\ N\}$,

$$
\left\{(\bar{P}^{[\mathrm{kal}]}(k))^{-1} + C^T(k)\left[D(k)D^T(k)\right]^{-1} C(k)\right\}^{-1} \leq \left\{P^{-1}(k) + C_i^T(k)C_i(k)\right\}^{-1}. \tag{6.29}
$$

Substitution of this inequality into Eq. (6.26) leads to

$$
\bar{P}^{[\mathrm{kal}]}(k+1) \leq A(k)\{P^{-1}(k) + C_i^T(k)C_i(k)\}^{-1}A^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k). \tag{6.30}
$$

We therefore have that

$$
\begin{aligned}
\bar{P}_{ii}^{[\mathrm{kal}]}(k+1) &= J_{\mathbf{y}i}^T \bar{P}^{[\mathrm{kal}]}(k+1) J_{\mathbf{y}i} \\
&\leq J_{\mathbf{y}i}^T \left\{A(k)\left[P^{-1}(k) + C_i^T(k)C_i(k)\right]^{-1} A^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k)\right\} J_{\mathbf{y}i} \\
&= A_i(k)\{P^{-1}(k) + C_i^T(k)C_i(k)\}^{-1}A_i^T(k) + B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i) \\
&= P_{ii}(k+1).
\end{aligned} \tag{6.31}
$$

This completes the proof. □

Recall that the matrix $\bar{P}^{[\mathrm{kal}]}(k+1)$ stands for the covariance matrix of the prediction errors of the lumped Kalman filter at the time instant $k+1$. Obviously, its ith diagonal block matrix represents the covariance matrix associated with the plant ith subsystem $\mathbf{\Sigma}_i$. Assume that at the time instant $k = 0$, the lumped Kalman filter and the CDOSSP start with completely the same estimate of the plant state vector. The above arguments imply that at every succeeding time instant and for each state of the plant, it is certain that the lumped Kalman filter has a higher prediction accuracy than the CDOSSP. As a matter of fact, by Eqs. (6.23), (6.26), and (6.27) and by the definitions of the matrices $A_i(k)|_{i=1}^{N}$ some straightforward algebraic manipulations show that, for an arbitrary $i \in \{1, 2, \cdots, N\}$,

$$\begin{aligned} & P_{ii}(k+1) - \bar{P}_{ii}^{[\text{kal}]}(k+1) \\ = & A_i(k) \left\{ \left[P^{-1}(k) + C_i^T(k) C_i(k) \right]^{-1} - \left[[\bar{P}^{[\text{kal}]}(k)]^{-1} + \sum_{j=1}^{N} C_j^T(k) C_j(k) \right]^{-1} \right\} A_i^T(k). \end{aligned}$$

Note that $\sum_{j=1}^{N} C_j^T(k) C_j(k) \geq C_i^T(k) C_i(k)$ for every $i = 1, 2, \cdots, N$. It is clear that when $P(k) \geq \bar{P}^{[\text{kal}]}(k)$, it is certain that

$$P^{-1}(k) + C_i^T(k) C_i(k) \leq [\bar{P}^{[\text{kal}]}(k)]^{-1} + \sum_{j=1}^{N} C_j^T(k) C_j(k).$$

Hence $P_{ii}(k+1) \geq \bar{P}_{ii}^{[\text{kal}]}(k+1)$. A repeated utilization of this relation leads to the aforementioned conclusions that the lumped Kalman filter always outperforms the CDOSSP in prediction accuracies, that is, compared to the lumped Kalman filter, the prediction accuracy of the CDOSSP is generally worse. Moreover, its convergence rate is usually slower. These conclusions can be considered reasonable noting that the predictor CDOSSP utilizes less information in estimating the states of every subsystem.

6.2.3 Robustification of the Distributed Predictor

A system model is usually an approximation of its dynamics, which means that there are often discrepancies between the output behaviors of the model and the actual outputs of the plant, even when they are stimulated by the same input signal. For plants working in various different environments, when their working mechanisms are not very clear, these discrepancies may be very large. Well-known examples include chemical processes, biochemical processes, and so on. These imply that if a state estimator is very sensitive to modeling errors, then it may not work in line with its original design purposes. Although the Kalman filter has been successfully applied to many engineering problems, such as target tracking, global positioning, hydrological modeling, atmospheric monitoring, economic data analysis, automated drug delivery, and so on, there also exist situations in which its performances are greatly sacrificed by modeling errors. To overcome these drawbacks, various approaches have been suggested, such as the H_∞ estimator, set-valued predictions, guaranteed cost designs, and so on [20,21]. In [22,23], a sensitivity penalization-based method is derived, which can deal with nonlinear parametric modeling errors, and has been extended to situations in which there are random data transmission failures in [24,25]. An explicit formula of this robust state estimator and its derivations are already given in Chapter 4. More specifically, in the derivations of this robust estimator, an interpretation of the Kalman filter is utilized, which is based on deterministic least squares/maximum likelihood estimations. Moreover, to increase the robustness of the obtained estimator, a penalty is introduced into the cost function on the sensitivity of the so-called innovation process to modeling errors. An outstanding characteristic of this robust estimator is that it can be recursively realized without verifying any conditions. Specifically, this robust estimator has almost the same form as that of the Kalman filter.

In this subsection, we discuss a robustification of the CDOSSP utilizing its relations to the Kalman filter given in the previous subsection.

To achieve this purpose, define the cost function

$$\begin{aligned} J(x(k), d(k)) \;=\; & \frac{1}{2}\Big\{[x(k) - \bar{\hat{x}}(k)]^T (P_i^{[\mathrm{kal}]}(k))^{-1}[x(k) - \bar{\hat{x}}(k)] + d^T(k)d(k) \\ & + [y(k,i) - \bar{C}_i(k)x(k)]^T [D(k,i)D^T(k,i)]^{-1}[y(k,i) - \bar{C}_i(k)x(k)]\Big\}, \end{aligned} \tag{6.32}$$

and let $x^{\mathrm{opt}}(k)$ and $d^{\mathrm{opt}}(k)$ denote respectively the state vector and external disturbance vector that make this cost function achieve its minimum. Obviously, the function $J(x(k), d(k))$ is convex about the variables $x(k)$ and $d(k)$. On the other hand, from the assumptions that both $w(k,i)$ and $d(k,i)$ are normally distributed and independent of each other and from Eq. (6.5) we can easily understand that the prediction result of the Kalman filter at the previous time instant can be interpreted as that $x(k)$ has the normal distribution with expectation $\bar{\hat{x}}(k)$ and covariance matrix $P_i^{[\mathrm{kal}]}(k)$ [18]. With the availability of a new measurement of the plant output vector, that is, $y(k,i)$, some new information about the plant state vector $x(k)$ is obtained, and its estimate should be enhanced by this new measurement, provided that the estimator is optimal. In this enhancement, the aforementioned distribution becomes an a priori knowledge about the plant state vector $x(k)$, and the newly obtained information is used to reduce uncertainty of its estimate. From these aspects, $x^{\mathrm{opt}}(k)$ and $d^{\mathrm{opt}}(k)$ can be respectively interpreted as the newly obtained measurement $y(k,i)$ based maximum likelihood estimates for $x(k)$ and $d(k)$ [18,22]. Hence, the one-step prediction of the plant state vector $x(k+1)$ provided by the Kalman filter using the plant output measurements $y(j,i)|_{j=0}^{k}$, that is, $\bar{\hat{x}}(k+1)$ of Eq. (6.19), is equal to $A(k)x^{\mathrm{opt}}(k) + B_{\mathbf{T}}(k)d^{\mathrm{opt}}(k)$.

To reduce the influence of modeling errors on this one-step ahead state prediction, we adopt the same approach as that utilized in the robustification of the Kalman filter through sensitivity reductions, which is discussed in Chapter 4, that is, to make the associated cost function less selective to modeling errors or, in other words, to reduce the variation magnitude of the desirable values of the variables to be optimized when some or all parameters defining this cost function deviate from their normal values. Obviously, this can be achieved through adding a penalty into the cost function defined by Eq. (6.32) on the derivative vector of the innovation process $y(k,i) - \bar{C}_i(k)x(k)$ with respect to parametric modeling errors, noting that it is the unique factor in the cost function that depends on the parameters of the adopted system model.

This approach is firstly adopted in [22] and [23] to robustify the lumped Kalman filter. In this subsection, we apply these ideas in the robustification of the distributed state predictor derived in Subsection 6.2.1.

Specifically, let $\varepsilon_j(k,i)$, $j = 1, 2, \cdots, m_i$, represent parametric errors in the model of the ith plant subsystem $\mathbf{\Sigma}_i$ at the time instant k, and let $\gamma(k,i)$ be a number belonging to the interval $(0,\ 1)$. Assume that connections among plant subsystems are still described by Eq. (6.2), but there may exist some errors in the transmissions of the plant internal output signals, which make the subsystem connection matrix $\Phi(k)$ no longer exact. In addition, parametric errors are also permitted in the state space model of each plant subsystem, which are independent of each other. More precisely, it is assumed that the dynamics of the plant ith subsystem is described by

$$\begin{bmatrix} x(k+1,i) \\ z(k,i) \\ y(k,i) \end{bmatrix}$$
$$= \begin{bmatrix} A_{\mathbf{xx}}(k,i,\varepsilon(k,i)) & A_{\mathbf{xv}}(k,i,\varepsilon(k,i)) & B_{\mathbf{x}}(k,i,\varepsilon(k,i)) & 0 \\ A_{\mathbf{zx}}(k,i,\varepsilon(k,i)) & A_{\mathbf{zv}}(k,i,\varepsilon(k,i)) & 0 & 0 \\ C_{\mathbf{x}}(k,i,\varepsilon(k,i)) & C_{\mathbf{v}}(k,i,\varepsilon(k,i)) & 0 & D(k,i,\varepsilon(k,i)) \end{bmatrix} \begin{bmatrix} x(k,i) \\ v(k,i) \\ d(k,i) \\ w(k,i) \end{bmatrix}, \tag{6.33}$$

where $\varepsilon(k,i)$ stands for the vector consisting of deviations of the parameters of the ith plant subsystem from their nominal values. This vector usually contains several independent elements. In particular, we assume for a clear presentation that the ith plant subsystem has m_i uncertain and independent parameters and that $\varepsilon(k,i) = \mathbf{col}\{\varepsilon_j(k,i)|_{j=1}^{m_i}\}$. In addition to these, it is also assumed that every subsystem matrix has at least the first-order derivative with respect to each associated parametric error. This assumption is adopted only for avoiding awkward statements. Moreover, the number m_i can also be permitted to vary with the time instant k. This dependence is omitted to make expressions concise although the results remain valid.

Let $\varepsilon_{j,\phi}(k)$, $j = 1, 2, \cdots, m_\phi$, represent parametric errors in the subsystem connection matrix $\Phi(k)$. Similar to the plant state vector $x(k)$, define the parametric error vector

$$\varepsilon(k) = \mathbf{col}\left\{ \mathbf{col}\{\varepsilon_j(k,i)|_{j=1}^{m_i}\}\Big|_{i=1}^{N},\ \mathbf{col}\{\varepsilon_{j,\phi}(k)|_{j=1}^{m_\phi}\}\right\}.$$

Moreover, define every involved matrix, for example, $A_{\mathbf{TT}}(k, \varepsilon(k))$, $A(k, \varepsilon(k))$, and so on, completely as its counterpart without modeling errors. When $\varepsilon_j(k,i) = 0$ for each $j = 1, 2, \cdots, m_i$ and/or for each $i = 1, 2, \cdots, N$, and/or $\varepsilon_{j,\phi}(k) = 0$ for every $j = 1, 2, \cdots, m_\phi$, to avoid complicated expressions, the variables representing modeling errors in these matrices are deleted, and these matrices are abbreviated to their counterparts without modeling errors. Let $P_i^{\mathrm{rob}}(k)$ represent the pseudo-covariance matrix of prediction errors of the robustified OSSP of the dynamic system $\bar{\Sigma}_i$ at time k, and let $\tilde{\hat{x}}(k)$ be the corresponding state prediction. Then, the modified cost function related to this robustified OSSP can be written as

$$\begin{aligned}\bar{J}(x(k), d(k)) &= \gamma(k,i) J(x(k), d(k)) \\ &\quad + \frac{1-\gamma(k,i)}{2}\left[(\star)^T \times \frac{\partial(y(k,i) - \bar{C}_i(k, \varepsilon(k))x(k))}{\partial \varepsilon(k)}\right]\Bigg|_{\varepsilon(k)=0}.\end{aligned} \tag{6.34}$$

Here, $P_i^{[\mathrm{kal}]}(k)$ and $\bar{\hat{x}}(k)$ of Eq. (6.32) are replaced respectively by $P_i^{\mathrm{rob}}(k)$ and $\tilde{\hat{x}}(k)$. Using this modified cost function, we can derive a robust OSSP, which has a similar form as that of the Kalman filter. Its derivation is given in the appendix.

Theorem 6.4. *Denote the matrix* $\mathbf{col}\left\{ \mathbf{col}\left\{ \frac{\partial \bar{C}_i(k,\varepsilon(k))}{\partial \varepsilon_l(k,j)}\Big|_{l=1}^{m_j}\right\}\Big|_{j=1}^{N},\ \frac{\partial \bar{C}_i(k,\varepsilon(k))}{\partial \varepsilon_{l,\phi}(k)}\Big|_{l=1}^{m_\phi}\right\}\Bigg|_{\varepsilon(k)=0}$ *by $H(k,i)$. Assume that the invertible matrix $P_i^{\mathrm{rob}}(k)$ is given. Then, an OSSP of the dynamic system of equation (6.33), which is robust against parametric errors represented by the vector $\varepsilon(k)$, is given as*

$$\tilde{\hat{x}}(k+1) = \hat{A}(k)\tilde{\hat{x}}(k) + K_i^{\mathrm{rob}}(k)[y(k,i) - \bar{C}_i(k)\tilde{\hat{x}}(k)], \tag{6.35}$$

where

$$K_i^{\mathrm{rob}}(k) = \hat{A}(k)\left[(P_i^{\mathrm{rob}}(k))^{-1} + C_i^T(k)C_i(k)\right]^{-1} C_i^T(k)[D(k,i)D^T(k,i)]^{-1/2}, \tag{6.36}$$

$$\begin{aligned}\hat{A}(k) &= A(k)\left[(P_i^{\mathrm{rob}}(k))^{-1} + C_i^T(k)C_i(k) + \frac{1-\gamma(k,i)}{\gamma(k,i)}H^T(k,i)H(k,i)\right]^{-1} \\ &\qquad \times \left[(P_i^{\mathrm{rob}}(k))^{-1} + C_i^T(k)C_i(k)\right].\end{aligned} \tag{6.37}$$

In addition, $P_i^{\mathrm{rob}}(k+1)$ can be recursively computed as

$$P_i^{\mathrm{rob}}(k+1) = \hat{A}(k)\left[(P_i^{\mathrm{rob}}(k))^{-1} + C_i^T(k)C_i(k)\right]^{-1}\hat{A}^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k). \tag{6.38}$$

Through a comparison of Eqs. (6.35)–(6.38) with Eqs. (6.16)–(6.18), we can observe that except the matrix $\hat{A}(k)$, these two one-step ahead state predictors have completely the same form. On the other hand, the matrix $\hat{A}(k)$ can be computed from system matrices and the matrix $P_i^{\mathrm{rob}}(k)$. Moreover, this matrix equals $A(k)$ when the design parameter $\gamma(k,i)$ is chosen as $\gamma(k,i)=1$; that is, when robustness against modeling errors is not taken into account, this state predictor reduces to the Kalman filter. This conclusion is in a good agreement with Eq. (6.34), in which the associated cost function equals to that when $\gamma(k,i)=1$.

From these observations and the relations given in the previous subsection among the CDOSSP and from the Kalman filter and least squares/maximum likelihood estimations we can declare that when the plant model with parametric errors is described by Eq. (6.33), a reasonable approach to robustify the CDOSSP of Eq. (6.3), is to assign its gain matrix $K_{\mathbf{T}}(k,i)$ as $K_{\mathbf{T}}(k,i)=J_{\mathbf{x}i}^T K_i^{\mathrm{rob}}(k)$, $i=1,2,\cdots,N$. On the basis of these gain matrices, we can derive a recursive formula similar to that of Eq. (6.17) for the associated pseudo-covariance matrix of prediction errors. This means that the robustified OSSP can still be realized in a distributed way and it is feasible to scale it to a linear time-varying plant with a large number of subsystems.

6.3 Distributed State Filtering

In the previous section, we derived a distributed state predictor for a plant with several subsystems using the criteria of unbiasedness and minimal covariance matrix of local estimation errors. In this section, we investigate distributed state filtering with similar ideas.

To this purpose, for each time instant k and every subsystem $\mathbf{\Sigma}_i$, define the matrices

$$\begin{aligned}
&\bar{A}_{\mathbf{xv}}(k,i)=\left[0_{m_{\mathbf{x}i}\times m_{\mathbf{v}i}}\ \ A_{\mathbf{xv}}(k,i)\right], \quad \bar{A}_{\mathbf{zx}}(k,i)=\left[\begin{array}{c} A_{\mathbf{zx}}(k+1,i)A_{\mathbf{xx}}(k,i) \\ A_{\mathbf{zx}}(k,i) \end{array}\right], \\
&\bar{A}_{\mathbf{zv}}(k,i)=\left[\begin{array}{cc} A_{\mathbf{zv}}(k+1,i) & A_{\mathbf{zx}}(k+1,i)A_{\mathbf{xv}}(k,i) \\ 0_{m_{\mathbf{z}i}\times m_{\mathbf{v}i}} & A_{\mathbf{zv}}(k,i) \end{array}\right], \\
&\bar{C}_{\mathbf{x}}(k,i)=C_{\mathbf{x}}(k+1,i)A_{\mathbf{xx}}(k,i), \quad \bar{C}_{\mathbf{v}}(k,i)=\left[C_{\mathbf{v}}(k+1,i)\ \ C_{\mathbf{x}}(k+1,i)A_{\mathbf{xv}}(k,i)\right].
\end{aligned}$$

Clearly, these matrices are well defined and uniquely determined by the parameters of the plant ith subsystem.

To get an estimate for the states $x(k,i)|_{i=1}^N$ of the dynamical system $\mathbf{\Sigma}$ from the output measurements $y(k,i)|_{i=1}^N$, we construct the following observer $\bar{\mathbf{\Sigma}}$. This observer is also constituted from N subsystems, but the state space model of its ith subsystem $\bar{\mathbf{\Sigma}}_i$ is given by

$$\begin{bmatrix} \hat{x}(k+1,i) \\ \hat{z}(k,i) \\ \hat{y}(k+1,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k,i) & \bar{A}_{\mathbf{xv}}(k,i) & K_{\mathbf{x}}(k,i) \\ \bar{A}_{\mathbf{zx}}(k,i) & \bar{A}_{\mathbf{zv}}(k,i) & 0 \\ \bar{C}_{\mathbf{x}}(k,i) & \bar{C}_{\mathbf{v}}(k,i) & 0 \end{bmatrix} \begin{bmatrix} \hat{x}(k,i) \\ \hat{v}(k,i) \\ y(k+1,i) - \hat{y}(k+1,i) \end{bmatrix}. \tag{6.39}$$

Moreover, these subsystems $\bar{\mathbf{\Sigma}}_i|_{i=1}^N$ are connected through the following relation using the subsystem connection matrix of the plant at the time instants k and $k+1$:

$$\hat{v}(k) = \bar{\Phi}(k)\hat{z}(k), \quad \bar{\Phi}(k) = \mathbf{diag}\{\Phi(k+1), \quad \Phi(k)\}, \tag{6.40}$$

where $\hat{z}(k)$ and $\hat{v}(k)$ are defined in a similar way as those $z(k)$ and $v(k)$ of the plant. More specifically, $\hat{z}(k) = \mathbf{col}\left\{\hat{z}_1(k,i)|_{i=1}^N, \ \hat{z}_2(k,i)|_{i=1}^N\right\}$ and $\hat{v}(k) = \mathbf{col}\left\{\hat{v}_1(k,i)|_{i=1}^N, \ \hat{v}_2(k,i)|_{i=1}^N\right\}$ with $\hat{z}_1(k,i)$ and $\hat{z}_2(k,i)$ standing respectively for the vectors consisting of the first and last $m_{\mathbf{z}i}$ elements of the internal output vector $\hat{z}(k,i)$ and $\hat{v}_1(k,i)$ and $\hat{v}_2(k,i)$ those of the first and last $m_{\mathbf{v}i}$ elements of the internal input vector $\hat{v}(k,i)$.

Similar to those in the one-step ahead predictor designs, the objective of this section is to find the optimal gain matrix $K_{\mathbf{x}}(k,i)$, $k = 0, 1, 2, \cdots, i = 1, 2, \cdots, N$, such that for every subsystem, the state estimate $\hat{x}(k,i)$ is unbiased and the covariance matrix of its estimation errors is minimized together with a recursive formula for its realization.

For this purpose and to simplify mathematical formulas, the same hypotheses are adopted as those of the previous section on plant process disturbances and measurement errors; that is, $d(k_1,i_1)$ and $w(k_2,i_2)$ are assumed to be independent of each other for all k_1, k_2, i_1, and i_2, and the covariance matrix between arbitrary $d(k_1,i_1)$ and $d(k_2,i_2)$, or between arbitrary $w(k_1,i_1)$ and $w(k_2,i_2)$, is equal to zero whenever $k_1 \neq k_2$ or $i_1 \neq i_2$. In addition, the covariance matrices of $d(k,i)$ and $w(k,i)$ are assumed to be equal to the identity matrix. Furthermore, mathematical expectations of these process disturbances and measurement errors are assumed to be zero.

Comparisons between Eqs. (6.4) and (6.40) make it clear that although the observer to be designed has formally the same structure as that of the system to be estimated, their subsystem connection matrices are different from each other. More precisely, the dimensions of the internal input/output vectors $\hat{v}(k,i)$ and $\hat{z}(k,i)$ of the state estimator are respectively twice as those of the plant corresponding vectors $v(k,i)$ and $z(k,i)$. Moreover, the system matrices $\bar{A}_{\mathbf{xv}}(k,i)$, $\bar{A}_{\mathbf{zx}}(k,i)$, and so on are different from their counterparts of the plant. In addition, the subsystem connection matrix $\bar{\Phi}(k)$ depends not only on the current subsystem connection matrix $\Phi(k+1)$ of the plant, but also on its subsystem connection matrix at the previous time instant, that is, $\Phi(k)$. This is different from the one-step state predictor derived in the previous section and from traditional observer designs and the well-known Kalman filter [18,20,26]. These differences make distributed state filtering mathematically more involved.

To have concise expressions, the following matrix representations are adopted in the rest of this section: $A_{\mathbf{*\#}}(k) = \mathbf{diag}\left\{A_{\mathbf{*\#}}(k,i)|_{i=1}^{N}\right\}$, $\tilde{A}_{\mathbf{*\#}}(k) = \mathbf{diag}\left\{\bar{A}_{\mathbf{*\#}}(k,i)|_{i=1}^{N}\right\}$, $B_{*}(k) = \mathbf{diag}\left\{B_{*}(k,i)|_{i=1}^{N}\right\}$, $C_{*}(k) = \mathbf{diag}\left\{C_{*}(k,i)|_{i=1}^{N}\right\}$, $\tilde{C}_{*}(k) = \mathbf{diag}\left\{\bar{C}_{*}(k,i)|_{i=1}^{N}\right\}$, $D(k) = \mathbf{diag}\left\{D(k,i)|_{i=1}^{N}\right\}$, and $K_{\mathbf{x}}(k) = \mathbf{diag}\left\{K_{\mathbf{x}}(k,i)|_{i=1}^{N}\right\}$, where $\mathbf{*}, \mathbf{\#} = \mathbf{x},\ \mathbf{v}$. Moreover, denote $\mathbf{col}\left\{d(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{w(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{y(k,i)|_{i=1}^{N}\right\}$, and $\mathbf{col}\left\{\hat{y}(k,i)|_{i=1}^{N}\right\}$ respectively by $d(k)$, $w(k)$, $y(k)$, and $\hat{y}(k)$. Then, the dynamics of the plant $\mathbf{\Sigma}$ and its state estimator given by Eqs. (6.1), (6.2), (6.39), and (6.40) can be equivalently expressed as

$$\begin{bmatrix} x(k+1) \\ z(k) \\ y(k) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k) & A_{\mathbf{xv}}(k) & B_{\mathbf{x}}(k) & 0 \\ A_{\mathbf{zx}}(k) & A_{\mathbf{zv}}(k) & 0 & 0 \\ C_{\mathbf{x}}(k) & C_{\mathbf{v}}(k) & 0 & D(k) \end{bmatrix} \begin{bmatrix} x(k) \\ v(k) \\ d(k) \\ w(k) \end{bmatrix}, \tag{6.41}$$

$$\begin{bmatrix} \hat{x}(k+1) \\ \hat{z}(k) \\ \hat{y}(k+1) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(k) & \tilde{A}_{\mathbf{xv}}(k) & K_{\mathbf{x}}(k) \\ \tilde{A}_{\mathbf{zx}}(k) & \tilde{A}_{\mathbf{zv}}(k) & 0 \\ \tilde{C}_{\mathbf{x}}(k) & \tilde{C}_{\mathbf{v}}(k) & 0 \end{bmatrix} \begin{bmatrix} \hat{x}(k) \\ \hat{v}(k) \\ y(k+1) - \hat{y}(k+1) \end{bmatrix}. \tag{6.42}$$

On the basis of these representations, through canceling the internal input/output vectors $z(k)$ and $v(k)$ of the plant, straightforward matrix manipulations show that the input–output relations of the dynamic system $\mathbf{\Sigma}$ can be further expressed by Eq. (6.5), which is rewritten here for references in deriving the distributed state estimator:

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \left\{ \begin{bmatrix} A_{\mathbf{xx}}(k) & B_{\mathbf{x}}(k) & 0 \\ C_{\mathbf{x}}(k) & 0 & D(k) \end{bmatrix} + \begin{bmatrix} A_{\mathbf{xv}}(k) \\ C_{\mathbf{v}}(k) \end{bmatrix} \Phi(k)\,[\,I - A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}\,[A_{\mathbf{zx}}(k)\ \ 0\ \ 0] \right\} \begin{bmatrix} x(k) \\ d(k) \\ w(k) \end{bmatrix}. \tag{6.43}$$

Define the matrices

$$A(k) = A_{\mathbf{xx}}(k) + A_{\mathbf{xv}}(k)\Phi(k)\,[I - A_{\mathbf{zv}}(k)\Phi(k)]^{-1}\,A_{\mathbf{zx}}(k)$$

and

$$C(k) = C_{\mathbf{x}}(k) + C_{\mathbf{v}}(k)\Phi(k)\,[I - A_{\mathbf{zv}}(k)\Phi(k)]^{-1}\,A_{\mathbf{zx}}(k).$$

Then, from Eq. (6.43) we have that

$$\begin{aligned} x(k+1) &= A(k)x(k) + B_{\mathbf{x}}(k)d(k), \\ y(k+1) &= C(k+1)x(k+1) + D(k+1)w(k+1) \end{aligned} \tag{6.44}$$

$$= \quad C(k+1)A(k)x(k) + C(k+1)B_{\mathbf{x}}(k)d(k) + D(k+1)w(k+1). \tag{6.45}$$

These equalities make it clear that whereas the relations between the plant state vectors of two successive time instants k and $k+1$ only depend on the subsystem interconnection matrix $\Phi(k)$, the relations between the plant output vector $y(k+1)$ and the plant state vector $x(k)$ depend not only on this matrix at the time constant $k+1$, but also on its value at the previous time instant k. This means that to predict the plant outputs at the time instant $k+1$ using its current state estimates, information on both $\Phi(k)$ and $\Phi(k+1)$ is necessary. This may mean that the structure of the suggested state estimator given in Eqs. (6.39) and (6.40) is reasonable.

On the other hand, from Eqs. (6.39) and (6.40) we can obtain the following input–output relations for the suggested plant state estimator (derivations are deferred to the appendix):

$$\hat{x}(k+1) \quad = \quad A(k)\hat{x}(k) + K_{\mathbf{x}}(k)[y(k+1) - \hat{y}(k+1)], \tag{6.46}$$

$$\hat{y}(k+1) \quad = \quad C(k+1)A(k)\hat{x}(k). \tag{6.47}$$

Denote $\hat{x}(k,i) - x(k,i)$ and $\mathbf{col}\left\{\tilde{x}(k,i)|_{i=1}^{N}\right\}$ respectively by $\tilde{x}(k,i)$ and $\tilde{x}(k)$. Then, straightforward matrix operations from these relations show that

$$\begin{aligned}\tilde{x}(k+1) \quad = \quad & [I - K_{\mathbf{x}}(k)C(k+1)]A(k)\tilde{x}(k) + K_{\mathbf{x}}(k)D(k+1)w(k+1) \\ & - [I - K_{\mathbf{x}}(k)C(k+1)]B_{\mathbf{x}}(k)d(k).\end{aligned} \tag{6.48}$$

From this recursive expression for estimation errors and the adopted hypotheses on process disturbances $w(k,i)$ and measurement errors $d(k,i)$ we can further prove that

$$\mathbf{E}\{\tilde{x}(k+1)\} = [I - K_{\mathbf{x}}(k)C(k+1)]A(k)\mathbf{E}\{\tilde{x}(k)\}, \tag{6.49}$$

$$\begin{aligned}\mathbf{E}\{\tilde{x}(k+1)\tilde{x}^T(k+1)\} = {} & [I - K_{\mathbf{x}}(k)C(k{+}1)]A(k)\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}A^T(k)[I - K_{\mathbf{x}}(k)C(k{+}1)]^T \\ & + K_{\mathbf{x}}(k)D(k+1)D^T(k+1)K_{\mathbf{x}}^T(k) \\ & + [I - K_{\mathbf{x}}(k)C(k+1)]B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k)[I - K_{\mathbf{x}}(k)C(k+1)]^T.\end{aligned} \tag{6.50}$$

Clearly, if the estimate at the time instant k is unbiased, then the estimator of Eqs. (6.39) and (6.40) certainly provides an unbiased estimate at the next time instant. On the other hand, as $\mathbf{E}\{\tilde{x}(k+1)\tilde{x}^T(k+1)\}$ depends quadratically on each update gain matrix $K_{\mathbf{x}}(k,i)$ of the state estimator, a globally optimal gain matrix, denote it by $K_{\mathbf{x}}^{\mathrm{opt}}(k,i)$, can be obtained under the condition that $\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}$ is available. These results are established on the basis of a relation between local and global optimality of the update gain matrix, which are similar to those on the one-step predictor design discussed in [4] and Section 6.2. The details are omitted. An interested reader can establish this relation by mimicking Lemma 6.1.

From the relation given by Eq. (6.50) we can obtain explicit formulas for the optimal gain matrix and the covariance matrix of the corresponding estimation error. These results are parallel to those on state predictions given further in Theorems 6.5 and 6.6.

Denote $\mathbf{E}\{\tilde{x}(k)\tilde{x}^T(k)\}$ and $\mathbf{E}\{\tilde{x}(k,i)\tilde{x}^T(k,j)\}$ respectively by $P(k)$ and $P_{ij}(k)$, $i,j = 1,2,\cdots,N$. Moreover, define integer $M_{\star i}$ as $M_{\star i}=0$ when $i=1$ and $M_{\star i}=\sum_{k=1}^{i-1} m_{\star k}$ when $2 \le i \le N$. Let $J_{\star i}$ denote the matrix $\mathbf{col}\left\{0_{M_{\star i}\times m_{\star i}}, I_{m_{\star i}}, 0_{(m_\star - M_{\star,i+1})\times m_{\star i}}\right\}$, where $\star = \mathbf{x}, \mathbf{y}, \mathbf{v}$. Then, from the definitions of the matrices $P(k)$ and $P_{ij}(k)$ we can straightforwardly prove that

$$P_{ij}(k) = J_{\mathbf{x}i}^T P(k) J_{\mathbf{x}j}, \quad \forall\, i,j = 1,2,\cdots,N. \tag{6.51}$$

From this relation we obtain a recursive expression for the optimal update gain matrix $K_{\mathbf{x}}^{\text{opt}}(k,i)$.

Theorem 6.5. *Let $C_i(k+1)$ represent $J_{\mathbf{y}i}^T C(k+1)$. If the matrix $D(k+1,i)$ is of full row rank, then, for every subsystem $\mathbf{\Sigma}_i$ and every time instant k, the optimal observer gain matrix $K_{\mathbf{x}}^{\text{opt}}(k,i)$ minimizing $P_{ii}(k+1)$ can be expressed as*

$$K_{\mathbf{x}}^{\text{opt}}(k,i) = J_{\mathbf{x}i}^T \Xi(k) C_i^T(k+1)\left\{C_i(k+1)\Xi(k)C_i^T(k+1) + D(k+1,i)D^T(k+1,i)\right\}^{-1}, \tag{6.52}$$

where $\Xi(k) = A(k)P(k)A^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k)$.

Proof. From Eq. (6.50) straightforward algebraic manipulations show that

$$\begin{aligned} P(k+1) \;&=\; K_{\mathbf{x}}(k)[C(k+1)\Xi(k)C^T(k+1) + D(k+1)D^T(k+1)]K_{\mathbf{x}}^T(k) \\ &\quad - \Xi(k)C^T(k+1)K_{\mathbf{x}}^T(k) - K_{\mathbf{x}}(k)C(k+1)\Xi(k) + \Xi(k). \end{aligned} \tag{6.53}$$

Note that, for every $i = 1,2,\cdots,N$,

$$\begin{aligned} J_{\mathbf{x}i}^T K_{\mathbf{x}}(k) \;&=\; [0 \;\cdots\; K_{\mathbf{x}}(k,i)\;\; 0 \;\cdots\; 0] \\ &=\; K_{\mathbf{x}}(k,i) J_{\mathbf{y}i}^T. \end{aligned} \tag{6.54}$$

Combining this relation with Eq. (6.51), we have that

$$\begin{aligned} P_{ii}(k+1) \;&=\; J_{\mathbf{x}i}^T\{K_{\mathbf{x}}(k)[C(k+1)\Xi(k)C^T(k+1) + D(k+1)D^T(k+1)]K_{\mathbf{x}}^T(k) \\ &\quad - \Xi(k)C^T(k+1)K_{\mathbf{x}}^T(k) - K_{\mathbf{x}}(k)C(k+1)\Xi(k) + \Xi(k)\}J_{\mathbf{x}i} \end{aligned}$$

$$\begin{aligned}
&= K_{\mathbf{x}}(k,i)J_{\mathbf{y}i}^T[C(k+1)\Xi(k)C^T(k+1)+D(k+1)D^T(k+1)]J_{\mathbf{y}i}K_{\mathbf{x}}^T(k,i)\\
&\quad - J_{\mathbf{x}i}^T\Xi(k)C^T(k+1)J_{\mathbf{y}i}K_{\mathbf{x}}^T(k,i) - K_{\mathbf{x}}(k,i)J_{\mathbf{y}i}^TC(k+1)\Xi(k)J_{\mathbf{x}i}\\
&\quad + J_{\mathbf{x}i}^T\Xi(k)J_{\mathbf{x}i}\\
&= K_{\mathbf{x}}(k,i)[C_i(k+1)\Xi(k)C_i^T(k+1)+D(k+1,i)D^T(k+1,i)]K_{\mathbf{x}}^T(k,i)\\
&\quad - J_{\mathbf{x}i}^T\Xi(k)C_i^T(k+1)K_{\mathbf{x}}^T(k,i) - K_{\mathbf{x}}(k,i)C_i(k+1)\Xi(k)J_{\mathbf{x}i} + J_{\mathbf{x}i}^T\Xi(k)J_{\mathbf{x}i}\\
&= [K_{\mathbf{x}}(k,i)-K_{\mathbf{x}}^{\mathrm{opt}}(k,i)][C_i(k+1)\Xi(k)C_i^T(k+1)\\
&\quad + D(k+1,i)D^T(k+1,i)][K_{\mathbf{x}}(k,i)-K_{\mathbf{x}}^{\mathrm{opt}}(k,i)]^T\\
&\quad - J_{\mathbf{x}i}^T\Xi(k)C_i^T(k+1)[C_i(k+1)\Xi(k)C_i^T(k+1)\\
&\quad + D(k+1,i)D^T(k+1,i)]C_i(k+1)\Xi(k)J_{\mathbf{x}i} + J_{\mathbf{x}i}^T\Xi(k)J_{\mathbf{x}i}.
\end{aligned} \tag{6.55}$$

From the definition of the matrix $\Xi(k)$ and the assumption that the matrix $D(k+1,i)$ is of full row rank it is obvious that the matrix $C_i(k+1)\Xi(k)C_i^T(k+1)+D(k+1,i)D^T(k+1,i)$ is positive definite. We can therefore declare that the matrix $K_{\mathbf{x}}^{\mathrm{opt}}(k,i)$ is the unique optimal update gain matrix for the ith subsystem of the plant.

This completes the proof. □

When this optimal gain matrix is adopted in the state estimator, the covariance matrix of the corresponding estimation errors can also be recursively computed. The results are given in the next theorem.

Theorem 6.6. *Assume that the optimal gain matrices $K_{\mathbf{x}}^{\mathrm{opt}}(k,i)|_{i=1}^N$ are adopted in the distributed state observer. Then, for each $i,\ j=1,2,\cdots,N$, the submatrix $P_{ij}(k+1)$ of the covariance matrix of the estimation errors can be expressed as follows,*

$$P_{ij}(k+1) = \begin{cases} J_{\mathbf{x}i}^T\left[\Xi^{-1}(k)+\bar{C}_i^T(k+1)\bar{C}_i(k+1)\right]^{-1}J_{\mathbf{x}i}, \\ \quad i=j, \\ J_{\mathbf{x}i}^T\left[\Xi^{-1}(k)+\bar{C}_i^T(k+1)\bar{C}_i(k+1)\right]^{-1}\Xi^{-1}(k)\left[\Xi^{-1}(k)+\bar{C}_j^T(k+1)\bar{C}_j(k+1)\right]^{-1}J_{\mathbf{x}j}, \\ \quad i\neq j, \end{cases} \tag{6.56}$$

where $\bar{C}_i(k)=[D(k,i)D^T(k,i)]^{-1/2}C_i(k)$ and $\bar{D}_i(k)=[D(k,i)D^T(k,i)]^{-1/2}J_{\mathbf{y}i}^TD(k)$.

Proof. From Eqs. (6.51) and (6.53) we can establish the following equality for all $i,\ j=1,2,\cdots,N$:

$$P_{ij}(k+1) = J_{\mathbf{x}i}^T\{K_{\mathbf{x}}(k)[C(k+1)\Xi(k)C^T(k+1)+D(k+1)D^T(k+1)]K_{\mathbf{x}}^T(k)$$

$$
\begin{aligned}
&- \Xi(k)C^T(k+1)K_{\mathbf{x}}^T(k) - K_{\mathbf{x}}(k)C(k+1)\Xi(k) + \Xi(k)\}J_{\mathbf{x}j} \\
&= K_{\mathbf{x}}(k,i)J_{\mathbf{y}i}^T[C(k+1)\Xi(k)C^T(k+1) + D(k+1)D^T(k+1)]J_{\mathbf{y}j}K_{\mathbf{x}}^T(k,j) \\
&- J_{\mathbf{x}i}^T\Xi(k)C^T(k+1)J_{\mathbf{y}j}K_{\mathbf{x}}^T(k,j) - K_{\mathbf{x}}(k,i)J_{\mathbf{y}i}^T C(k+1)\Xi(k)J_{\mathbf{x}j} + J_{\mathbf{x}i}^T\Xi(k)J_{\mathbf{x}j}.
\end{aligned}
\tag{6.57}
$$

In this equation, replace the submatrices $K_{\mathbf{x}}(k,i)$ and $K_{\mathbf{x}}(k,j)$ of the state update gain matrix $K_{\mathbf{x}}(k)$ respectively by their optimal values, that is, $K_{\mathbf{x}}^{\text{opt}}(k,i)$ and $K_{\mathbf{x}}^{\text{opt}}(k,j)$. Then direct matrix operations show that

$$
\begin{aligned}
P_{ij}(k+1) = J_{\mathbf{x}i}^T \Big\{ &\Xi(k)\bar{C}_i^T(t+1)[\bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1)+I]^{-1}[\bar{C}_i(k+1)\Xi(k)\bar{C}_j^T(k+1) \\
&+ \bar{D}(k+1,i)\bar{D}^T(k+1,j)][\bar{C}_j(k+1)\Xi(k)\bar{C}_j^T(k+1)+I]^{-1}\bar{C}_j(k+1)\Xi(k) \\
&+ \Xi(k) - \Xi(k)\bar{C}_i^T(k+1)[\bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1)+I]^{-1}\bar{C}_i(k+1)\Xi(k) \\
&- \Xi(k)\bar{C}_j^T(k+1)[\bar{C}_j(k+1)\Xi(k)\bar{C}_j^T(k+1)+I]^{-1}\bar{C}_j(k+1)\Xi(k)\Big\} J_{\mathbf{x}j}.
\end{aligned}
\tag{6.58}
$$

Note that, for every $k = 1, 2, \cdots, N$,

$$
\begin{aligned}
&\Xi(k)\bar{C}_k^T(k+1)[\bar{C}_k(k+1)\Xi(k)\bar{C}_k^T(k+1)+I]^{-1}\bar{C}_k(k+1) \\
=\ &[\Xi(k)\bar{C}_k^T(t+1)\bar{C}_k(k+1)+I]^{-1}\Xi(k)\bar{C}_k^T(t+1)\bar{C}_k(k+1) \\
=\ &I - [\Xi(k)\bar{C}_k^T(t+1)\bar{C}_k(k+1)+I]^{-1} \\
=\ &I - [\bar{C}_k^T(t+1)\bar{C}_k(k+1)+\Xi^{-1}(k)]^{-1}\Xi^{-1}(k)
\end{aligned}
\tag{6.59}
$$

and

$$
\bar{D}_i(k+1)\bar{D}_j^T(k+1) = \begin{cases} I_{m_{\mathbf{y}i}}, & i = j, \\ 0_{m_{\mathbf{y}i}\times m_{\mathbf{y}j}}, & i \neq j. \end{cases}
\tag{6.60}
$$

The proof can now be completed through substituting Eqs. (6.59) and (6.60) into Eq. (6.58) and some straightforward algebraic manipulations. □

Although the theorem gives a recursive and relatively compact expression for the covariance matrix of estimation errors for the optimal distributed state estimator, it is not convenient to be utilized in state estimations of a large-scale networked system. Note that for these systems, the dimension of the plant state vector and therefore the dimension of the associated matrix are generally great. This large dimension of $\Xi(k)$ may cause numerical instability problems and prohibitive computational complexities in computing the inverse of the matrices Ξ_k and $\Xi^{-1}(k) + \bar{C}_i^T(k+1)\bar{C}_i(k+1)$, $i = 1, 2, \cdots, N$. To overcome these difficulties, we derive another expression for this covariance matrix through some algebraic manipulations.

Corollary 6.2. *Assume that the optimal distributed state observer is adopted. Then, for each $i = 1, 2, \cdots, N$, the submatrix $P_{ii}(k+1)$ of the covariance matrix $P(k+1)$ can be equivalently represented as*

$$\begin{aligned} P_{ii}(k+1) &= J_{\mathbf{x}i}^T \left\{ \Xi(k) - \Xi(k)\bar{C}_i^T(k+1)\left[\bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1) + I\right]^{-1} \right. \\ &\quad \left. \times\, \bar{C}_i(k+1)\Xi(k) \right\} J_{\mathbf{x}i}. \end{aligned} \tag{6.61}$$

Proof. Note that

$$\begin{aligned} &\left[\Xi^{-1}(k) + \bar{C}_i^T(k+1)\bar{C}_i(k+1) + I\right]^{-1} \\ = \; & \Xi(k)\left[I + \bar{C}_i^T(k+1)\bar{C}_i(k+1)\Xi(k)\right]^{-1} \\ = \; & \Xi(k)\left\{I - \left[I + \bar{C}_i^T(k+1)\bar{C}_i(k+1)\Xi(k)\right]^{-1}\bar{C}_i^T(k+1)\bar{C}_i(k+1)\Xi(k)\right\} \\ = \; & \Xi(k)\left\{I - \bar{C}_i^T(k+1)\left[I + \bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1)\right]^{-1}\bar{C}_i(k+1)\Xi(k)\right\} \\ = \; & \Xi(k) - \Xi(k)\bar{C}_i^T(k+1)\left[I + \bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1)\right]^{-1}\bar{C}_i(k+1)\Xi(k). \end{aligned} \tag{6.62}$$

The proof can now be completed through substituting this relation into Eq. (6.56). □

From Eqs. (6.56) and (6.61) it is clear that in calculating the covariance matrix $P(k+1)$, only the inverse of the matrix $\bar{C}_i(k+1)\Xi(k)\bar{C}_i^T(k+1) + I$, $i = 1, 2, \cdots, N$, is required. Note that rather than the dimension of the state vector of the whole system, the dimension of this matrix is equal to that of the external output vector of the plant ith subsystem $\mathbf{\Sigma}_i$. Clearly, this dimension does not increase with the number of plant subsystems. This means that the expression of the submatrix $P_{ii}(k)$ in Eqs. (6.61) is more attractive in state estimations for a large-scale system.

6.4 Asymptotic Property of the Distributed Observers

As declared in Chapter 5, the Kalman filter is the optimal state estimator when the plant is linear and the external disturbances, including both the process noises and measurement errors, are normally distributed. On the other hand, the state predictor and estimator developed respectively in Sections 6.2 and 6.3 aim to optimally estimate plant subsystem states using only local output measurements. An interesting theoretical issue, which is also of great engineering significance, is that whether or not there exist situations, under which these state estimators have an estimation accuracy as high as that of the Kalman filter. Whereas this is in general still a challenging theoretical problem, we clarify these situations in this section under the condition that the one-step ahead state predictor given in Section 6.2 converges to a steady state, for which its steady prediction accuracy is equal to that of the Kalman filter.

To achieve this purpose, the following results are first derived, which reveal differences in estimation accuracy between the developed CDOSSP and the Kalman filter.

Lemma 6.2. *Assume that the matrix $D_{\mathbf{w}}(t)$ is of full row rank. Moreover, assume that the lumped Kalman filter is applied to the dynamic system $\mathbf{\Sigma}$. Let $P^{[\mathrm{kal}]}(k)$ denote the covariance matrix of its estimation errors at the time instant k, and let $P^{[\mathrm{kal}]}_{ij}(t)$ be its ith row jth column block submatrix. Then*

$$\begin{aligned}
& P_{ii}(k+1) - P^{[\mathrm{kal}]}_{ii}(k+1) \\
= \; & A_i(k)\left\{[P^{[\mathrm{kal}]}(k)]^{-1} + \sum_{j=1}^{N}\bar{C}_j^T(k)\bar{C}_j(k)\right\}^{-1} \left\{[P^{[\mathrm{kal}]}(k)]^{-1} - P^{-1}(k)\right. \\
& \left. + \sum_{j=1, j\neq i}^{N} \bar{C}_j^T(k)\bar{C}_j(k)\right\}[P^{-1}(t) + \bar{C}_i^T(k)\bar{C}_i(k)]^{-1}A_i^T(k).
\end{aligned} \tag{6.63}$$

Based on Lemma 6.2, we can derive a necessary and sufficient condition on system matrices for the equivalence between the estimation accuracy of the CDOSSP and that of the Kalman filter, which is given in the following theorem. The associated proof is deferred to the appendix of this chapter.

Theorem 6.7. *Assume that $P(k) = P^{[\mathrm{kal}]}(k)$ at some time instant k. Then, at the next time instant $k+1$, for the ith plant subsystem $\mathbf{\Sigma}_i$, the CDOSSP has a covariance matrix of estimation errors equal to that of the Kalman filter if and only if for each $j = 1, 2, \cdots, N,\ j \neq i$,*

$$A_i(k)P(k)\bar{C}^T(k)[I + \bar{C}(k)P(k)\bar{C}^T(k)]^{-1}J_{\mathbf{y}j} = 0, \tag{6.64}$$

where $\bar{C}(k) = \mathbf{col}\{\bar{C}_i(k)|_{i=1}^{N}\}$. Moreover, let $K^{[\mathrm{kal}]}(k)$ denote the update gain matrix of the Kalman filter, and let $K^{[\mathrm{kal}]}_{ij}(k)$ be its ith row jth column block. Then, when this condition is satisfied, $P_{ij}(k+1) = P^{[\mathrm{kal}]}_{ij}(k+1)$ and $K^{[\mathrm{kal}]}_{ij}(k) = 0$ are simultaneously valid for arbitrary $j \neq i$.

In general, the condition $P(k) = P^{[\mathrm{kal}]}(k)$ of the theorem cannot be easily verified. However, when only the steady estimation accuracy of the CDOSSP is required to be compared to that of the Kalman filter, this condition becomes necessary. Note that in the steady state of an estimator, the covariance matrix of its estimation errors does not vary with the temporal variable.

Note also that when system $\boldsymbol{\Sigma}$ is time invariant, the convergence of its Kalman filter is guaranteed by its controllability and observability, which can be verified through its subsystem dynamics and subsystem connection matrix using the results of Chapter 3. Based on these observations and Theorem 6.7, characteristics of systems are clarified whose steady estimation accuracy with the CDOSSP is equal to that of the Kalman filter.

Theorem 6.8. *Assume that both the subsystem parameter matrices of system $\boldsymbol{\Sigma}$ and its subsystem connection matrix are time invariant. Moreover, assume that the covariance matrix of estimation errors of its Kalman filter converges to a constant and positive definite matrix P. Then, the CDOSSP can have the same steady estimation accuracy as the Kalman filter if and only if for every $i = 1, 2, \cdots, N$,*

$$A_i P \bar{C}^T [I + \bar{C} P \bar{C}^T]^{-1} J_{\mathbf{y}j} = 0, \quad \forall j \neq i. \tag{6.65}$$

A proof of these results is given in the appendix of this chapter.

Theorems 6.7 and 6.8 make it clear that if the Kalman filter of system $\boldsymbol{\Sigma}$ converges, then to guarantee that the CDOSSP has the same steady estimation accuracy as the Kalman filter, it is necessary and sufficient that the Kalman filter has a steady block diagonal update gain matrix. The sufficiency of this condition can be easily imagined, as it simply means that the Kalman filter can be realized in a distributed way, which is certainly optimal. Its necessity, however, clarifies that this is the only situation that these two estimators have the same steady estimation accuracy. Although this condition may be severe in practice, numerical simulations in [27] show that element magnitude of the matrix $A_i P \bar{C}^T [I + \bar{C} P \bar{C}^T]^{-1} J_{\mathbf{y}j}$ appears to be a good indicator on the accuracy difference between these two estimators. These phenomena have been observed in various numerical examples, including one in the next section.

6.5 Distributed State Estimation Through Neighbor Information Exchanges

State estimations with local measurements discussed in the previous sections reveal that a coordination unit is generally necessary to make the estimates optimal. In [28], it has been shown that if a controller is required to meet some constraints, then a so-called quadratic invariance condition on the system is both necessary and sufficient for the constraint set to be preserved under feedback. Although this quadratic invariance condition is stringent in general and can hardly be satisfied in actual applications, that paper also shows that in case that a distributed controller communicates among its subsystems faster than the propagation of the plant dynamics, an optimal stabilizing controller can be efficiently computed. Note that control and estimation have been proven to be dual to each other, and feedback also exists in state estimations [8,19]. This is also clear from the structure of the Luenberger observer given in Fig. 4.1. The results in [28] open great opportunities to achieve an approximate global optimization in state estimations through introducing interactions between communication and state estimate updates. Numerical simulations in [29] have confirmed efficiency of this approach in distributed state estimations for an interconnected system described by Eqs. (6.1) and (6.2), but no theoretical analysis has been provided.

In this section, a distributed estimation problem is discussed in which only plant measurements are utilized without taking the plant dynamics into account. This problem is attacked in [30] and reflects many important characteristics in the interplays between communications and state estimations. One of the attractive properties of the results given in [30] is that under some situations, iterations for computing the global optimum end in a finite number of steps.

In particular, consider a system $\mathbf{\Sigma}$ constituted from N subsystems with the outputs of its ith subsystem $\mathbf{\Sigma}_i$ described as

$$y_i = \sum_{j=1}^{N} C_{ij} x_j + w_i, \quad i = 1, 2, \cdots, N, \tag{6.66}$$

where $x_i \in \mathcal{R}^{m_{\mathbf{x}i}}$, $y_i \in \mathcal{R}^{m_{\mathbf{y}i}}$, and $w_i \in \mathcal{R}^{m_{\mathbf{w}i}}$ represent respectively the state vector, output vector, and measurement error vector of the ith system. We assume that w_i and w_j are independent of each other whenever $i \neq j$ and that

$$w_i \sim \mathcal{N}(0,\ R_i)$$

with R_i being an $(m_{\mathbf{y}i} \times m_{\mathbf{y}i})$-dimensional positive definite matrix.

As dynamics of the plant is not considered in state estimations, the temporal variable k has not been included in the above system model.

Define the vectors

$$x = \mathbf{col}\{x_i|_{i=1}^{N}\}, \quad y = \mathbf{col}\{y_i|_{i=1}^{N}\}, \quad w = \mathbf{col}\{w_i|_{i=1}^{N}\}. \tag{6.67}$$

Moreover, define the matrices

$$C = \left[C_{ij}|_{i,j=1}^{N} \right] \quad \text{and} \quad R = \mathbf{diag}\{R_i|_{i=1}^{N}\}. \tag{6.68}$$

Then the relation between the system states and its output measurements can be rewritten as

$$y = Cx + w, \quad w \sim \mathcal{N}(0,\ R). \tag{6.69}$$

Moreover, the maximum likelihood estimate of the plant states using its output measurements can be written as

$$\begin{aligned}\hat{x} &= \arg\min_{x} (y - Cx)^T R^{-1}(y - cx) \\ &= (C^T R^{-1} C)^{-1} C^T R^{-1} y,\end{aligned} \tag{6.70}$$

provided that the matrix C is of full column rank.

This estimate is also called a weighted least squares estimate [19,20].

The rank condition on the matrix C is reasonable. If this condition is not satisfied, then there are infinitely many vectors x satisfying Eq. (6.69), which makes state estimation meaningless in actual engineering. To satisfy this condition, it is necessary that $\sum_{i=1}^{N} m_{\mathbf{x}i} \leq \sum_{i=1}^{N} m_{\mathbf{y}i}$, that is, the number of sensors in the system must not be smaller than the dimension of the system state vector.

Note that the vector $\hat{x}$ in Eq. (6.70) can also be interpreted as the solution to the linear equation

$$\gamma \Pi (C^T R^{-1} C)\hat{x} = \gamma \Pi C^T R^{-1} y, \tag{6.71}$$

which can be further equivalently rewritten as

$$\hat{x} = \left[I - \gamma \Pi (C^T R^{-1} C) \right] \hat{x} + \gamma \Pi C^T R^{-1} y, \tag{6.72}$$

where γ is an arbitrary nonzero real scalar, and Π is an arbitrary invertible real matrix.

Note that the matrix R is block diagonal according to its definition. On the other hand, for a large-scale system, the number of subsystems that directly affect a subsystem is generally not very large, which implies that the matrix C may be sparse in the sense that most of its submatrices C_{ij}, $i, j = 1, 2, \cdots, N$, are in fact equal to a zero matrix. Hence, if the matrix Π also takes a block diagonal form with the dimensions of its submatrices compatible with those of the matrix R, then sparseness of the matrix C can be well preserved by both matrices $I - \gamma \Pi (C^T R^{-1} C)$ and $\Pi C^T R^{-1}$. This is significantly different from the matrix $(C^T R^{-1} C)^{-1}$ and therefore from the matrix $(C^T R^{-1} C)^{-1} C^T R^{-1}$, which are usually dense even if the matrix C is sparse and the matrix R is block diagonal.

On the basis of these observations, a distributed method is suggested in [30] using the so-called Richardson method for the computation of the optimal state estimate $\hat{x}$.

Construct the matrix Π as

$$\Pi = \mathbf{diag}\{\Pi_i|_{i=1}^{N}\}, \quad \Pi_i \in \mathcal{R}^{m_{\mathbf{x}i} \times m_{\mathbf{x}i}}, \quad \Pi > 0, \tag{6.73}$$

and select γ satisfying

$$0 < \gamma < \frac{2}{\sigma_{\max}\left(\Pi^{1/2} C^T R^{-1} C \Pi^{1/2}\right)}. \tag{6.74}$$

Then we can straightforwardly prove that, for each $\Pi_i > 0$, $i = 1, 2, \cdots, N$,

$$\left|\left|I - \gamma \Pi C^T R^{-1} C \Pi\right|\right|_2 = \sigma_{\max}\left(I - \gamma \Pi C^T R^{-1} C \Pi\right) < 1, \tag{6.75}$$

that is, the mapping $f(\cdot) : \mathcal{R}^{m_{\mathbf{x}}} \to \mathcal{R}^{m_{\mathbf{x}}}$ defined by the matrix $I - \gamma \Pi C^T R^{-1} C \Pi$ as $f(z) = I - \gamma \Pi C^T R^{-1} C \Pi z$ is strictly contractive. Here $m_{\mathbf{x}} = \sum_{i=1}^{N} m_{\mathbf{x}i}$.

Accordingly, from the Richardson's method [31] and Eq. (6.72) we can declare that, starting from an arbitrary real vector $\hat{x}^{[0]}$ of a compatible dimension, the iterations

$$\hat{x}^{[k+1]} = \left[I - \gamma \Pi (C^T R^{-1} C)\right] \hat{x}^{[k]} + \gamma \Pi C^T R^{-1} y \tag{6.76}$$

converge to the optimal state estimate $\hat{x}$ with the increment of the iteration index k, that is,

$$\lim_{k \to \infty} \left|\left|\hat{x}^{[k+1]} - \hat{x}\right|\right|_2 = 0 \qquad \text{for all } \hat{x}^{[0]} \in \mathcal{R}^{m_{\mathbf{x}}}.$$

To clarify the distributed computation characteristics of the above iteration, for each $i = 1, 2, \cdots, N$, let $\mathcal{I}_i$ and $\mathcal{O}_i$ denote respectively the set of subsystems whose states directly affect the outputs of the subsystem $\mathbf{\Sigma}_i$ and the set of subsystems whose outputs are directly affected by the states of the subsystem $\mathbf{\Sigma}_i$,[1] that is,

$$\mathcal{I}_i = \left\{j \left| C_{ij} \neq 0\right.\right\}, \quad \mathcal{O}_i = \left\{j \left| C_{ji} \neq 0\right.\right\}.$$

For simplicity of expressions, define the set $\mathcal{N}_i$ for each subsystem $\mathbf{\Sigma}_i$ with $i = 1, 2, \cdots, N$ as

$$\mathcal{N}_i = \mathcal{I}_i \bigcup \mathcal{O}_i.$$

Then the iterations of Eq. (6.76) can be described as follows. In these iterations, each subsystem $\mathbf{\Sigma}_i$ estimates *only* its own state vector utilizing *only* its own output measurements and information from the subsystems with their indices belonging to the set $\mathcal{I}_i$ or the set $\mathcal{O}_i$.

[1] The symbols adopted here for these sets are similar to those of [30] but have completely contrary meanings. The objectives for these content changes are to make them more consistent with the in/out degrees of the graph associated with a networked system.

Algorithm 6.5.1. Distributed Computations of the Estimate $\hat{x}$ (Prototype)

Initialization. For each subsystem $\mathbf{\Sigma}_j$ with $j = 1, 2, \cdots, N$,

- *set $\hat{x}_j^{[0]} = 0$;*
- *compute $\alpha_{jk} = C_{kj}^T R_k^{-1} y_k$ for every $k \in \{1, 2, \cdots, N\}$ satisfying $C_{kj} \neq 0$. Moreover, transfer α_{jk} to each subsystem $\mathbf{\Sigma}_k$ with $k \in \{1, 2, \cdots, N\}$ and satisfying $j \in \mathcal{I}_k$;*
- *compute the vector α_j as*

$$\alpha_j = \sum_{k \in \mathcal{O}_j} \alpha_{jk}.$$

Iterations. For each subsystem $\mathbf{\Sigma}_j$ with $j = 1, 2, \cdots, N$,

- *send its current computed value $\hat{x}_j^{[k]}$, which is an approximation to the estimate $\hat{x}_j$ on its own state vector x_j, to each subsystem $\mathbf{\Sigma}_i$ with index i satisfying $i \in \mathcal{N}_j$;*
- *for each subsystem $\mathbf{\Sigma}_i$ with index i satisfying $i \in \mathcal{I}_j$, compute*

$$\hat{x}_{ij}^{[k]} = \sum_{l \in \mathcal{I}_j} C_{il}^T R_l^{-1} C_{lj} \hat{x}_l^{[k]};$$

- *compute its $(k+1)$th approximation $\hat{x}_j^{[k+1]}$ to $\hat{x}_j$ as*

$$\hat{x}_j^{[k+1]} = \hat{x}_j^{[k+1]} - \gamma \Pi_j \left(\sum_{i \in \mathcal{O}_j} \hat{x}_{il}^{[k]} - \alpha_i \right);$$

- *set $k + 1 \rightarrow k$ and go to the next iteration.*

From the definitions of the sets $\mathcal{I}_i$ and $\mathcal{O}_i$ it is clear that $|\mathcal{I}_i|$ and $|\mathcal{O}_i|$ are respectively the in-degree and out-degree of the subsystem $\mathbf{\Sigma}_i$, $i = 1, 2, \cdots, N$. Under some situations, both of these degrees may obey the so-called power law for a large-scale system [15,32]. This means that for most of the subsystems in a plant constituted from a great amount of subsystems, it is possible that only a few summations are required in each iteration of the above algorithm.

In [30], optimizations have also been investigated on the selection of the scalar parameter γ and the parameter matrices Π_i, $i = 1, 2, \cdots, N$, with the objective of increasing the convergence speed of the above algorithm.

When the system under investigation has some special structure, an algorithm is developed in [30] that is capable of reaching the exact $\hat{x}$ in finite iterations. To illustrate this algorithm, define the set $\mathcal{B}_i$ and the set $\mathcal{K}_{ij}$ for subsystems $\mathbf{\Sigma}_i$ and $\mathbf{\Sigma}_j$ with $i, j = 1, 2, \cdots, N$ respectively as

$$\mathcal{B}_i = \left\{ j \,\middle|\, \mathcal{N}_i \bigcap \mathcal{N}_j \neq \emptyset \right\} \quad \text{and} \quad \mathcal{K}_{ij} = \{ k \,|\, i, j \in \mathcal{I}_k \},$$

that is, $\mathcal{B}_i$ is the set of the indices of subsystems that shares at least one neighbor with subsystem $\mathbf{\Sigma}_i$, whereas $\mathcal{K}_{ij}$ is the set of the indices of the subsystems whose state vector directly and simultaneously affects the output measurements of both subsystems $\mathbf{\Sigma}_i$ and $\mathbf{\Sigma}_j$.

For each $i, j = 1, 2, \cdots, N$, define the matrix

$$\Phi_{ij} = \sum_{k \in \mathcal{K}_{ij}} C_{ki}^T R_k^{-1} C_{kj}.$$

Moreover, define the vectors α_i with $i = 1, 2, \cdots, N$ as in Algorithm 6.5.1. The following algorithm is also suggested in [30] for the calculation of the state estimate $\hat{x}$.

Algorithm 6.5.2. Finite Step Distributed Computations of the Estimate $\hat{x}$ (Prototype)

Initialization. For each subsystem $\mathbf{\Sigma}_i$ with $i = 1, 2, \cdots, N$, set

- $\Sigma_i^{[0]} = \Phi_{ii}^{-1}$, $\hat{x}_i^{[0]} = \Sigma_i^{[0]} \alpha_i$,
- $\Sigma_{ij}^{[0]} = \Sigma_i^{[0]}$, $\hat{x}_{ij}^{[0]} = \hat{x}_i^{[0]}$ *for each* $j \in \mathcal{B}_i \backslash \{i\}$.

Iterations. For each subsystem $\mathbf{\Sigma}_i$ with $i = 1, 2, \cdots, N$ and each $k = 1, 2, \cdots$,

- *for each* $j \in \mathcal{B}_i \backslash \{i\}$,
  - *compute* $\gamma_{ij}^{[k]} = \Phi_{ji} \hat{x}_i^{[k-1]}$ *and* $\Gamma_{ij}^{[k]} = \Phi_{ji} \Sigma_i^{[k-1]} \Phi_{ij}$;
  - *send both* $\gamma_{ij}^{[k]}$ *and* $\Gamma_{ij}^{[k]}$ *to the subsystem* $\mathbf{\Sigma}_j$.
- *at the subsystem* $\mathbf{\Sigma}_i$,
  - *compute sequentially*

$$\Sigma_i^{[k]} = \left[\Phi_{ii} - \sum_{j \in \mathcal{B}_i \backslash \{i\}} \Gamma_{ji}^{[k-1]} \right]^{-1}, \quad \hat{x}_i^{[k]} = \Sigma_i^{[k]} \left[\alpha_i - \sum_{j \in \mathcal{B}_i \backslash \{i\}} \gamma_{ji}^{[k-1]} \right];$$

  - *for each* $j \in \mathcal{B}_i \backslash \{i\}$, *compute sequentially*

$$\Sigma_{ij}^{[k]} = \left[\Phi_{ii} - \sum_{j \in \mathcal{B}_i \backslash \{i,\, j\}} \Gamma_{ji}^{[k-1]} \right]^{-1}, \quad \hat{x}_{ij}^{[k]} = \Sigma_{ij}^{[k]} \left[\alpha_i - \sum_{j \in \mathcal{B}_{i,\, j} \backslash \{i\}} \gamma_{ji}^{[k-1]} \right];$$

- *set* $k + 1 \to k$ *and go to the next iteration.*

Compared with Algorithm 6.5.1, Algorithm 6.5.2 can reach the exact $\hat{x}$ in finitely many iteration steps when the networked system has a special structure. On the other hand, in this algorithm, neither the parameter γ nor the matrix Π is adopted to increase the convergence speed of the iterative computations. In particular, the following conclusions have been established in [30].

Theorem 6.9. *Assume that the graph associated with the networked system $\mathbf{\Sigma}$ described by Eq. (6.66) is acyclic. Then*

1. *All the matrices $\Sigma_i^{[k]}$ and $\Sigma_{ij}^{[k]}$ with $i, j = 1, 2, \cdots, N$ and $k = 0, 1, 2, \cdots$, adopted in Algorithm 6.5.2, are well defined;*
2. *For each $i = 1, 2, \cdots, N$, let ρ_i denote the maximum distance between the subsystem $\mathbf{\Sigma}_i$ and any other subsystem in the networked system $\mathbf{\Sigma}$. Then, for each $k \geq \rho_i$,*

$$\hat{x}_i^{[k]} = \hat{x}_i.$$

It is worth pointing out that for a networked system described by Eqs. (6.1) and (6.2) in which subsystem interactions are realized through their internal input and output vectors, the associated output matrix C in Eq. (6.68) is generally dense, even though the subsystem connection matrix Φ is sparse. This asks further investigations on distributed state estimations for these systems.

6.6 Bibliographic Notes

Distributed estimations are extensively studied in various literatures. Basically, two types of applications exist in this problem. One is to estimate the plant states using a great amount of sensors, which may be distributed in various places that are far from each other, and in which each sensor has its own computation capability and exchange information with its neighbors about its estimate on the plant states [33]. The other one is to use local plant output measurements to estimate the states of a plant subsystem. In this chapter, only the latter one is investigated. Results of this chapter are mainly based on works in [4,5,27]. There are also other methods dealing with this problem, in which different model is adopted for a networked system. For instance, a Jacobi over-relaxation-based method is combined in [2] with dynamic average consensus algorithms under the framework of Bayesian estimations. In addition, augmented Lagrangian formulation and price-decomposition-coordination is used in [34] to develop a distributed estimation algorithm for a large-scale networked system.

Appendix 6.A

6.A.1 Proof of Theorem 6.1

Let $J_{\mathbf{y}i}$, $J_{\mathbf{d}i}$, and $J_{\mathbf{w}i}$, $i = 1, 2, \cdots, N$, denote the matrices defined in the same way as $J_{\mathbf{x}i}$ and $J_{\mathbf{v}i}$ using respectively the dimensions of the vectors $y(k, j)|_{j=1}^{N}$, $d(k, j)|_{j=1}^{N}$, and $w(k, j)|_{j=1}^{N}$.

Note that, for arbitrary $i = 1, 2, \cdots, N$, we have

$$\begin{aligned}
J_{\mathbf{x}i}^T\,[-K_{\mathbf{x}}(k)\;\; I] &= \Big[0 \cdots 0\, I_{m_{\mathbf{x}i}}\, 0 \cdots 0\Big]\Big[-\mathbf{diag}\Big\{K_{\mathbf{x}}(k,i)|_{i=1}^{N}\Big\}\;\; \mathbf{diag}\Big\{I_{m_{\mathbf{x}i}}|_{i=1}^{N}\Big\}\Big] \\
&= \Big[\; [0 \cdots 0 \,-\, K_{\mathbf{x}}(k,i)\, 0 \cdots 0] \;\; \Big[0 \cdots 0\, I_{m_{\mathbf{x}i}}\, 0 \cdots 0\Big] \;\Big] \\
&= \Big[-K_{\mathbf{x}}(k,i) J_{\mathbf{y}i}^T \;\; J_{\mathbf{x}i}^T\Big] \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} J_{\mathbf{y}i}^T & 0 \\ 0 & J_{\mathbf{x}i}^T \end{bmatrix}. \qquad (6.\text{A}.1)
\end{aligned}$$

Hence,

$$\begin{aligned}
& J_{\mathbf{x}i}^T\,[-K_{\mathbf{x}}(k)\;\; I]\begin{bmatrix} C(k) \\ A(k) \end{bmatrix} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} J_{\mathbf{y}i}^T & 0 \\ 0 & J_{\mathbf{x}i}^T \end{bmatrix}\left\{\begin{bmatrix} C_{\mathbf{x}}(k) \\ A_{\mathbf{x}\mathbf{x}}(k) \end{bmatrix} + \begin{bmatrix} C_{\mathbf{v}}(k) \\ A_{\mathbf{x}\mathbf{v}}(k) \end{bmatrix}\Phi(k)\,[\, I - A_{\mathbf{z}\mathbf{v}}(k)\Phi(k)\,]^{-1}\, A_{\mathbf{z}\mathbf{x}}(k)\right\} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\left\{\begin{bmatrix} J_{\mathbf{y}i}^T C_{\mathbf{x}}(k) \\ J_{\mathbf{x}i}^T A_{\mathbf{x}\mathbf{x}}(k) \end{bmatrix} + \begin{bmatrix} J_{\mathbf{y}i}^T C_{\mathbf{v}}(k) \\ J_{\mathbf{x}i}^T A_{\mathbf{x}\mathbf{v}}(k) \end{bmatrix}\Phi(k)\,[\, I - A_{\mathbf{z}\mathbf{v}}(k)\Phi(k)\,]^{-1}\, A_{\mathbf{z}\mathbf{x}}(k)\right\} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\left\{\begin{bmatrix} C_{\mathbf{x}}(k,i) \\ A_{\mathbf{x}\mathbf{x}}(k,i) \end{bmatrix} J_{\mathbf{x}i}^T + \begin{bmatrix} C_{\mathbf{v}}(k,i) \\ A_{\mathbf{x}\mathbf{v}}(k,i) \end{bmatrix} J_{\mathbf{v}i}^T \Phi(k)\,[\, I - A_{\mathbf{z}\mathbf{v}}(k)\Phi(k)\,]^{-1}\, A_{\mathbf{z}\mathbf{x}}(k)\right\} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} C_{\mathbf{x}}(k,i) & C_{\mathbf{v}}(k,i) \\ A_{\mathbf{x}\mathbf{x}}(k,i) & A_{\mathbf{x}\mathbf{v}}(k,i) \end{bmatrix}\begin{bmatrix} J_{\mathbf{x}i}^T \\ J_{\mathbf{v}i}^T \Phi(k)\,[\, I - A_{\mathbf{z}\mathbf{v}}(k)\Phi(k)\,]^{-1}\, A_{\mathbf{z}\mathbf{x}}(k) \end{bmatrix} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} C(k,i) \\ A_{\mathbf{x}}(k,i) \end{bmatrix} W(k,i). \qquad (6.\text{A}.2)
\end{aligned}$$

Moreover,

$$\begin{aligned}
J_{\mathbf{x}i}^T\,[-K_{\mathbf{x}}(k)\;\; I]\begin{bmatrix} D(k) & 0 \\ 0 & B_{\mathbf{x}}(k) \end{bmatrix} &= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} J_{\mathbf{y}i}^T & 0 \\ 0 & J_{\mathbf{x}i}^T \end{bmatrix}\begin{bmatrix} D(k) & 0 \\ 0 & B_{\mathbf{x}}(k) \end{bmatrix} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} D(k,i) J_{\mathbf{w}i}^T & 0 \\ 0 & B_{\mathbf{x}}(k,i) J_{\mathbf{d}i}^T \end{bmatrix} \\
&= \Big[-K_{\mathbf{x}}(k,i)\;\; I_{m_{\mathbf{x}i}}\Big]\begin{bmatrix} D(k,i) & 0 \\ 0 & B_{\mathbf{x}}(k,i) \end{bmatrix}\begin{bmatrix} J_{\mathbf{w}i}^T & 0 \\ 0 & J_{\mathbf{d}i}^T \end{bmatrix}. \\
& \qquad (6.\text{A}.3)
\end{aligned}$$

On the other hand, from Eqs. (6.9) and (6.14) we obtain the following relation:

$$\begin{aligned}
P_{ii}(k+1) &= J_{\mathbf{x}i}^T \left\{ [-K_{\mathbf{x}}(k)\ \ I] \left(\begin{bmatrix} C(k) \\ A(k) \end{bmatrix} P(k) \begin{bmatrix} C(k) \\ A(k) \end{bmatrix}^T \right.\right. \\
&\quad \left.\left. + \begin{bmatrix} D(k)D^T(k) & 0 \\ 0 & B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k) \end{bmatrix} \right) \begin{bmatrix} -K_{\mathbf{x}}^T(k) \\ I \end{bmatrix} \right\} J_{\mathbf{x}i} \\
&= \left(J_{\mathbf{x}i}^T [-K_{\mathbf{x}}(k)\ \ I] \begin{bmatrix} C(k) \\ A(k) \end{bmatrix} \right) P(k)\, (\star)^T \\
&\quad + \left(J_{\mathbf{x}i}^T [-K_{\mathbf{x}}(k)\ \ I] \begin{bmatrix} D(k) & 0 \\ 0 & B_{\mathbf{x}}(k) \end{bmatrix} \right) (\star)^T . \qquad (6.A.4)
\end{aligned}$$

Substitute Eqs. (6.A.2) and (6.A.3) into Eq. (6.A.4). From the definitions of the matrices $A_{\mathbf{x}}(k,i)$, $C(k,i)$, and $W(k,i)$ we can further prove that

$$\begin{aligned}
P_{ii}(k+1) &= [-K_{\mathbf{x}}(k,i)\ \ I_{m_{\mathbf{x}i}}] \left\{ \left(\begin{bmatrix} C(k,i) \\ A_{\mathbf{x}}(k,i) \end{bmatrix} W(k,i) \right) P(k)\, (\star)^T \right. \\
&\quad \left. + \begin{bmatrix} D(k,i)D^T(k,i) & 0 \\ 0 & B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i) \end{bmatrix} \right\} \begin{bmatrix} -K_{\mathbf{x}}^T(k,i) \\ I_{m_{\mathbf{x}i}} \end{bmatrix} \\
&= [-K_{\mathbf{x}}(k,i)\ \ I_{m_{\mathbf{x}i}}] \left[\begin{matrix} C(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i) + D(k,i)D^T(k,i) \\ A_{\mathbf{x}}(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i) \end{matrix} \right. \\
&\qquad \left. \begin{matrix} C(k,i)W(k,i)P(k)W^T(k,i)A_{\mathbf{x}}^T(k,i) \\ A_{\mathbf{x}}(k,i)W(k,i)P(k)W^T(k,i)A_{\mathbf{x}}^T(k,i) + B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i) \end{matrix} \right] \\
&\quad \times \begin{bmatrix} -K_{\mathbf{x}}^T(k,i) \\ I_{m_{\mathbf{x}i}} \end{bmatrix}. \qquad (6.A.5)
\end{aligned}$$

From the positive semidefiniteness of the matrix $P(k)$ and the assumption about the regularity of the matrix $D(k,i)D^T(k,i)$ we can directly declare that the matrix $C(k,i)W(k,i)P(k) \times W^T(k,i)C^T(k,i) + D(k,i)D^T(k,i)$ is positive definite.

On the basis of Lemma 2.3 and Eq. (6.A.5), we can claim that the optimal $K_{\mathbf{x}}(k,i)$, denoted by $K_{\mathbf{x}}^{\text{opt}}(k,i)$, that minimizes $P_{ii}(k+1)$ is unique and can be expressed as

$$\begin{aligned}
K_{\mathbf{x}}^{\text{opt}}(k,i) &= A_{\mathbf{x}}(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i) \\
&\quad \times \left\{ C(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i) + D(k,i)D^T(k,i) \right\}^{-1} \\
&= A_{\mathbf{x}}(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i) \left[D(k,i)D^T(k,i) \right]^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\times \left\{C(k,i)W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}+I\right\}^{-1}\\
= \; & A_{\mathbf{x}}(k,i)\left\{I+W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1}\\
&\times W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}. \qquad (6.A.6)
\end{aligned}
$$

This completes the proof. □

6.A.2 Proof of Theorem 6.2

To simplify mathematical expressions, denote the matrix $W(k,i)P(k)W^T(k,i)C^T(k,i)\times\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)$ by $Z(k,i)$. From the definitions of the matrices $A(k)$, $C(k)$, and $K_{\mathbf{x}}(k)$ it is clear that

$$
\begin{aligned}
A(k)-K_{\mathbf{x}}(k)C(k) &= \mathbf{diag}\{A_{\mathbf{Tx}}(k,i)-K_{\mathbf{x}}(k,i)C_{\mathbf{x}}(k,i)\}+\mathbf{diag}\{A_{\mathbf{xv}}(k,i)-K_{\mathbf{x}}(k,i)C_{\mathbf{v}}(k,i)\}\\
&\quad\times\Phi(k)\,[\,I-A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}A_{\mathbf{zx}}(k)\\
&= \mathbf{col}\left\{[A_{\mathbf{Tx}}(k,i)-K_{\mathbf{x}}(k,i)C_{\mathbf{x}}(k,i)]J_{\mathbf{x}i}^T|_{i=1}^N\right\}\\
&\quad+\mathbf{col}\left\{[A_{\mathbf{xv}}(k,i)-K_{\mathbf{x}}(k,i)C_{\mathbf{v}}(k,i)]J_{\mathbf{v}i}^T|_{i=1}^N\right\}\\
&\quad\times\Phi(k)\,[\,I-A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}A_{\mathbf{zx}}(k)\\
&= \mathbf{col}\left\{[-K_{\mathbf{x}}(k,i)\;\; I]\begin{bmatrix} C_{\mathbf{x}}(k,i) & C_{\mathbf{v}}(k,i)\\ A_{\mathbf{Tx}}(k,i) & A_{\mathbf{xv}}(k,i)\end{bmatrix}\right.\\
&\quad\left.\times\begin{bmatrix} J_{\mathbf{x}i}^T\\ J_{\mathbf{v}i}^T\Phi(k)\,[I-A_{\mathbf{zv}}(k)\Phi(k)]^{-1}A_{\mathbf{zx}}(k)\end{bmatrix}\Bigg|_{i=1}^N\right\}\\
&= \mathbf{col}\left\{[-K_{\mathbf{x}}(k,i)\;\; I]\begin{bmatrix} C(k,i)\\ A_{\mathbf{x}}(k,i)\end{bmatrix}W(k,i)\Bigg|_{i=1}^N\right\}. \qquad (6.A.7)
\end{aligned}
$$

On the other hand, straightforwardly from the expression of the optimal gain matrix $K_{\mathbf{x}}^{\text{opt}}(k,i)$, we have

$$
\begin{aligned}
[-K_{\mathbf{x}}^{\text{opt}}(k,i)\;\; I]\begin{bmatrix} C(k,i)\\ A_{\mathbf{x}}(k,i)\end{bmatrix} &= A_{\mathbf{x}}(k,i)-K_{\mathbf{x}}^{\text{opt}}(k,i)C(k,i)\\
&= A_{\mathbf{x}}(k,i)-A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}Z(k,i)\\
&= A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}. \qquad (6.A.8)
\end{aligned}
$$

Denote the matrix $\mathbf{diag}\{K_{\mathbf{x}}^{\mathrm{opt}}(k,i)|_{i=1}^{N}\}$ by $K_{\mathbf{x}}^{\mathrm{opt}}(k)$. Then, from the last two equations we further get

$$\begin{aligned}
&[A(k)-K_{\mathbf{x}}^{\mathrm{opt}}(k)C(k)]P(k)[A(k)-K_{\mathbf{x}}^{\mathrm{opt}}(k)C(k)]^{T}\\
={}& \left(\mathbf{col}\left\{[-K_{\mathbf{x}}^{\mathrm{opt}}(k,i)\;\; I]\begin{bmatrix} C(k,i)\\ A_{\mathbf{x}}(k,i)\end{bmatrix}W(k,i)\bigg|_{i=1}^{N}\right\}\right)P(k)\\
&\times\left(\mathbf{col}\left\{[-K_{\mathbf{x}}^{\mathrm{opt}}(k,i)\;\; I]\begin{bmatrix} C(k,i)\\ A_{\mathbf{x}}(k,i)\end{bmatrix}W(k,i)\bigg|_{i=1}^{N}\right\}\right)^{T}\\
={}& \left(\mathbf{col}\left\{A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)\Big|_{i=1}^{N}\right\}\right)P(k)\\
&\times\left(\mathbf{col}\left\{A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)\Big|_{i=1}^{N}\right\}\right)^{T}\\
={}& \left\{A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)P(k)W^{T}(k,j)\left[I+Z^{T}(k,j)\right]^{-1}A_{\mathbf{x}}^{T}(k,j)\bigg|_{i,j=1}^{N}\right\}.
\end{aligned}\tag{6.A.9}$$

Note that

$$K_{\mathbf{x}}(k)D(k)D^{T}(k)K_{\mathbf{x}}^{T}(k)=\mathbf{diag}\left\{K_{\mathbf{x}}(k,i)D(k,i)D^{T}(k,i)K_{\mathbf{x}}^{T}(k,i)|_{i=1}^{N}\right\},\tag{6.A.10}$$

$$B_{\mathbf{x}}(k)B_{\mathbf{x}}^{T}(k)=\mathbf{diag}\left\{B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^{T}(k,i)|_{i=1}^{N}\right\}.\tag{6.A.11}$$

Moreover,

$$\begin{aligned}
&K_{\mathbf{x}}^{\mathrm{opt}}(k,i)D(k,i)D^{T}(k,i)(K_{\mathbf{x}}^{\mathrm{opt}}(k,i))^{T}\\
={}& \left(A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)P(k)W^{T}(k,i)C^{T}(k,i)\left[D(k,i)D^{T}(k,i)\right]^{-1}\right)\\
&\times D(k,i)D^{T}(k,i)\,(\star)^{T}\\
={}& A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)P(k)W^{T}(k,i)Z^{T}(k,i)\left[I+Z^{T}(k,i)\right]^{-1}A_{\mathbf{x}}^{T}(k,i)\\
={}& A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)P(k)W^{T}(k,i)A_{\mathbf{x}}^{T}(k,i)\\
&-A_{\mathbf{x}}(k,i)\,[I+Z(k,i)]^{-1}\,W(k,i)P(k)W^{T}(k,i)\left[I+Z^{T}(k,i)\right]^{-1}A_{\mathbf{x}}^{T}(k,i).
\end{aligned}\tag{6.A.12}$$

Recall that $P_{ij}(k+1)=J_{\mathbf{x}i}^{T}P(k+1)J_{\mathbf{x}j}$ for $i,j\in\{1,2,\cdots,N\}$. The proof can now be completed by multiplying both sides of Eq. (6.9) from their left sides and right sides respectively by the matrices $J_{\mathbf{x}i}^{T}$ and $J_{\mathbf{x}j}$, replacing the gain matrix $K_{\mathbf{x}}(k)$ by its optimal value $K_{\mathbf{x}}^{\mathrm{opt}}(k)$, and substituting Eqs. (6.A.9)–(6.A.12) into the right side of this equation. □

6.A.3 Proof of Theorem 6.3

To simplify mathematical expressions, denote the matrices $W^T(k,i)C^T(k,i)[D(k,i)\times D^T(k,i)]^{-1}C(k,i)W(k,i)$ and $A_{\mathbf{x}}(k,i)W(k,i)$ respectively by $Q(k,i)$ and $S(k,i)$. Then, it is obvious that the matrix $Q(k,i)$ is always symmetric, that is, $Q^T(k,i)=Q(k,i)$. On the other hand, from Theorem 6.2 we have that, for arbitrary $i\in\{1,2,\cdots,N\}$,

$$\begin{aligned}
P_{ii}(k+1) &= A_{\mathbf{x}}(k,i)\left\{I+W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1} \\
&\quad \times W(k,i)P(k)W^T(k,i)A_{\mathbf{x}}^T(k,i)+B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i) \\
&= [A_{\mathbf{x}}(k,i)W(k,i)]\left\{I+P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}\right. \\
&\quad \left.\times\, C(k,i)W(k,i)\right\}^{-1}P(k)[A_{\mathbf{x}}(k,i)W(k,i)]^T+B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i) \\
&= S(k,i)[I+P(k)Q(k,i)]^{-1}P(k)S^T(k,i)+B_{\mathbf{x}}(k,i)B_{\mathbf{x}}^T(k,i).
\end{aligned} \tag{6.A.13}$$

In addition, for all $i,j=1,2,\cdots,N$ with $i\neq j$,

$$\begin{aligned}
P_{ij}(k+1) &= A_{\mathbf{x}}(k,i)\left\{I+W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1} \\
&\quad \times W(k,i)P(k)W^T(k,j)\left\{I+C^T(k,j)\left[D(k,j)D^T(k,j)\right]^{-1}\right. \\
&\quad \left.\times\, C(k,j)W(k,j)P(k)W^T(k,j)\right\}^{-1}A_{\mathbf{x}}^T(k,j) \\
&= [A_{\mathbf{x}}(k,i)W(k,i)]\left\{I+P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}\right. \\
&\quad \left.\times\, C(k,i)W(k,i)\right\}^{-1}P(k)\left\{I+W^T(k,j)C^T(k,j)\left[D(k,j)D^T(k,j)\right]^{-1}\right. \\
&\quad \left.\times\, C(k,j)W(k,j)P(k)\right\}^{-1}[A_{\mathbf{x}}(k,j)W(k,j)]^T \\
&= S(k,i)[I+P(k)Q(k,i)]^{-1}P(k)[I+Q(k,j)P(k)]^{-1}S^T(k,j).
\end{aligned} \tag{6.A.14}$$

From the definitions of the matrices $A_{\mathbf{x}}(k,i)$, $W(k,i)$, and $C(k,i)$ we have that

$$\begin{aligned}
S(k,i) &= [A_{\mathbf{xx}}(k,i)\ \ A_{\mathbf{xv}}(k,i)]\begin{bmatrix} J_{\mathbf{x}i}^T \\ J_{\mathbf{v}i}^T\Phi(k)\,[\,I-A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}A_{\mathbf{zx}}(k)\end{bmatrix} \\
&= J_{\mathbf{x}i}^T A_{\mathbf{xx}}(k)+J_{\mathbf{x}i}^T A_{\mathbf{xv}}(k)\Phi(k)\,[\,I-A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}A_{\mathbf{zx}}(k)
\end{aligned}$$

$$\begin{aligned}
&= J_{\mathbf{x}i}^T\{A_{\mathbf{xx}}(k) + A_{\mathbf{xv}}(k)\Phi(k)\,[\,I - A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}\,A_{\mathbf{zx}}(k)\} \\
&= A_i(k) \qquad (6.A.15)
\end{aligned}$$

and

$$\begin{aligned}
C(k,i)W(k,i) &= [C_{\mathbf{x}}(k,i)\;\; C_{\mathbf{v}}(k,i)]\begin{bmatrix} J_{\mathbf{x}i}^T \\ J_{\mathbf{v}i}^T\Phi(k)\,[\,I - A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}\,A_{\mathbf{zx}}(k) \end{bmatrix} \\
&= J_{\mathbf{y}i}^T C_{\mathbf{x}}(k) + J_{\mathbf{y}i}^T C_{\mathbf{v}}(k)\Phi(k)\,[\,I - A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}\,A_{\mathbf{zx}}(k) \\
&= J_{\mathbf{y}i}^T\{C_{\mathbf{x}}(k) + C_{\mathbf{v}}(k)\Phi(k)\,[\,I - A_{\mathbf{zv}}(k)\Phi(k)\,]^{-1}\,A_{\mathbf{zx}}(k)\} \\
&= J_{\mathbf{y}i}^T C(k). \qquad (6.A.16)
\end{aligned}$$

Hence

$$\begin{aligned}
[D(k,i)D^T(k,i)]^{-1/2}C(k,i)W(k,i) &= [D(k,i)D^T(k,i)]^{-1/2}\{J_{\mathbf{y}i}^T C(k)\} \\
&= J_{\mathbf{y}i}^T[D(k)D^T(k)]^{-1/2}C(k) \\
&= C_i(k). \qquad (6.A.17)
\end{aligned}$$

Moreover,

$$\begin{aligned}
Q(k,i) &= \left\{[D(k,i)D^T(k,i)]^{-1/2}C(k,i)W(k,i)\right\}^T\left\{[D(k,i)D^T(k,i)]^{-1/2}C(k,i)W(k,i)\right\} \\
&= C_i^T(k)C_i(k). \qquad (6.A.18)
\end{aligned}$$

Substituting Eqs. (6.A.15) and (6.A.18) into Eqs. (6.A.13) and (6.A.14), we obtain the desirable expression for $P_{ij}(k+1)$, $i, j = 1, 2, \cdots, N$.

On the other hand, from Theorem 6.1 we have that, for every $i = 1, 2, \cdots, N$,

$$\begin{aligned}
K_{\mathbf{x}}^{\text{opt}}(k,i) &= A_{\mathbf{x}}(k,i)\left\{I + W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}C(k,i)\right\}^{-1} \\
&\quad \times W(k,i)P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1} \\
&= [A_{\mathbf{x}}(k,i)W(k,i)]\left\{I + P(k)W^T(k,i)C^T(k,i)\left[D(k,i)D^T(k,i)\right]^{-1}\right. \\
&\quad \left.\times\, C(k,i)W(k,i)\right\}^{-1} P(k)\left\{\left[D(k,i)D^T(k,i)\right]^{-1/2}C(k,i)W(k,i)\right\}^T \\
&\quad \times\left[D(k,i)D^T(k,i)\right]^{-1/2}
\end{aligned}$$

$$= A_i(k)[P^{-1}(k) + C_i^T(k)C_i(k)]^{-1}C_i^T(k)\left[D(k,i)D^T(k,i)\right]^{-1/2}. \tag{6.A.19}$$

This completes the proof. □

6.A.4 Proof of Theorem 6.4

For brevity, let $\lambda(k,i)$ and $R(k,i)$ represent respectively the number $(1-\gamma(k,i))/\gamma(k,i)$ and the matrix $D(k,i)D^T(k,i)$. Note that, for an arbitrary time instant $k = 0, 1, \dots,$

$$\frac{\partial(y(k,i) - \bar{C}_i(k,\varepsilon(k))x(k))}{\partial\varepsilon_k(k,j)} = -\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_k(k,j)}x(k), \quad k = 1, 2, \cdots, m_j,\ j = 1, 2, \cdots, N, \tag{6.A.20}$$

$$\frac{\partial(y(k,i) - \bar{C}_i(k,\varepsilon(k))x(k))}{\partial\varepsilon_{k,\phi}(k)} = -\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_{k,\phi}(k)}x(k), \quad k = 1, 2, \cdots, m_\phi. \tag{6.A.21}$$

We can therefore declare that

$$\begin{aligned}
\bar{J}(x(k),d(k)) &= \gamma(k,i)J(x(k),d(k)) + \frac{1-\gamma(k,i)}{2}\sum_{j=1}^{N}\sum_{k=1}^{m_j}\left(\left.\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_k(k,j)}\right|_{\varepsilon(k)=0} \times x(k)\right)^T \\
&\quad \times \left(\left.\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_k(k,j)}\right|_{\varepsilon(k)=0} \times x(k)\right) \\
&\quad + \frac{1-\gamma(k,i)}{2}\sum_{k=1}^{m_\phi}\left(\left.\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_{k,\phi}(k)}\right|_{\varepsilon(k)=0} \times x(k)\right)^T \\
&\quad \times \left(\left.\frac{\partial\bar{C}_i(k,\varepsilon(k))}{\partial\varepsilon_{k,\phi}(k)}\right|_{\varepsilon(k)=0} \times x(k)\right) \\
&= \gamma(k,i)\left\{J(x(k),d(k)) + \frac{\lambda(k,i)}{2}x^T(k)H^T(k,i)H(k,i)x(k)\right\}.
\end{aligned} \tag{6.A.22}$$

Hence

$$\begin{aligned}
&\frac{\partial\bar{J}(x(k),d(k))}{\partial x(k)} \\
&= \gamma(k,i)\Big\{(P_i^{\text{rob}}(k))^{-1}[x(k) - \tilde{\hat{x}}(k)] - \bar{C}_i^T(k)R^{-1}(k,i)[y(k,i) - \bar{C}_i(k)x(k)] \\
&\qquad + \lambda(k,i)H^T(k,i)H(k,i)x(k)\Big\} \\
&= \gamma(k,i)\Big\{\Big[(P_i^{\text{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k) + \lambda(k,i)H^T(k,i)H(k,i)\Big]x(k) \\
&\qquad - (P_i^{\text{rob}}(k))^{-1}\tilde{\hat{x}}(k) - \bar{C}_i^T(k)R^{-1}(k,i)y(k,i)\Big\},
\end{aligned} \tag{6.A.23}$$

$$\frac{\partial \bar{J}(x(k), d(k))}{\partial d(k)} = \gamma(k,i)d(k). \tag{6.A.24}$$

Let $\tilde{x}^{\mathrm{opt}}(k)$ and $\tilde{d}^{\mathrm{opt}}(k)$ denote respectively the optimal $x(k)$ and $d(k)$ that minimize the cost function $\bar{J}(x(k), d(k))$ defined in Eq. (6.34). Then, from the first-order derivative condition on optimums and the assumption $\gamma(k,i) > 0$ we have that

$$\begin{aligned}
\tilde{x}^{\mathrm{opt}}(k) = {} & \left\{(P_i^{\mathrm{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k) + \lambda(k,i)H^T(k,i)H(k,i)\right\}^{-1} \\
& \times \left\{(P_i^{\mathrm{rob}}(k))^{-1}\tilde{\hat{x}}(k) + \bar{C}_i^T(k)R^{-1}(k,i)y(k,i)\right\} \\
= {} & \left\{(P_i^{\mathrm{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k) + \lambda(k,i)H^T(k,i)H(k,i)\right\}^{-1} \\
& \times \Big\{[(P_i^{\mathrm{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k)]\tilde{\hat{x}}(k) + \bar{C}_i^T(k)R^{-1}(k,i) \\
& \quad \times [y(k,i) - \bar{C}_i(k)\tilde{\hat{x}}(k)]\Big\},
\end{aligned} \tag{6.A.25}$$

$$\tilde{d}^{\mathrm{opt}}(k) = 0. \tag{6.A.26}$$

Hence, according to the weighted least squares interpretation of the Kalman filter, the optimal estimate $\tilde{\hat{x}}(k+1)$ can be expressed as

$$\begin{aligned}
\tilde{\hat{x}}(k+1) &= A(k)\tilde{x}^{\mathrm{opt}}(k) + B_{\mathbf{x}}(k)\tilde{d}^{\mathrm{opt}}(k) \\
&= A(k)\left\{(P_i^{\mathrm{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k) + \lambda(k,i)H^T(k,i)H(k,i)\right\}^{-1} \\
&\quad \times \Big\{[(P_i^{\mathrm{rob}}(k))^{-1} + \bar{C}_i^T(k)R^{-1}(k,i)\bar{C}_i(k)]\tilde{\hat{x}}(k) + \bar{C}_i^T(k)R^{-1}(k,i) \\
&\quad \times [y(k,i) - \bar{C}_i(k)\tilde{\hat{x}}(k)]\Big\} \\
&= \hat{A}(k)\tilde{\hat{x}}(k) + K_i^{\mathrm{rob}}(k)[y(k,i) - \bar{C}_i(k)\tilde{\hat{x}}(k)].
\end{aligned} \tag{6.A.27}$$

To derive the desirable expression for the pseudo-covariance matrix $P_i^{\mathrm{rob}}(k+1)$, consider the state prediction errors of the following dynamic system $\tilde{\mathbf{\Sigma}}_i$:

$$x(k+1) = \hat{A}(k)x(k) + B_{\mathbf{x}}(k)d(k), \qquad y(k,i) = \bar{C}_i(k)x(k) + D(k,i)w(k,i). \tag{6.A.28}$$

Then, clearly,

$$\begin{aligned}
& \tilde{\hat{x}}(k+1) - x(k+1) \\
&= \{\hat{A}(k)\tilde{\hat{x}}(k) + K_i^{\mathrm{rob}}(k)[\bar{C}_i(k)x(k) + D(k,i)w(k,i) - \bar{C}_i(k)\tilde{\hat{x}}(k)]\} \\
&\quad - \{\hat{A}(k)x(k) + B_{\mathbf{x}}(k)d(k)\}
\end{aligned}$$

$$= [\hat{A}(k) - K_i^{\text{rob}}(k)\bar{C}_i(k)][\tilde{\hat{x}}(k) - x(k)] + K_i^{\text{rob}}(k)D(k,i)w(k,i) - B_{\mathbf{x}}(k)d(k). \tag{6.A.29}$$

From this relation and the assumptions on external disturbances and measurement errors, straightforward but tedious algebraic operations show that

$$\begin{aligned} P_i^{\text{rob}}(k+1) &= \mathbf{E}\{[\tilde{\hat{x}}(k+1) - x(k+1)][\tilde{\hat{x}}(k+1) - x(k+1)]^T\} \\ &= [\hat{A}(k) - K_i^{\text{rob}}(k)\bar{C}_i(k)]\mathbf{E}\{[\tilde{\hat{x}}(k) - x(k)][\tilde{\hat{x}}(k) - x(k)]^T\}[\hat{A}(k) - K_i^{\text{rob}}(k)\bar{C}_i(k)]^T \\ &\quad + [K_i^{\text{rob}}(k)D(k,i)]\mathbf{E}\{w(k,i)w^T(k,i)\}[K_i^{\text{rob}}(k)D(k,i)]^T \\ &\quad + B_{\mathbf{x}}(k)\mathbf{E}\{d(k)d^T(k)\}B_{\mathbf{x}}^T(k) \\ &= \hat{A}(k)[(P_i^{\text{rob}}(k))^{-1} + C_i^T(k)C_i(k)]^{-1}\hat{A}^T(k) + B_{\mathbf{x}}(k)B_{\mathbf{x}}^T(k). \end{aligned} \tag{6.A.30}$$

This completes the proof. □

6.A.5 Derivation of Eqs. (6.46) and (6.47)

Define the vectors $\bar{z}(k) = \mathbf{col}\left\{\hat{z}(k,i)|_{i=1}^{N}\right\}$ and $\bar{v}(k) = \mathbf{col}\left\{\hat{v}(k,i)|_{i=1}^{N}\right\}$. Then

$$\hat{v}(k) = T_v\bar{v}(k), \quad \hat{z}(k) = T_z\bar{z}(k), \tag{6.A.31}$$

where for each $\star = v$ or $\star = z$,

$$T_\star = \begin{bmatrix} I_{m_{\star 1}} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & I_{m_{\star 2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I_{m_{\star N}} & 0 \\ 0 & I_{m_{\star 1}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & I_{m_{\star N}} \end{bmatrix}.$$

Obviously, both T_v and T_z are invertible.

From Eqs. (6.40) and (6.A.31) it is clear that

$$\bar{v}(k) = T_v^{-1}\bar{\Phi}(k)T_z\bar{z}(k). \tag{6.A.32}$$

Based on these relations and Eq. (6.42), we can prove through direct matrix manipulations that

$$\begin{bmatrix} \hat{x}(k+1) \\ \hat{y}(k+1) \end{bmatrix} = \left\{ \begin{bmatrix} A_{\mathbf{xx}}(k) & K_{\mathbf{x}}(k) \\ \tilde{C}_{\mathbf{x}}(k) & 0 \end{bmatrix} + \begin{bmatrix} \tilde{A}_{\mathbf{xv}}(k) \\ \tilde{C}_{\mathbf{v}}(k) \end{bmatrix} \tilde{\Phi}(k)\left[I - \tilde{A}_{\mathbf{zv}}(k)\tilde{\Phi}(k)\right]^{-1}\left[\tilde{A}_{\mathbf{zx}}(k) \ \ 0\right] \right\}$$

$$\times \begin{bmatrix} \hat{x}(k) \\ y(k+1) - \hat{y}(k+1) \end{bmatrix}, \tag{6.A.33}$$

where $\tilde{\Phi}(k) = T_v^{-1}\bar{\Phi}(k)T_z$.

Define the matrices

$$\bar{A}_{\mathbf{zx}}(k) = \begin{bmatrix} A_{\mathbf{zx}}(k+1)A_{\mathbf{Tx}}(k) \\ A_{\mathbf{zx}}(k) \end{bmatrix}, \quad \bar{A}_{\mathbf{zv}}(k) = \begin{bmatrix} A_{\mathbf{zv}}(k+1) & A_{\mathbf{zx}}(k+1)A_{\mathbf{xv}}(k) \\ 0_{m_{\mathbf{z}} \times m_{\mathbf{v}}} & A_{\mathbf{zv}}(k) \end{bmatrix},$$
$$\bar{A}_{\mathbf{xv}}(k) = \begin{bmatrix} 0_{m_{\mathbf{x}} \times m_{\mathbf{v}}} & A_{\mathbf{xv}}(k) \end{bmatrix}, \quad \bar{C}_{\mathbf{v}}(k) = [C_{\mathbf{v}}(k+1) \;\; C_{\mathbf{x}}(k+1)A_{\mathbf{xv}}(k)].$$

Then direct algebraic manipulations show that

$$\tilde{A}_{\mathbf{xv}}(k) = \bar{A}_{\mathbf{xv}}(k)T_v, \;\; T_z\tilde{A}_{\mathbf{zv}}(k) = \bar{A}_{\mathbf{zv}}(k)T_v, \;\; T_z\tilde{A}_{\mathbf{zx}}(k) = \bar{A}_{\mathbf{zx}}(k), \;\; \tilde{C}_{\mathbf{v}}(k) = \bar{C}_{\mathbf{v}}(k)T_v. \tag{6.A.34}$$

Therefore

$$\begin{aligned} & \tilde{A}_{\mathbf{xv}}(k)\tilde{\Phi}(k)\left[I - \tilde{A}_{\mathbf{zv}}(k)\tilde{\Phi}(k)\right]^{-1}\tilde{A}_{\mathbf{zx}}(k) \\ &= \tilde{A}_{\mathbf{xv}}(k)T_v^{-1}\bar{\Phi}(k)T_z\left[I - \tilde{A}_{\mathbf{zv}}(k)T_v^{-1}\bar{\Phi}(k)T_z\right]^{-1}\tilde{A}_{\mathbf{zx}}(k) \\ &= \tilde{A}_{\mathbf{xv}}(k)T_v^{-1}\bar{\Phi}(k)\left[I - T_z\tilde{A}_{\mathbf{zv}}(k)T_v^{-1}\bar{\Phi}(k)\right]^{-1}T_z\tilde{A}_{\mathbf{zx}}(k) \\ &= \bar{A}_{\mathbf{xv}}(k)\bar{\Phi}(k)\left[I - \bar{A}_{\mathbf{zv}}(k)\bar{\Phi}(k)\right]^{-1}\bar{A}_{\mathbf{zx}}(k), \end{aligned} \tag{6.A.35}$$

$$\begin{aligned} & \tilde{C}_{\mathbf{v}}(k)\tilde{\Phi}(k)\left[I - \tilde{A}_{\mathbf{zv}}(k)\tilde{\Phi}(k)\right]^{-1}\tilde{A}_{\mathbf{zx}}(k) \\ &= \tilde{C}_{\mathbf{v}}(k)T_v^{-1}\bar{\Phi}(k)T_z\left[I - \tilde{A}_{\mathbf{zv}}(k)T_v^{-1}\bar{\Phi}(k)T_z\right]^{-1}\tilde{A}_{\mathbf{zx}}(k) \\ &= \bar{C}_{\mathbf{v}}(k)\bar{\Phi}(k)\left[I - \bar{A}_{\mathbf{zv}}(k)\bar{\Phi}(k)\right]^{-1}\bar{A}_{\mathbf{zx}}(k). \end{aligned} \tag{6.A.36}$$

On the other hand, from the definitions of the matrices $\bar{\Phi}(k)$, $\bar{A}_{\mathbf{zv}}(k)$, and $\bar{A}_{\mathbf{zx}}(k)$ we have that

$$\begin{aligned} & \bar{\Phi}(k)\left[I - \bar{A}_{\mathbf{zv}}(k)\bar{\Phi}(k)\right]^{-1}\bar{A}_{\mathbf{zx}}(k) \\ = & \begin{bmatrix} \Phi(k+1) & 0 \\ 0 & \Phi(k) \end{bmatrix}\begin{bmatrix} I - A_{\mathbf{zv}}(k+1)\Phi(k+1) & -A_{\mathbf{zx}}(k+1)A_{\mathbf{xv}}(k)\Phi(k) \\ 0_{m_{\mathbf{z}} \times m_{\mathbf{v}}} & I - A_{\mathbf{zv}}(k)\Phi(k) \end{bmatrix}^{-1} \\ & \times \begin{bmatrix} A_{\mathbf{zx}}(k+1)A_{\mathbf{Tx}}(k) \\ A_{\mathbf{zx}}(k) \end{bmatrix} \\ = & \begin{bmatrix} \Phi(k+1)[I - A_{\mathbf{zv}}(k+1)\Phi(k+1)]^{-1}A_{\mathbf{zx}}(k+1)A(k) \\ \Phi(k)[I - A_{\mathbf{zv}}(k)\Phi(k)]^{-1}A_{\mathbf{zx}}(k) \end{bmatrix}. \end{aligned} \tag{6.A.37}$$

Hence

$$\bar{A}_{\mathbf{xv}}(k)\bar{\Phi}(k)\left[I-\bar{A}_{\mathbf{zv}}(k)\bar{\Phi}(k)\right]^{-1}\bar{A}_{\mathbf{zx}}(k)=A_{\mathbf{xv}}(k)\Phi(k)\left[I-A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1}A_{\mathbf{zx}}(k), \tag{6.A.38}$$

$$\begin{aligned}&\bar{C}_{\mathbf{v}}(k)\bar{\Phi}(k)\left[I-\bar{A}_{\mathbf{zv}}(k)\bar{\Phi}(k)\right]^{-1}\bar{A}_{\mathbf{zx}}(k)\\ =\,&C_{\mathbf{v}}(k+1)\Phi(k+1)\left[I-A_{\mathbf{zv}}(k+1)\Phi(k+1)\right]^{-1}A_{\mathbf{zx}}(k+1)A(k)\\ &+C_{\mathbf{x}}(k+1)A_{\mathbf{xv}}(k)\Phi(k)\left[I-A_{\mathbf{zv}}(k)\Phi(k)\right]^{-1}A_{\mathbf{zx}}(k).\end{aligned} \tag{6.A.39}$$

Note also that

$$\begin{aligned}\tilde{C}_{\mathbf{x}}(k)&=\mathbf{diag}\left\{C_{\mathbf{x}}(k+1,i)A_{\mathbf{Tx}}(k,i)|_{i=1}^{N}\right\}\\ &=C_{\mathbf{x}}(k+1)A_{\mathbf{Tx}}(k).\end{aligned} \tag{6.A.40}$$

Substituting the relations of Eqs. (6.A.38)–(6.A.40) into Eq. (6.A.33), we obtain the following equalities, which give the desired equations to be proved, that is, Eqs. (6.46) and (6.47):

$$\begin{aligned}\hat{x}(k+1)&=\left\{A_{\mathbf{Tx}}(k)+\tilde{A}_{\mathbf{xv}}(k)\tilde{\Phi}(k)\left[I-\tilde{A}_{\mathbf{zv}}(k)\tilde{\Phi}(k)\right]^{-1}\tilde{A}_{\mathbf{zx}}(k)\right\}\hat{x}(k)\\ &\quad+K_{\mathbf{x}}(k)[y(k+1)-\hat{y}(k+1)]\\ &=A(k)\hat{x}(k)+K_{\mathbf{x}}(k)[y(k+1)-\hat{y}(k+1)],\end{aligned} \tag{6.A.41}$$

$$\begin{aligned}\hat{y}(k+1)&=\left\{\tilde{C}_{\mathbf{x}}(k)+\tilde{C}_{\mathbf{v}}(k)\tilde{\Phi}(k)\left[I-\tilde{A}_{\mathbf{zv}}(k)\tilde{\Phi}(k)\right]^{-1}\tilde{A}_{\mathbf{zx}}(k)\right\}\hat{x}(k)\\ &=C(k+1)A(k)\hat{x}(k).\end{aligned} \tag{6.A.42}$$

This completes the proof. □

6.A.6 Proof of Theorem 6.7

For brevity, define the matrices

$$\hat{C}_i(k)=\mathbf{col}\{\bar{C}_j(k)|_{j=1,j\neq i}^{N}\},$$
$$X_i(k)=\left[P^{-1}(k)+\bar{C}_i^{T}(k)\bar{C}_i(k)\right]^{1/2}.$$

Take the value of the covariance matrix $P(k)$ as $P(k)=P^{[\mathrm{kal}]}(k)$. Then we can establish the following equality on the basis of Lemma 2.3:

$$P_{ii}(k+1)-P_{ii}^{[\mathrm{kal}]}(k+1)=A_i(k)\left\{X_i^{-2}(k)-\left[X_i^{2}(k)+\hat{C}_i^{T}(k)\hat{C}_i(k)\right]^{-1}\right\}A_i^{T}(k). \tag{6.A.43}$$

Note that

$$\begin{aligned} & X_i^{-2}(k) - \left[X_i^2(k) + \hat{C}_i^T(k)\hat{C}_i(k)\right]^{-1} \\ & = X_i^{-2}(k)\hat{C}_i^T(k)\left[I + \hat{C}_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\right]^{-1}\hat{C}_i(k)X_i^{-2}(k). \end{aligned} \tag{6.A.44}$$

We therefore have that

$$\begin{aligned} P_{ii}(k+1) - P_{ii}^{[\mathrm{kal}]}(k+1) = & \left\{A_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\left[I + \hat{C}_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\right]^{-1/2}\right\} \\ & \times \left\{A_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\left[I + \hat{C}_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\right]^{-1/2}\right\}^T. \end{aligned} \tag{6.A.45}$$

Hence, $P_{ii}(k+1) - P_{ii}^{[\mathrm{kal}]}(k+1) = 0$ if and only if

$$A_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\left[I + \hat{C}_i(k)X_i^{-2}(k)\hat{C}_i^T(k)\right]^{-1/2} = 0,$$

which is further equivalent to

$$A_i(k)\left[P^{-1}(k) + \bar{C}_i^T(k)\bar{C}_i(k)\right]^{-1}\hat{C}_i^T(k) = 0. \tag{6.A.46}$$

On the other hand, we can show through some direct algebraic manipulations that

$$\begin{aligned} & \left[P^{-1}(k) + \bar{C}_i^T(k)\bar{C}_i(k)\right]^{-1}\hat{C}_i^T(k) \\ & = \left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1}\hat{C}_i^T(k)\left\{I - \hat{C}_i(k)\left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1}\hat{C}_i^T(k)\right\}^{-1}. \end{aligned} \tag{6.A.47}$$

Combining Eqs. (6.A.46) and (6.A.47), we can claim that a necessary and sufficient condition for the equality $P_{ii}(k+1) = P_{ii}^{[\mathrm{kal}]}(k+1)$ is that

$$A_i(k)\left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1}\hat{C}_i^T(k) = 0. \tag{6.A.48}$$

From the definition of $\hat{C}_i(k)$ we straightforwardly see that this condition is equivalent to that, for all $j \neq i$,

$$A_i(k)\left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1}\bar{C}_j^T(k) = 0. \tag{6.A.49}$$

By definition $\bar{C}_j(k) = J_{\mathbf{y}j}^T \bar{C}(k)$. Direct matrix manipulations show that Eq. (6.A.49) is equivalent to Eq. (6.64).

On the other hand, from Eq. (6.23), through some tedious but straightforward algebraic operations, we can prove that when $P(k) = P^{[\mathrm{kal}]}(k)$ and $i \neq j$,

$$\begin{aligned}
& P_{ij}(k+1) - P_{ij}^{[\mathrm{kal}]}(k+1) \\
&= A_i(k)\left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1} \hat{C}_i^T(k)\hat{C}_i(k)X_i^{-2}(k)P^{-1}(k)X_j^{-2}(k)A_j^T(k) \\
&\quad - A_i(k)\left[P^{-1}(k) + \bar{C}^T(k)\bar{C}(k)\right]^{-1} \bar{C}_j^T(k)\bar{C}_j(k)X_j^{-2}(k)A_j^T(k).
\end{aligned} \tag{6.A.50}$$

From Eqs. (6.A.48) and (6.A.49) we can declared that if $P_{ii}(k+1) = P_{ii}^{[\mathrm{kal}]}(k+1)$, then for any integer j belonging to the set $\{1,\ 2,\ \cdots,\ N\}/\{i\}$,

$$P_{ij}(k+1) = P_{ij}^{[\mathrm{kal}]}(k+1).$$

In addition, from the theory of Kalman filtering [19,22] we have that its update gain matrix can also be expressed as

$$K^{[\mathrm{kal}]}(k) = A(k)\left[(P^{[\mathrm{kal}]}(k))^{-1} + \bar{C}^T(k)\bar{C}(k)\right]^{-1} \bar{C}^T(k).$$

Then from the definition of the matrices $J_{\mathbf{x}i}$ and $J_{\mathbf{y}i}$ we have that its ith row jth column block $K_{ij}^{[\mathrm{kal}]}(k)$ can be further rewritten as

$$K_{ij}^{[\mathrm{kal}]}(k) = A_i(k)\left[(P^{[\mathrm{kal}]}(k))^{-1} + \bar{C}^T(k)\bar{C}(k)\right]^{-1} \bar{C}_j^T(k).$$

We can therefore declare from Eq. (6.A.49) that if $P_{ii}(k+1) = P_{ii}^{[\mathrm{kal}]}(k+1)$, then

$$K_{ij}^{[\mathrm{kal}]}(k) = \begin{cases} A_i(k)\left[(P^{[\mathrm{kal}]}(k))^{-1} + \bar{C}^T(k)\bar{C}(k)\right]^{-1} \bar{C}_i^T(k), & i = j, \\ 0, & i \neq j. \end{cases} \tag{6.A.51}$$

This completes the proof. □

6.A.7 Proof of Theorem 6.8

Assume that the CDOSSP and the Kalman filter have the same steady covariance matrix of estimation errors. Denote it by P; that is,

$$\lim_{t\to\infty} P(k) = \lim_{t\to\infty} P^{[\mathrm{kal}]}(k) = P.$$

Then, for arbitrary $i = 1, \cdots, N$, we have

$$\lim_{t\to\infty} [P_{ii}(k) - P_{ii}^{[\mathrm{kal}]}(k)] = 0.$$

On the other hand, based on Lemma 2.3 and Eqs. (6.A.43)–(6.A.45), we can directly prove that

$$\begin{aligned}
&\lim_{t\to\infty} \left[P_{ii}(k+1) - P_{ii}^{[\mathrm{kal}]}(k+1)\right] \\
&= \left\{A_i[P^{-1} + \bar{C}_i^T\bar{C}_i]^{-1}\hat{C}_i^T\left[I + \hat{C}_i(P^{-1} + \bar{C}_i^T\bar{C}_i)^{-1}\hat{C}_i^T\right]^{-1/2}\right\}\{\star\}^T . \qquad (6.A.52)
\end{aligned}$$

Therefore, these two state estimators have the same steady estimation accuracy only if

$$A_i[P^{-1} + \bar{C}_i^T\bar{C}_i]^{-1}\hat{C}_i^T = 0, \quad i = 1, \cdots, N. \qquad (6.A.53)$$

Using the same arguments as those in Eqs. (6.A.47)–(6.A.49), we can show that if $i \neq j$ and the condition of Eq. (6.A.53) is satisfied, then

$$A_i P\bar{C}^T[I + \bar{C}P\bar{C}^T]^{-1}J_{\mathbf{y}j} = 0, \quad i, j = 1, \cdots, N. \qquad (6.A.54)$$

On the contrary, assume that system $\mathbf{\Sigma}$ is time invariant and its system matrices satisfy the condition of Eq. (6.A.54). Then, using completely the same arguments as in Theorem 6.7, we can prove that the Kalman filter has a steady block diagonal update gain matrix, that is, for an arbitrary integer pair i and j with $i, j \in \{1, 2, \cdots, N\}$ and $j \neq i$, we certainly have

$$\lim_{t\to\infty} K_{ij}^{[\mathrm{kal}]}(k) = 0. \qquad (6.A.55)$$

Initialize both the Kalman filter and the CDOSSP with the steady covariance matrix of the Kalman filter. Then, by Theorem 6.7 and Eq. (6.A.55), satisfaction of Eq. (6.65), which is completely the same as Eq. (6.A.54), means that the update gain matrix of the Kalman filter is always block diagonal.

Note that the update gain matrix of the CDOSSP is proven to be optimal among all the block diagonal ones. We can declare that the covariance matrix of its estimation errors must not exceed that of the Kalman filter. On the other hand, as Kalman filter is optimal for linear plants with normal external disturbances [19,22], its estimation accuracy must be higher than that of the CDOSSP. Therefore, these two estimators must always have the same covariance matrix. As the Kalman filter is assumed to be convergent, the CDOSSP must also converge with this initial condition. Hence they must have the same steady covariance matrix of prediction errors.

These arguments can be easily modified to situations in which the Kalman filter takes another initial covariance matrix, which essentially only requires an appropriate transformation between the temporal variables of these two predictors. In fact, let k_{kal} and k_d denote the temporal variables respectively for the Kalman filter and the CDOSSP, and assign k_d as $k_d = k_{\mathrm{kal}} + \delta_k$. Then, based on Theorem 6.7 and taking the limit as $\delta_k \to \infty$, completely the same results can be established through similar arguments. This completes the proof. □

References

[1] W.A. Porter, J.L. Aravena, State estimation in discrete m-D systems, IEEE Transactions on Automatic Control 31 (1986) 280–283.

[2] J. Cortes, Distributed kriged Kalman filters for spatial estimation, IEEE Transactions on Automatic Control 54 (2009) 2816–2827.

[3] H.Y. Liang, T. Zhou, Distributed state estimation for spatially interconnected systems and its convergence analysis, ACTA Automatica Sinica 36 (2010) 720–730 (in Chinese).

[4] T. Zhou, Coordinated one-step optimal distributed state prediction for a networked dynamical system, IEEE Transactions on Automatic Control 58 (2013) 2756–2771.

[5] T. Zhou, Optimal distributed observer design for networked dynamical systems, in: Proceedings of the 53rd IEEE Conference on Decision and Control, Los Angeles, California, USA, pp. 3358–3363.

[6] R. D'Andrea, G.E. Dullerud, Distributed control design for spatially interconnected systems, IEEE Transactions on Automatic Control 48 (2003) 1478–1495.

[7] J. Schuppen, O. Boutin, P.L. Kempker, J. Komenda, T. Masopust, N. Pambakian, A.C.M. Ran, Control of distributed systems: tutorial and overview, European Journal of Control 17 (2011) 579–602.

[8] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[9] T. Zhou, Stability and stability margin for a two-dimensional system, IEEE Transactions on Signal Processing 54 (2006) 3483–3488.

[10] J.B. Hiriart-Urruty, C. Lemarechal, Fundamentals of Convex Analysis, Springer, Berlin, Germany, 2001.

[11] D.P. Bertsekas, Convex Optimization Theory, Athena Scientific, Boston, USA, 2009.

[12] H. Abou-Kandil, G. Freiling, V. Ionescu, G. Jank, Matrix Riccati Equations in Control and System Theory, Birkhäuser Verlag, Basel, 2003.

[13] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[14] M.S. Andersen, S.K. Pakazad, A. Hansson, A. Rantzer, Robust stability of sparsely interconnected uncertain systems, IEEE Transactions on Automatic Control 59 (2014) 2151–2156.

[15] E.D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, 2009.

[16] D.D. Siljak, Large-Scale Dynamic Systems: Stability and Structure, North-Holland Books, New York, USA, 1978.

[17] T. Zhou, Further results on the boundedness of multidimensional systems, Systems & Control Letters 58 (2009) 818–825.

[18] A.E. Bryson, Y.C. Ho, Applied Optimal Control: Optimization, Estimation and Control, Tarlor & Francis, New York, USA, 1975.

[19] A.T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[20] D. Simon, Optimal State Prediction: Kalman, H_∞ and Nonlinear Approaches, Wiley-Interscience, A John Wiley & Sons, Inc., Publication, Hoboken, New Jersey, USA, 2006.

[21] J. George, Robust Kalman–Bucy filter, IEEE Transactions on Automatic Control 58 (2013) 174–180.

[22] T. Zhou, Sensitivity penalization based robust state estimation for uncertain linear systems, IEEE Transactions on Automatic Control 55 (2010) 1018–1024.

[23] T. Zhou, H.Y. Liang, On asymptotic behaviors of a sensitivity penalization based robust state estimator, Systems & Control Letters 60 (2011) 174–180.

[24] T. Zhou, Robust recursive state estimation with random measurement droppings, IEEE Transactions on Automatic Control 61 (2016) 156–171.

[25] T. Zhou, Asymptotic behavior of recursive state estimations with intermittent measurements, IEEE Transactions on Automatic Control 61 (2016) 400–415.

[26] R.E. Kalman, A new approach to linear filtering and prediction problems, Transactions of the American Society of Mechanical Engineer-Journal of Basic Engineering (Series D) 82 (1960) 34–45.

[27] T. Zhou, On the controllability and observability of networked dynamic systems, Automatica 52 (2015) 63–75.

[28] M. Rotkowitz, S. Lall, A characterization of convex problems in decentralized control, IEEE Transactions on Automatic Control 51 (2006) 274–286.

[29] J.L. Ma, Distributed State Estimations for a Large Scale System, Bachlor Thesis, Department of Automation, Tsinghua University, China, 2015 (in Chinese).

[30] D.E. Marelli, M.Y. Fu, Distributed weighted least-squares estimation with fast convergence for large-scale systems, Automatica 51 (2015) 27–39.

[31] D.P. Bertsekas, J.N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Athena Scientific, Boston, USA, 1997.

[32] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Review 51 (2009) 661–703.

[33] Y. Bar-Shalom, X.R. Li, Multitarget-Multisensor Tracking: Principles and Techniques, YBS Publishing, Storrs, CT, USA, 1995.

[34] D. Georges, G. Besancon, J.F. Dulhoste, A decentralized optimal LQ state observer based on an augmented Lagrangian approach, Automatica 50 (2014) 1451–1458.

CHAPTER 7

Stability and Robust Stability of a Large-Scale NCS

7.1 Introduction

To guarantee a satisfactory work of a system, it is necessary that its behavior can automatically return to its original trajectory after being perturbed by some external disturbances and/or when the system is not been originally set to the desirable states at the initial time instant. A system with this property is usually called stable. In other words, stability must be first guaranteed for a system to accomplish its expected tasks properly. In addition, as modeling errors are generally unavoidable, stability of a system must be kept even if some of its model parameters deviate from their nominal values and/or even if some other dynamics enter into the system behavior that are not included in its model. Due to the importance of these properties, verification of the stability and robust stability attracted extensive attentions for a long time in various fields, especially in the fields of systems and control [1]. Well-known results include Lyapunov stability theory, the Routh–Horwitz criterion for continuous-time systems, the Jury criterion for discrete-time systems, and so on. Although numerous milestone conclusions have been established, various important issues still require further investigations.

Especially, when a large-scale networked system is concerned, development of less conservative, more computationally efficient criteria is still theoretically challenging. Stimulated by the development of network and communication technologies and so on, renewed interests in this problem have extensively raised recently [2–4].

In [2], by means of introducing a spatial $\mathcal{Z}$-transformation, some sufficient conditions have been derived, which are computationally attractive for the verification of the stability of spatially distributed systems. On the other hand, in [5], situations are clarified under which these conditions become also necessary for a spatially distributed system to be stable. Through adopting some parameter-dependent Lyapunov functions, in [6], a less conservative sufficient condition for this stability verification was derived. In addition, [7] reveals some important relations between the stability of a large-scale networked system and the structured singular value of a matrix. To apply these results, however, it is required that every subsystem has the same dynamics and that all subsystems are connected regularly and along some particular spatial directions. On the other hand, when a large-scale system has a so-called sequentially

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00007-5

semiseparable structure, an iterative method is successfully developed in [8] for its stability verifications.

The concept of diagonal stability has also been utilized in the stability verification of a large-scale system. Particularly, stability of a networked system is investigated in [3] for systems consisting of only passive subsystems and having special structures. Influence of time delays on the stability of a large-scale system has been studied in [9] using a description of the so-called integral quadratic constraints. On the basis of the same descriptions for modeling errors, a sufficient condition is derived in [10] for the verification of the robust stability of the large-scale networked system described by Eqs. (6.1) and (6.2) with both its subsystem parameters and its subsystem connection matrix being time invariant. When modeling errors affect system input–output behaviors in a linear fractional way, some results are established in [11] on influence of subsystem interactions on the stability and robust stability of a networked system. Necessary and sufficient conditions are expressed there explicitly using the subsystem connection matrix of a networked system, and influence of the out-degree of a subsystem on the stability and robust stability of a networked system have also been clarified.

In this chapter, we summarize the results developed in [10,11] for the stability and robust stability of a large-scale NCS. Some necessary and sufficient conditions are given that explicitly take system structures into account, as well as some necessary or sufficient conditions that depend separately only on parameters of each subsystem and the subsystem connection matrix. Although the latter conditions are not necessary and sufficient, they can be verified independently for each individual subsystem and are therefore attractive for a large-scale NCS from the viewpoint of both numerical stability and computational costs. In addition, when the robust stability of an NCS is under investigation, both parametric modeling errors and unmodeled dynamics are permitted.

7.2 A Networked System With Discrete-Time Subsystems

In this section, we attack stability and robust stability of a networked system under the condition that each of its subsystem is described by a discrete-time model. Parallel results can be developed for a system with its subsystem dynamics described by a continuous-time model, that is, a set of first-order differential equations.

7.2.1 System Description

The NCS $\mathbf{\Sigma}$ investigated in this section is the same as that of the previous chapters, which has also been adopted in [10–12]. This system is constituted from N linear time-invariant dynamic subsystems, whereas the dynamics of its ith subsystem $\mathbf{\Sigma}_i$ is described by the following state space model like representation:

$$\begin{bmatrix} x(k+1,i) \\ z(k,i) \\ y(k,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} \begin{bmatrix} x(k,i) \\ v(k,i) \\ u(k,i) \end{bmatrix}, \tag{7.1}$$

where $i = 1, 2, \cdots, N$. Moreover, its subsystems are connected through

$$v(k) = \Phi z(k). \tag{7.2}$$

As in the previous chapters, here $z(k)$ and $v(k)$ are respectively defined as $z(k) = \textbf{col}\left\{z(k,i)|_{i=1}^{N}\right\}$ and $v(k) = \textbf{col}\left\{v(k,i)|_{i=1}^{N}\right\}$. Moreover, k and i stand respectively for the temporal variable and the index number of a subsystem, $x(k,i)$ represents the state vector of the ith subsystem $\mathbf{\Sigma}_i$ at the time instant k, $z(k,i)/v(k,i)$ represent its output/input vector to/from other subsystems, and $y(k,i)$ and $u(k,i)$ represent respectively its output and input vectors. Once again, as in the previous chapters, to distinguish $z(k,i)$ and $v(k,i)$ respectively from $y(k,i)$ and $u(k,i)$, $z(k,i)$ and $v(k,i)$ are called internal output/input vectors, whereas $y(k,i)$ and $u(k,i)$ are called external output/input vectors.

Throughout this section, the dimensions of the vectors $x(k,i)$, $v(k,i)$, $u(k,i)$, $z(k,i)$, and $y(k,i)$, are assumed respectively to be $m_{\mathbf{x}i}$, $m_{\mathbf{v}i}$, $m_{\mathbf{u}i}$, $m_{\mathbf{z}i}$, and $m_{\mathbf{y}i}$. Moreover, we assume that every row of the matrix Φ has only one nonzero element, which is equal to one. As argued in the previous chapters, this assumption does not sacrifice any generality of the adopted system model. Note also that an approximate power-law degree distribution exists extensively in science and engineering systems, such as gene regulation networks, protein interaction networks, internet, electrical power systems, and so on. For these systems, interactions among subsystems are sparse, and the matrix Φ usually has a dimension significantly smaller than that of its state vector [10,12–14]. Under such a situation, results given in this section work well in general.

7.2.2 Stability of a Networked System

To develop a computationally attractive criterion for the stability of system $\mathbf{\Sigma}$ and for its robust stability against both parametric modeling errors and unmodeled dynamics, the following results are at first introduced, which are well known in system theories [1,15–17].

Lemma 7.1. *Concerning a discrete LTI system with its state space model being $x(k+1) = Ax(k) + Bu(k)$, $y(k) = Cx(k) + Du(k)$, it is stable if and only if $\rho(A) < 1$.*

Note that well-posedness is essential for a system to work properly. In fact, a plant that does not satisfy this requirement is usually hard to be controlled, and/or its states are in general difficult to be estimated [1,4,12]. It is therefore assumed in the following discussions that the networked system $\mathbf{\Sigma}$ under investigation is well-posed, which can be explicitly expressed as a requirement that the associated matrix is invertible.

When each its subsystem is linear and time-invariant and the subsystem connection matrix is also time invariant, the networked system $\mathbf{\Sigma}$ itself is also linear and time invariant. From Lemma 7.1 it is clear that the requirement that the networked system $\mathbf{\Sigma}$ is stable can be equivalently expressed as that all the eigenvalues of its state transition matrix are smaller than 1 in magnitude. To make mathematical derivations more concise, we first define the following matrices: $A_{\mathbf{*\#}} = \mathbf{diag}\left\{A_{\mathbf{*\#}}(i)|_{i=1}^{N}\right\}$, $B_{\mathbf{*}} = \mathbf{diag}\left\{B_{\mathbf{*}}(i)|_{i=1}^{N}\right\}$, $C_{\mathbf{*}} = \mathbf{diag}\left\{C_{\mathbf{*}}(i)|_{i=1}^{N}\right\}$, and $D_{\mathbf{u}} = \mathbf{diag}\left\{D_{\mathbf{u}}(i)|_{i=1}^{N}\right\}$, where $\mathbf{*}, \# = \mathbf{x}$, $\mathbf{v}$ or $\mathbf{z}$. Moreover, denote $\mathbf{col}\left\{u(k,i)|_{i=1}^{N}\right\}$, $\mathbf{col}\left\{x(k,i)|_{i=1}^{N}\right\}$, and $\mathbf{col}\left\{y(k,i)|_{i=1}^{N}\right\}$ respectively by $u(k)$, $x(k)$, and $y(k)$. Then straightforward algebraic manipulations show that the well-posedness of the networked system $\mathbf{\Sigma}$ is equivalent to the regularity of the matrix $I - A_{\mathbf{zv}}\Phi$, that is, this matrix is invertible. Moreover, when the networked system $\mathbf{\Sigma}$ is well-posed, its dynamics can be equivalently described by the following state space model:

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \left\{ \begin{bmatrix} A_{\mathbf{xx}} & B_{\mathbf{x}} \\ C_{\mathbf{x}} & D_{\mathbf{u}} \end{bmatrix} + \begin{bmatrix} A_{\mathbf{xv}} \\ C_{\mathbf{v}} \end{bmatrix} \Phi \left[\, I - A_{\mathbf{zv}}\Phi \,\right]^{-1} \left[A_{\mathbf{zx}} \;\; B_{\mathbf{z}}\right] \right\} \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}. \tag{7.3}$$

This expression gives a lumped state space model of the networked system $\mathbf{\Sigma}$ and is completely the same as that adopted in Lemma 7.1. This property can be understood more easily if we define the matrices A, B, C, and D respectively as

$$A = A_{\mathbf{xx}} + A_{\mathbf{xv}}\Phi\left(I - A_{\mathbf{zv}}\Phi\right)^{-1} A_{\mathbf{zx}}, \quad B = B_{\mathbf{x}} + A_{\mathbf{xv}}\Phi\left(I - A_{\mathbf{zv}}\Phi\right)^{-1} B_{\mathbf{z}},$$
$$C = C_{\mathbf{x}} + C_{\mathbf{v}}\Phi\left(\, I - A_{\mathbf{zv}}\Phi \,\right)^{-1} A_{\mathbf{zx}}, \quad D = D_{\mathbf{u}} + C_{\mathbf{v}}\Phi\left(I - A_{\mathbf{zv}}\Phi\right)^{-1} B_{\mathbf{z}}.$$

Clearly, all these matrices are time invariant. Moreover, the input–output relation of the networked system $\mathbf{\Sigma}$ has been rewritten completely in the same form as that in Lemma 7.1. As in the verification of controllability and observability of the networked system, this relation enables application of Lemma 7.1 to the stability and robust stability analysis of the networked system $\mathbf{\Sigma}$. However, it is worth noting that a networked system usually has a great amount of subsystems, which means that the dimensions of the associated matrices in the lumped model are in general high. Hence, when a large-scale networked system is under investigation, it is usually not numerically feasible to straightforwardly apply the results of Lemma 7.1 to its analysis and synthesis. In addition, note that the inverse of the matrix $I - A_{\mathbf{zv}}\Phi$ is required in the aforementioned expressions, and this inversion is in general not numerically stable when the dimension of this matrix is high and/or when this matrix is nearly singular. These imply that calculations of the parameter matrices A, B, C, and D of the lumped model itself might not be reliable. As discussed in Chapter 3, these difficulties happen also to verifications of the controllability and/or the observability of the networked system $\mathbf{\Sigma}$.

To develop a computationally feasible and numerically reliable criterion for the stability of the networked system $\mathbf{\Sigma}$, the following properties of the matrix A is first established.

Lemma 7.2. *Assume that the networked system $\mathbf{\Sigma}$ is well-posed. Then, a complex number λ is not an eigenvalue of the matrix A if and only if $|I - \Phi G(\lambda)| \neq 0$. Here, the transfer function matrix $G(\lambda)$ is defined as $G(\lambda) = A_{\mathbf{zv}} + A_{\mathbf{zx}}(\lambda I - A_{\mathbf{xx}})^{-1} A_{\mathbf{xv}}$.*

Proof. When system $\mathbf{\Sigma}$ is well-posed, we have that $|I - A_{\mathbf{zv}}\Phi| \neq 0$, which is equivalent to $|I - \Phi A_{\mathbf{zv}}| \neq 0$. From this inequality and the definition of the matrix A, direct algebraic manipulations show that for an arbitrary complex number λ,

$$\left|\begin{bmatrix} \lambda I - A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ \Phi A_{\mathbf{zx}} & I - \Phi A_{\mathbf{zv}} \end{bmatrix}\right| = |I - \Phi A_{\mathbf{zv}}| \times |\lambda I - A|. \tag{7.4}$$

Hence $|\lambda I - A| \neq 0$ can be equivalently expressed as

$$\left|\begin{bmatrix} \lambda I - A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ \Phi A_{\mathbf{zx}} & I - \Phi A_{\mathbf{zv}} \end{bmatrix}\right| \neq 0. \tag{7.5}$$

Note that $\lambda I - A_{\mathbf{xx}}$ is of full normal rank. On the basis of the definition of the transfer function matrix $G(\lambda)$, the following formal expression can be derived straightforwardly:

$$\left|\begin{bmatrix} \lambda I - A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ \Phi A_{\mathbf{zx}} & I - \Phi A_{\mathbf{zv}} \end{bmatrix}\right| = |\lambda I - A_{\mathbf{xx}}| \times |I - \Phi G(\lambda)|. \tag{7.6}$$

Assume that $|\lambda I - A_{\mathbf{xx}}| \neq 0$. Then, we can declare from Eqs. (7.5) and (7.6) that $|\lambda I - A| \neq 0$ if and only if $|I - \Phi G(\lambda)| \neq 0$.

Note that the dimension of the matrix $A_{\mathbf{xx}}$ is finite. We can claimed that its eigenvalues only take isolated values. Hence, if $|\lambda I - A_{\mathbf{xx}}| = 0$, then there always exists $\varepsilon > 0$ such that, for each $\delta \in (-\varepsilon,\ 0)\bigcup(0,\ \varepsilon)$, $|(\lambda + \delta)I - A_{\mathbf{xx}}| \neq 0$. Assume further that $|\lambda I - A| \neq 0$. Then, by Eq. (7.4),

$$\begin{aligned} \lim_{\delta \to 0}\left|\begin{bmatrix} (\lambda + \delta)I - A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ \Phi A_{\mathbf{zx}} & I - \Phi A_{\mathbf{zv}} \end{bmatrix}\right| &= \left|\begin{bmatrix} \lambda I - A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ \Phi A_{\mathbf{zx}} & I - \Phi A_{\mathbf{zv}} \end{bmatrix}\right| \\ &= |I - \Phi A_{\mathbf{zv}}| \times |\lambda I - A| \\ &\neq 0. \end{aligned} \tag{7.7}$$

Therefore, we can declare from Eq. (7.6) that

$$|I - \Phi G(\lambda)| = \lim_{\delta \to 0} |I - \Phi G(\lambda + \delta)| \neq 0. \tag{7.8}$$

On the contrary, assume that $|I - \Phi G(\lambda)| \neq 0$. If $|\lambda I - A| = 0$, then we can declare from Eq. (7.4) that there exist vectors α and β such that $\mathbf{col}\{\alpha,\ \beta\} \neq 0$ and

$$(\lambda I - A_{\mathbf{xx}})\alpha + A_{\mathbf{xv}}\beta = 0, \tag{7.9}$$
$$\Phi A_{\mathbf{zx}}\alpha + (I - \Phi A_{\mathbf{zv}})\beta = 0. \tag{7.10}$$

From Eq. (7.9) it is clear that $A_{\mathbf{xv}}\beta$ belongs to the image of the matrix $\lambda I - A_{\mathbf{xx}}$. We can therefore declare from matrix theories [18] that even though $|\lambda I - A_{\mathbf{xx}}| = 0$, the vector α can still be expressed as $\alpha = -(\lambda I - A_{\mathbf{xx}})^{\dagger} A_{\mathbf{xv}}\beta$, where $(\cdot)^{\dagger}$ denotes the pseudo-inverse of a matrix. We can claim from this expression that $\beta \neq 0$.

Substituting the expression for α into Eq. (7.10), we have that

$$\left\{I - \Phi\left[A_{\mathbf{zv}} + A_{\mathbf{zx}}(\lambda I - A_{\mathbf{xx}})^{\dagger} A_{\mathbf{xv}}\right]\right\}\beta = 0.$$

This further leads to

$$\left|I - \Phi\left[A_{\mathbf{zv}} + A_{\mathbf{zx}}(\lambda I - A_{\mathbf{xx}})^{\dagger} A_{\mathbf{xv}}\right]\right| = 0.$$

Hence from the definition of the transfer function matrix $G(\lambda)$ and consistencies between the inverse and the pseudo-inverse of a matrix we can declare that $|I - \Phi G(\lambda)| = 0$. This is a contradiction with the assumption. We can therefore claim that $|\lambda I - A| \neq 0$. This completes the proof. □

On the basis of Eq. (7.5), it can be shown that the networked system $\mathbf{\Sigma}$ is stable, only when all the unstable modes of each of its subsystems are both controllable and observable. This conclusion can be established through a similarity transformation that divide the state space of each subsystem into a stable subspace and an unstable subspace, utilizing the relation of the following Eq. (7.11).

From Lemmas 7.1 and 7.2 it is clear that a necessary and sufficient condition for the networked system $\mathbf{\Sigma}$ to be stable is that, for every complex number λ satisfying $|\lambda| \geq 1$, the matrix $I - \Phi G(\lambda)$ is regular. However, it is worth noting that when the networked system under investigation is constructed from a large amount of subsystems, although the subsystem connection matrix Φ is usually sparse, its dimension is still usually high. Moreover, the associated inequality must be checked for each complex number λ satisfying $|\lambda| \geq 1$. This is generally impossible through straightforward computations. By these observations it is safe to declare that some further efforts are still required to make the results of Lemma 7.2 applicable to a large-scale networked system.

To achieve these objectives, we recall the following property for the subsystem connection matrix Φ stated and proven in Chapter 3.

Let $m(i)$ stand for the number of subsystems that is directly affected by the ith element of the vector $z(k)$, $i=1,2,\ldots,M_{\mathbf{z}}$. Define the matrices $\Theta(j)$, $j=1,2,\ldots,N$, and Θ respectively as $\Theta(j)=\mathbf{diag}\{\sqrt{m(i)}|_{i=M_{\mathbf{z},j-1}+1}^{M_{\mathbf{z},j}}\}$ and $\Theta=\mathbf{diag}\{\sqrt{m(i)}|_{i=1}^{M_{\mathbf{z}}}\}$. It has been proven in [11] and Chapter 3 that

$$\Phi^T\Phi=\Theta^2=\mathbf{diag}\left\{\Theta^2(j)\Big|_{j=1}^{N}\right\}. \tag{7.11}$$

On the basis of these results, we derive a necessary and sufficient condition for the stability of the networked system $\mathbf{\Sigma}$, which explicitly takes the structure of the networked system into account and is computationally attractive.

Theorem 7.1. *System $\mathbf{\Sigma}$ is stable if and only if there exists a positive definite matrix (PDM) P such that*

$$\begin{aligned}&\begin{bmatrix} P-A_{\mathbf{xx}}^TPA_{\mathbf{xx}}+A_{\mathbf{zx}}^T\Theta^2A_{\mathbf{zx}} & A_{\mathbf{zx}}^T\Theta^2A_{\mathbf{zv}}-A_{\mathbf{xx}}^TPA_{\mathbf{xv}} \\ A_{\mathbf{zv}}^T\Theta^2A_{\mathbf{zx}}-A_{\mathbf{xv}}^TPA_{\mathbf{xx}} & A_{\mathbf{zv}}^T\Theta^2A_{\mathbf{zv}}-A_{\mathbf{xv}}^TPA_{\mathbf{xv}}+I\end{bmatrix}\\ &\quad-\begin{bmatrix} 0 & A_{\mathbf{zx}}^T\Phi^T \\ \Phi A_{\mathbf{zx}} & \Phi A_{\mathbf{zv}}+A_{\mathbf{zv}}^T\Phi^T\end{bmatrix}>0.\end{aligned} \tag{7.12}$$

Proof. Note that $|I-\Phi G(\lambda)|\neq 0$ is equivalent to $[I-\Phi G(\lambda)]^H[I-\Phi G(\lambda)]>0$. Moreover,

$$I-\Phi G(\lambda)=[-\Phi A_{\mathbf{zx}}\ \ I-\Phi A_{\mathbf{zv}}]\begin{bmatrix}(\lambda I-A_{\mathbf{xx}})^{-1}A_{\mathbf{xv}}\\ I\end{bmatrix},$$

and $|\lambda|\geq 1$ is equivalent to

$$\begin{bmatrix}\lambda\\1\end{bmatrix}^H\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}\lambda\\1\end{bmatrix}\geq 0.$$

On the other hand, straightforward matrix operations show that

$$[I\ \ -\lambda I]\begin{bmatrix}A_{\mathbf{xx}}&A_{\mathbf{xv}}\\I&0\end{bmatrix}\begin{bmatrix}(\lambda I-A_{\mathbf{xx}})^{-1}A_{\mathbf{xv}}\\ I\end{bmatrix}=0,$$

which implies that, for each fixed λ, a vector, say α, that can be expressed as

$$\alpha=\begin{bmatrix}A_{\mathbf{xx}}&A_{\mathbf{xv}}\\I&0\end{bmatrix}\begin{bmatrix}(\lambda I-A_{\mathbf{xx}})^{-1}A_{\mathbf{xv}}\\ I\end{bmatrix}\xi,$$

where ξ is a complex vector with a compatible dimension, belongs to the null space of the matrix $[I \quad -\lambda I]$. In other words, when the Laplacian variable λ is prescribed to a particular value, we have the following relation:

$$\mathcal{S}pan\left\{\begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ I & 0 \end{bmatrix}\begin{bmatrix} (\lambda I - A_{\mathbf{xx}})^{-1}A_{\mathbf{xv}} \\ I \end{bmatrix}\right\} \subseteq \mathcal{N}ull\left\{[I \quad -\lambda I]\right\}.$$

In addition, note also that

$$\mathcal{S}pan\left\{\begin{bmatrix} \lambda I \\ I \end{bmatrix}\right\} = \mathcal{N}ull\left\{[I \quad -\lambda I]\right\},$$

which can be directly proven from Definitions 2.5 and 2.6.

On the basis of these relations, similar arguments as those in the proof of Theorems 1 and 3 of [19] and those in [5,10] show that $|I - \Phi G(\lambda)| \neq 0$, $\forall |\lambda| \geq 1$, can be equivalently expressed as the existence of a positive definite matrix P such that

$$\begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ I & 0 \end{bmatrix}^T \begin{bmatrix} -P & 0 \\ 0 & P \end{bmatrix}\begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ I & 0 \end{bmatrix} + [-\Phi A_{\mathbf{zx}} \quad I - \Phi A_{\mathbf{zv}}]^T [-\Phi A_{\mathbf{zx}} \quad I - \Phi A_{\mathbf{zv}}] > 0. \tag{7.13}$$

On the other hand, direct algebraic manipulations show that

$$\begin{aligned} &[-\Phi A_{\mathbf{zx}} \quad I - \Phi A_{\mathbf{zv}}]^T [-\Phi A_{\mathbf{zx}} \quad I - \Phi A_{\mathbf{zv}}] \\ &= \begin{bmatrix} A_{\mathbf{zx}}^T \Theta^2 A_{\mathbf{zx}} & A_{\mathbf{zx}}^T \Theta^2 A_{\mathbf{zv}} - A_{\mathbf{zx}}^T \Phi^T \\ A_{\mathbf{zv}}^T \Theta^2 A_{\mathbf{zx}} - \Phi A_{\mathbf{zx}} & A_{\mathbf{zv}}^T \Theta^2 A_{\mathbf{zv}} + I - \Phi A_{\mathbf{zv}} - A_{\mathbf{zv}}^T \Phi^T \end{bmatrix}, \end{aligned} \tag{7.14}$$

$$\begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ I & 0 \end{bmatrix}^T \begin{bmatrix} -P & 0 \\ 0 & P \end{bmatrix}\begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ I & 0 \end{bmatrix} = -\begin{bmatrix} A_{\mathbf{xx}}^T P A_{\mathbf{xx}} - P & A_{\mathbf{xx}}^T P A_{\mathbf{xv}} \\ A_{\mathbf{xv}}^T P A_{\mathbf{xx}} & A_{\mathbf{xv}}^T P A_{\mathbf{xv}} \end{bmatrix}. \tag{7.15}$$

The proof can now be completed by substituting Eqs. (7.14) and (7.15) into Eq. (7.13). □

Compared with the available results, such as those of [2,3,6,7], an attractive characteristic of Theorem 7.1 is that there exist no restrictions on the model of the networked system $\mathbf{\Sigma}$ and the condition is both necessary and sufficient. On the other hand, note that the left-hand side of Eq. (7.12) depends linearly on the matrix P and that the structure of the networked system $\mathbf{\Theta}$, which is represented by the subsystem connection matrix Φ, is explicitly included in this equation. For a small size problem, feasibility of this matrix inequality can in general be easily verified using available linear matrix inequality (LMI) solvers. On the other hand, note that all the matrices $A_{*\#}$ with $*, \# = \mathbf{x}, \mathbf{z}$ are block diagonal and a networked system usually has a sparse structure, which is reflected by the subsystem connection matrix Φ. In addition, efficient methods have been developed for sparse semidefinite programming, such as those in [10,20,21]. It is expected that the above condition works well also for a moderate size problem. This expectation has been confirmed through some numerical simulations reported in [11].

From its proof it is clear that although the matrices $A_{*\#}$ with $*, \# = \mathbf{x}, \mathbf{v}$, or $\mathbf{z}$ are block diagonal, the matrix P is usually dense. In addition, when the square of the matrix Θ in inequality (7.12) is replaced by $\Phi^T\Phi$, the results of Theorem 7.1 become valid for an arbitrary subsystem connection matrix Φ. On the other hand, a necessary condition for the feasibility of this inequality obviously is

$$P - A_{\mathbf{xx}}^T P A_{\mathbf{xx}} + A_{\mathbf{zx}}^T \Theta^2 A_{\mathbf{zx}} > 0,$$

which can be proved to be equivalent to the existence of positive definite matrices $P(i)$ for $i = 1, 2, \cdots, N$ such that

$$P(i) - A_{\mathbf{xx}}^T(i) P(i) A_{\mathbf{xx}}(i) + A_{\mathbf{zx}}^T(i)\Theta_i^2 A_{\mathbf{zx}}(i) > 0.$$

The last inequality can be verified for each subsystem independently.

It is worth pointing out that from Lyapunov stability theory we can directly declare that system $\mathbf{\Sigma}$ is stable if and only if there exists a positive definite matrix P such that $P - A^T PA > 0$, which is also a linear matrix inequality. However, as argued in [10,12] and the previous chapters, although the subsystem connection matrix Φ is generally sparse, the state transition matrix A is usually dense. This implies that computational complexity for feasibility verification of this equation is usually significantly greater than that of inequality (7.12), especially when the plant consists of a large amount of subsystems. Similar observations have also been reported in [10] for robust stability verification of system $\mathbf{\Sigma}$ with IQC described uncertainties.

When a system is of a very large-scale, numerical difficulties may still arise in verifying the condition of Theorem 7.1. To overcome these difficulties, one possibility is to find some matrices $\Theta_{ij}^{[*]} = \mathbf{diag}\{\Theta_{ij}^{[*]}(k)|_{k=1}^{N}\}$ with $i, j = 1, 2$ and $* = l, h$ that have dimensions and partitions compatible with the matrix $A_{\mathbf{xx}}$ such that

$$\begin{bmatrix} \Theta_{11}^{[l]} & \Theta_{21}^{[l]T} \\ \Theta_{21}^{[l]} & \Theta_{22}^{[l]} \end{bmatrix} \leq \begin{bmatrix} 0 & A_{\mathbf{zx}}^T\Phi^T \\ \Phi A_{\mathbf{zx}} & \Phi A_{\mathbf{zv}} + A_{\mathbf{zv}}^T\Phi^T \end{bmatrix} \leq \begin{bmatrix} \Theta_{11}^{[h]} & \Theta_{21}^{[h]T} \\ \Theta_{21}^{[h]} & \Theta_{22}^{[h]} \end{bmatrix}. \tag{7.16}$$

With availability of these matrices, direct algebraic operations show that a necessary condition for the stability of system $\mathbf{\Sigma}$ is the existence of a PDM $P(i)$ for each $1 \le i \le N$ such that

$$\begin{bmatrix} P(i) - A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xx}}(i) + A_{\mathbf{zx}}^T(i)\Theta^2(i)A_{\mathbf{zx}}(i) - \Theta_{11}^{[l]}(i) & A_{\mathbf{zx}}^T(i)\Theta^2(i)A_{\mathbf{zv}}(i) - A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xv}}(i) - \Theta_{21}^{[l]T}(i) \\ A_{\mathbf{zv}}^T(i)\Theta^2(i)A_{\mathbf{zx}}(i) - A_{\mathbf{xv}}^T(i)P(i)A_{\mathbf{xx}}(i) - \Theta_{21}^{[l]}(i) & A_{\mathbf{zv}}^T(i)\Theta^2(i)A_{\mathbf{zv}}(i) - A_{\mathbf{xv}}^T(i)P(i)A_{\mathbf{xv}}(i) + I - \Theta_{22}^{[l]}(i) \end{bmatrix} > 0.$$

Moreover, when the superscript l is replaced by h, these inequalities become a sufficient condition. Clearly, these inequalities depend linearly on $P(i)$, and their dimensions are completely and independently determined by each individual system. These properties make them much more computationally attractive than Eq. (7.12). However, further efforts are required to find the matrices $\Theta_{ij}^{[*]}(k)$ in Eq. (7.16) with $i, j = 1, 2$, $* = h, l$, and $k = 1, 2, \cdots, N$ such that the associated inequalities are both tight and computationally attractive.

Based on the properties of the SCM Φ, another computationally attractive sufficient condition is derived for the stability of system $\mathbf{\Sigma}$.

Theorem 7.2. *Denote the TFM $A_{\mathbf{zv}}(i) + A_{\mathbf{zx}}(i)[\lambda I - A_{\mathbf{xx}}(i)]^{-1}A_{\mathbf{xv}}(i)$ by $G_i(\lambda)$ and assume that $||\Theta_i G_i(\lambda)||_\infty < 1$ for each subsystem. Then system $\mathbf{\Sigma}$ is stable.*

Proof. From the definitions of the TFMs $G_i(\lambda)$ and $G(\lambda)$ it is obvious that $G(\lambda) = \mathbf{diag}\{G_i(\lambda)|_{i=1}^N\}$. Based on this relation and Eq. (7.11), we can directly prove that

$$\begin{aligned} [\Phi G(\lambda)]^H \Phi G(\lambda) &= G^H(\lambda)\Phi^T\Phi G(\lambda) \\ &= \mathbf{diag}\left\{ [\Theta_i G_i(\lambda)]^H \Theta_i G_i(\lambda)\Big|_{i=1}^N \right\}. \end{aligned} \tag{7.17}$$

From the definition of the H_∞ norm of a TFM we can declare that if $||\Theta_i G_i(\lambda)||_\infty < 1$, then $\bar{\sigma}(\Theta_i G_i(\lambda)) < 1$ for $|\lambda| \ge 1$. Note that $\rho(\star) \le \bar{\sigma}(\star)$ for every square matrix [18]. We can therefore further claim that when $|\lambda| \ge 1$ and $||\Theta_i G_i(\lambda)||_\infty < 1$, $i = 1, 2, \cdots, N$, it is certain that

$$\rho(\Phi G(\lambda)) \le \bar{\sigma}(\Phi G(\lambda)) = \max_{1 \le i \le N} \bar{\sigma}(\Theta_i G_i(\lambda)) < 1, \tag{7.18}$$

and hence $|I - \Phi G(\lambda)| \ne 0$. This completes the proof. □

Note that $\Theta_i G_i(\lambda)$ is completely determined by the SCM Φ and the parameters of the ith subsystem, and efficient methods exist for computing an upper bound of the H_∞ norm of a TFM. Therefore, the condition of Theorem 7.2 can be easily verified, and its computational complexity increases only linearly with the increment of the subsystem number N. This means that for a large-scale networked system, the computation cost of Theorem 7.2 is in general significantly lower than that of Theorem 7.1. However, it is worth emphasizing that this condition is only sufficient and its conservativeness is still not clear. On the other hand, the numerical simulation results reported in [11] show that with the increment of the subsystem number and/or the magnitude of the subsystem matrix elements, conservativeness of Theorem 7.2 usually increases.

Note also that the finiteness of $||\Theta_i G_i(\lambda)||_\infty$ implies the stability of the subsystem $\mathbf{\Sigma}_i$. Moreover, from the definition of the TFM H_∞ norm we can see that $||\Theta_i G_i(\lambda)||_\infty$ decreases monotonically with the decrement of any diagonal element of the matrix Θ_i. Theorem 7.2 therefore also makes it clear that for an NS consisting of stable subsystems, sparse connections are helpful in maintaining its stability, and subsystem interaction reduction can make an unstable system stable.

On the basis of the results given in [19], here we provide a linear matrix inequality-based condition for the verification of the H_∞ norm-based condition.

Note that $\bar{\sigma}(\Theta_i G_i(\lambda)) < 1$ is equivalent to

$$\begin{bmatrix} \Theta_i G_i(\lambda) \\ I \end{bmatrix}^H \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Theta_i G_i(\lambda) \\ I \end{bmatrix} > 0. \tag{7.19}$$

It can be directly proved from the structure of $\mathbf{col}\{\Theta_i G_i(\lambda), I\}$ that the condition of Theorem 7.2 can be equivalently expressed as the following matrix inequality:

$$[\star]^H \begin{bmatrix} -A_{\mathbf{zx}}^T(i)\Theta_i^2 A_{\mathbf{zx}}(i) & -A_{\mathbf{zx}}^T(i)\Theta_i^2 A_{\mathbf{zv}}(i) \\ -A_{\mathbf{zv}}^T(i)\Theta_i^2 A_{\mathbf{zx}}(i) & -A_{\mathbf{zv}}^T(i)\Theta_i^2 A_{\mathbf{zv}}(i)+I \end{bmatrix} \begin{bmatrix} (\lambda I - A_{\mathbf{xx}}(i))^{-1} A_{\mathbf{xv}}(i) \\ I \end{bmatrix} > 0 \tag{7.20}$$

for every complex number λ such that $|\lambda| \geq 1$.

Recall that $|\lambda| \geq 1$ is equivalent to

$$\begin{bmatrix} \lambda \\ 1 \end{bmatrix}^H \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \geq 0.$$

Based on the definition of the H_∞ norm of a transfer function matrix and Theorems 1 and 3 of [19], we can directly claim that $||\Theta_i G_i(\lambda)||_\infty < 1$ is equivalent to the existence of a positive definite matrix $P(i)$ such that

$$[\star]^T \begin{bmatrix} P(i) & 0 \\ 0 & -P(i) \end{bmatrix} \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xv}}(i) \\ I & 0 \end{bmatrix} - \begin{bmatrix} -A_{\mathbf{zx}}^T(i)\Theta_i^2 A_{\mathbf{zx}}(i) & -A_{\mathbf{zx}}^T(i)\Theta_i^2 A_{\mathbf{zv}}(i) \\ -A_{\mathbf{zv}}^T(i)\Theta_i^2 A_{\mathbf{zx}}(i) & -A_{\mathbf{zv}}^T(i)\Theta_i^2 A_{\mathbf{zv}}(i)+I \end{bmatrix} < 0, \tag{7.21}$$

which can be reexpressed as

$$\left[\begin{array}{cc} P(i)-A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xx}}(i)-A_{\mathbf{zx}}^T(i)\Theta_i^2A_{\mathbf{zx}}(i) & -A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xv}}(i)-A_{\mathbf{zx}}^T(i)\Theta_i^2A_{\mathbf{zv}}(i) \\ -A_{\mathbf{xv}}^T(i)P(i)A_{\mathbf{xx}}(i)-A_{\mathbf{zv}}^T(i)\Theta_i^2A_{\mathbf{zx}}(i) & I-A_{\mathbf{xv}}^T(i)P(i)A_{\mathbf{xv}}(i)-A_{\mathbf{zv}}^T(i)\Theta_i^2A_{\mathbf{zv}}(i) \end{array}\right]>0. \tag{7.22}$$

Clearly, the left-hand side of this matrix inequality depends linearly on the matrix $P(i)$, and its dimension is completely determined by that of the ith subsystem $\mathbf{\Sigma}_i$. Therefore, feasibility of this inequality can be effectively verified in general.

Note that a necessary condition for the feasibility of inequality (7.22) is the existence of a positive definite matrix $P(i)$ such that

$$P(i)-A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xx}}(i)-A_{\mathbf{zx}}^T(i)\Theta_i^2A_{\mathbf{zx}}(i)>0.$$

This inequality further leads to

$$P(i)-A_{\mathbf{xx}}^T(i)P(i)A_{\mathbf{xx}}(i)>0.$$

According to the Lyapunov stability theory [1,2,4], this means that the matrix $A_{\mathbf{xx}}(i)$ is stable. In other words, a necessary condition for the feasibility of inequality (7.22) is that when every subsystem of system $\mathbf{\Sigma}$ is completely isolated from interactions with all the other subsystems of the plant, then all they are stable. Clearly, this is generally not required to guarantee the stability of the whole system, noting that a controllable and observable open loop unstable system can be stabilized by an appropriately designed feedback controller. From these aspects it is also clear that the associated condition is generally conservative.

7.2.3 Robust Stability of a Networked System

Modeling errors are unavoidable in practical engineering. A general requirement about a system property is that it is not sensitive to modeling errors, which is widely known as robustness of this property. This section investigates the robust stability of a networked system when there exist both parametric modeling errors and unmodeled dynamics in a state space model of each of its subsystems. To deal with this problem, an extra input vector and an extra output vector are introduced into the description of the dynamics of its ith subsystem $\mathbf{\Sigma}_i$ with the purpose of reflecting influences of modeling errors while the subsystem connections remain unchanged. More precisely, the state space model of Eq. (7.1) is modified into the following form:

$$\begin{bmatrix} x(k+1,i) \\ w(k,i) \\ z(k,i) \\ y(k,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) & A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) & A_{\mathbf{wv}}(i) & B_{\mathbf{w}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) & A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{d}}(i) & C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} \begin{bmatrix} x(k,i) \\ d(k,i) \\ v(k,i) \\ u(k,i) \end{bmatrix}. \tag{7.23}$$

In addition, the unmodeled dynamics in this subsystem and the parametric errors in the matrices $A_{*\#}(i)$ with $*$, $\# = \mathbf{x}, \mathbf{d}, \mathbf{v}, \mathbf{w}, \mathbf{z}$ are respectively described by

$$d(k,i) = \Delta_u(q,i)w(k,i), \quad ||\Delta_u(q,i)||_\infty \leq 1, \tag{7.24}$$

$$A(i) = A_0(i) + E_i \Delta_p(i) F_i, \quad \bar{\sigma}(\Delta_p(i)) \leq 1, \tag{7.25}$$

$$A(i) = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) & A_{\mathbf{xv}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) & A_{\mathbf{wv}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) & A_{\mathbf{zv}}(i) \end{bmatrix}. \tag{7.26}$$

Here q stands for the one-step forward shift operator, $\Delta_u(q,i)$ and $\Delta_p(i)$ have respectively a prescribed structure represented by $\boldsymbol{\Delta}_u(i)$ and $\boldsymbol{\Delta}_p(i)$. Moreover, E_i and F_i are known matrices with compatible dimensions. In this description, the matrices E_i and F_i are introduced to reflect how parametric errors influence subsystem parameters, whereas $\Delta_u(q,i)$ and $\Delta_p(i)$ denote respectively unmodeled dynamics and parametric errors existing in the state space model. This description is general enough to describe a large class of system dynamics. Moreover, a widely adopted structure for modeling errors is that they are block diagonal [1,22].

With this uncertainty model, the following results are obtained for the robust stability of the networked system $\boldsymbol{\Sigma}$. Their proof is given in the appendix attached to this chapter.

Theorem 7.3. *Let ζ denote the variable of $\mathcal{Z}$-transformation. Define the matrix G_i and the uncertainty structure $\boldsymbol{\Delta}(i)$ respectively as*

$$G_i = \begin{bmatrix} 0 & F_i \\ \mathbf{diag}\{I_{m_{\mathbf{x}i}},\ I_{m_{\mathbf{w}i}},\ \Theta_i\}[E_i & A_0(i)] \end{bmatrix},$$

$$\boldsymbol{\Delta}(i) = \left\{ \Delta \;\middle|\; \begin{array}{l} \Delta = \mathbf{diag}\{\Delta_p(i),\ \delta I_{m_{\mathbf{x}i}},\ \Delta_u(\zeta,i),\ \Delta(\zeta)\} \\ \quad \Delta_p(i) \in \boldsymbol{\Delta}_p(i),\ \delta \in \mathcal{C} \\ \quad \Delta_u(\zeta,i) \in \boldsymbol{\Delta}_u(i),\ \Delta(\zeta) \in \mathcal{H}_\infty \end{array} \right\}.$$

Assume that all subsystems $\boldsymbol{\Sigma}_i|_{i=1}^{N}$ are well posed. If $\mu_{\boldsymbol{\Delta}(i)}(G_i) < 1$ for each $i = 1, 2, \cdots, N$, then the networked system $\boldsymbol{\Sigma}$ is robustly stable against all the modeling errors described by Eqs. (7.24)–(7.26).

Note that the condition of Theorem 7.3 can be verified independently for each subsystem. This means that the well-developed SSV analysis methods can be directly applied to its verification in general. Compared with the results of [10], Theorem 7.3 is valid for a wider class of uncertainty models and have a lower computational complexity. As a matter of fact, its computational complexity increases only linearly with the increment of the subsystem number N.

On the other hand, from Eq. (7.A.4) we can derive another SSV-based sufficient condition for the robust stability of system $\boldsymbol{\Sigma}$ simultaneously using all subsystem parameter matrices and the subsystem connection matrix. However, the corresponding matrix and the uncertainty structure generally have a high dimension, which is not appropriate to be applied to a large-scale networked system.

7.3 A Networked System With Continuous-Time Subsystems

In this section, we investigate the robust stability of a networked system when its subsystem is described by a set of first-order differential equations and its modeling errors are described by several IQCs, which stand for integral quadratic constraint, and is to some extent a little abstract description. As in the previous section, similar arguments can lead to parallel results for am NCS with discrete-time subsystems.

7.3.1 Modeling Errors Described by IQCs

In the description of modeling errors, a well-known but relatively abstract expression is the so-called integral quadratic constraints (IQCs) [23,24]. As its name indicates, in this description, both integration and a quadratic form are utilized. More precisely, let $\Pi(\cdot, t)$ be a bounded self-adjoint operator, and let $\Delta(\cdot, t)$ be a bounded and causal operator that maps a p-dimensional signal to a q-dimensional signal. If for each p-dimensional signal $v(t)$ with finite energy, that is, $v(t) \in \mathcal{L}_2^p$,

$$\int_0^\infty \begin{bmatrix} v(t) \\ \Delta(v(t), t) \end{bmatrix}^T \Pi\left(\begin{bmatrix} v(t) \\ \Delta(v(t), t) \end{bmatrix}, t\right) dt \geq 0, \tag{7.27}$$

then we say that the operator $\Delta(\cdot, t)$ satisfies the IQC defined by the operator $\Pi(\cdot, t)$, which is usually denoted by $\Delta \in \mathcal{IQC}(\Pi(\cdot, t))$. This inequality is capable of describing many modeling errors adopted in system analysis and synthesis. For example, a $q \times p$-dimensional parametric modeling error Δ with $\bar{\sigma}(\Delta) < \gamma$ can be easily seen to satisfy

$$\int_0^\infty \begin{bmatrix} v \\ \Delta(v) \end{bmatrix}^T \begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_q \end{bmatrix} \begin{bmatrix} v \\ \Delta(v) \end{bmatrix} dt \geq 0, \quad \forall v(t) \in \mathcal{L}_2^p,$$

whereas the $q \times p$-dimensional unmodeled dynamics $\Delta(\cdot)$ with $||\Delta||_\infty < \gamma$ can be straightforwardly proved to meet the IQC

$$\int_0^\infty \begin{bmatrix} v \\ \Delta(v) \end{bmatrix}^T \begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_q \end{bmatrix} \begin{bmatrix} v \\ \Delta(v) \end{bmatrix} dt \geq 0, \quad \forall v(t) \in \mathcal{L}_2^p.$$

In addition to these, IQCs are also able to describe system properties like passivity, and combine several IQCs into a single IQC [25,26].

Utilization of IQCs in system analysis and synthesis can be traced back to the 1960s, in which they are extensively applied to the absolute stability analysis of a nonlinear system, whereas their applications to the analysis and synthesis of a system with modeling errors were started around the beginning of the 1990s [23–26]. A particularly important property of IQCs is that they have a clear frequency domain interpretation, which makes engineering intuition in system analysis and synthesis much clearer and easier to be understood. In particular, if the operator $\Pi(\cdot, t)$ of Eq. (7.27) is linear and time invariant and has a transfer function matrix representation, then the IQC of Eq. (7.27) can be equivalently rewritten as

$$\int_{-\infty}^{\infty} \begin{bmatrix} v(j\omega) \\ \Delta(v)(j\omega) \end{bmatrix}^H \Pi(j\omega) \begin{bmatrix} v(j\omega) \\ \Delta(v)(j\omega) \end{bmatrix} d\omega \geq 0. \tag{7.28}$$

In addition to this, an IQC constraint is also capable of describing nonlinear dynamic uncertainties, which makes the associated system analysis and synthesis methods able to cope with nonlinear modeling errors.

The following results are fundamental in robust stability analysis of a system with IQC described uncertainties [10,25]. They also play central roles in the derivations of a criterion for the robust stability of a networked system, which is the main objective of this section.

Lemma 7.3. *Consider a feedback system with $y(s) = G(s)u(s)$ and $u(t) = \Delta(y(t))$, where $G(s) \in \mathcal{RH}_{\infty}^{p\times q}$ and $\Delta(\cdot) \in \mathcal{IQC}(\Pi(\cdot))$. Moreover, let $\Pi(\cdot)$ be a bounded and self-adjoint operator. Suppose that*

- *for each $\delta \in [0,\ 1]$, the feedback system with $u(t) = \Delta(y(t))$ replaced by $u(t) = \delta\Delta(y(t))$ is well-posed;*
- *$\delta\Delta(\cdot) \in \mathcal{IQC}(\Pi(\cdot))$ for arbitrary $\delta \in [0,\ 1]$;*
- *there exists a positive number ε such that for an arbitrary angular frequency $\omega \in [0,\ \infty]$, the following inequality is satisfied:*

$$\begin{bmatrix} G(j\omega) \\ I \end{bmatrix}^H \Pi(j\omega) \begin{bmatrix} G(j\omega) \\ I \end{bmatrix} \leq -\varepsilon I. \tag{7.29}$$

Then, the feedback system is robustly stable with respect to the uncertain operator $\Delta(\cdot)$.

Remark 7.1. *For a particular uncertain operator $\Delta(\cdot)$, there may exist various bounded and self-adjoint operators $\Pi(\cdot)$ such that $\Delta(\cdot) \in \mathcal{IQC}(\Pi(\cdot))\mathbf{\Pi}_{\Delta}$. This is a well-encountered case in actual system analysis and synthesis [23]. Under such a situation, Lemma 7.3 declares that among these bounded and self-adjoint operators $\Pi(\cdot)$, if one of them satisfies inequality (7.29), then the feedback system consisting of $y(s) = G(s)u(s)$ and $u(t) = \Delta(y(t))$ is robustly stable. Note that the conditions of Lemma 7.3 are only sufficient. An appropriate selection of this bounded self-adjoint operator $\Pi(\cdot)$ is important in the reduction of the conservativeness of these conditions.*

7.3.2 Robust Stability With IQC-Described Modeling Errors

In this section, we investigate once again the robust stability of a networked system with dynamics described similarly as in Eqs. (7.23) and modeling errors for each subsystem described by an IQC. Rather than a discrete-time system, a continuous-time system is dealt with, that is, interactions among its subsystems are still described by an equation similar to Eq. (7.2), and the input–output relation of its ith subsystem by an equation similar to Eq. (7.23) with the signal transfer from $w(t,i)$ to $d(t,i)$ described by

$$d(t,i) = \Delta(w(t,i), i, t). \tag{7.30}$$

Moreover, there exists a bounded self-adjoint operator $\Pi(\cdot, i, t)$, such that $\Delta(\cdot, i, t) \in \mathcal{IQC}(\Pi(\cdot, i, t))$. To avoid awkward presentations, we assume that there are no longer any other modeling uncertainties in the subsystem matrices $A_{*\#}(i)$ with $*$, $\# = \mathbf{x}, \mathbf{d}, \mathbf{v}, \mathbf{w}, \mathbf{z}$.

More precisely, the dynamics of the ith subsystem $\mathbf{\Sigma}_i$ is described by

$$\begin{bmatrix} \frac{dx(t,i)}{dt} \\ w(t,i) \\ z(t,i) \\ y(t,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) & A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) & A_{\mathbf{wv}}(i) & B_{\mathbf{w}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) & A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{d}}(i) & C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} \begin{bmatrix} x(t,i) \\ d(t,i) \\ v(t,i) \\ u(t,i) \end{bmatrix}, \tag{7.31}$$

whereas the subsystem interactions are described by

$$v(t) = \Phi z(t), \tag{7.32}$$

where the vectors $v(t)$ and $z(t)$ are defined similarly to the vectors $v(k)$ and $z(k)$ of Eq. (7.2), which are stacked respectively by the internal input vectors $v(t,i)|_{i=1}^{N}$ and the internal output vectors $z(t,i)|_{i=1}^{N}$ row by row.

When the modeling errors in each subsystem of the networked system are described by Eq. (7.30), define the operators $\Delta(\cdot)$ and $\Pi(\cdot)$ as

$$\Delta(w(t)) = \mathbf{col}\left\{\Delta(w(t,i),i,t)|_{i=1}^{N}\right\} \quad \text{and} \quad \Pi(\cdot) = \mathbf{diag}\left\{\Pi(\cdot,i,t)|_{i=1}^{N}\right\}, \tag{7.33}$$

where $w(t) = \mathbf{col}\left\{w(t,i)|_{i=1}^{N}\right\}$. Moreover, denote the vector $\mathbf{col}\left\{d(t,i)|_{i=1}^{N}\right\}$ by $d(t)$. Then it is clear from these definitions that

$$d(t) = \Delta(w(t)). \tag{7.34}$$

Note that

$$\int_0^{\infty} \begin{bmatrix} w(t) \\ d(t) \end{bmatrix}^T \Pi\left(\begin{bmatrix} w(t) \\ d(t) \end{bmatrix}, t\right) dt$$

$$\begin{aligned} &= \int_0^{\infty} \begin{bmatrix} w(t) \\ \Delta(w(t),t) \end{bmatrix}^T \Pi\left(\begin{bmatrix} w(t) \\ \Delta(w(t),t) \end{bmatrix}, t\right) dt \\ &= \sum_{i=1}^{N} \int_0^{\infty} \begin{bmatrix} w(t,i) \\ \Delta(w(t,i),i,t) \end{bmatrix}^T \Pi\left(\begin{bmatrix} w(t,i) \\ \Delta(w(t,i),i,t) \end{bmatrix}, t, i\right) dt. \end{aligned} \tag{7.35}$$

It is clear that if the modeling errors of each subsystem can be described by an IQC, then the total modeling errors of the whole networked system can also be described by an IQC. More precisely, we can straightforwardly prove from the definition of an integral quadratic constraint that the operator $\Delta(\cdot)$ belongs to the set $\mathcal{IQC}\left(\Pi(\cdot,t)\right)$ if and only if for each $i \in \{1, 2, \cdots, N\}$, the uncertainty operator $\Delta((\cdot),i)$ of the subsystem $\boldsymbol{\Theta}_i$ belongs to the set $\mathcal{IQC}\left(\Pi(\cdot,i,t)\right)$. The proof is quite obvious and is therefore omitted.

The above relations can be expressed in a more explicit and concise way if the IQC associated inequality is described in the frequency domain for the modeling errors of each subsystem. More precisely, assume that for the ith subsystem $\boldsymbol{\Sigma}_i$, its modeling error $d(t,i) = \Delta(w(t,i),i,t)$ is described by

$$\int_{-\infty}^{\infty} \begin{bmatrix} w(j\omega,i) \\ d(j\omega,i) \end{bmatrix}^H \Pi(j\omega,i) \begin{bmatrix} w(j\omega,i) \\ d(j\omega,i) \end{bmatrix} d\omega \geq 0, \tag{7.36}$$

where $\Pi(j\omega,i)$ is a complex matrix-valued function that is Hermitian at each angular frequency ω. Partition this matrix-valued function as

$$\Pi(j\omega,i) = \begin{bmatrix} \Pi_{ww}(j\omega,i) & \Pi_{wd}(j\omega,i) \\ \Pi_{dw}(j\omega,i) & \Pi_{dd}(j\omega,i) \end{bmatrix}, \tag{7.37}$$

where $\Pi_{pq}(j\omega,i)$ is an $(m_{\mathbf{p}i} \times m_{\mathbf{q}i})$-dimensional complex matrix-valued function. Here $p,q = w,d$. Using these functions, define the complex matrix-valued functions $\Pi_{pq}(j\omega) = \mathbf{diag}\left\{ \Pi_{pq}(j\omega,i)\big|_{i=1}^{L} \right\}$ for $p,q = w,d$. Moreover, on the basis of these complex matrix-valued functions, define the other complex matrix-valued function

$$\Pi(j\omega) = \begin{bmatrix} \Pi_{ww}(j\omega) & \Pi_{wd}(j\omega) \\ \Pi_{dw}(j\omega) & \Pi_{dd}(j\omega) \end{bmatrix}. \tag{7.38}$$

Then it is obvious that this complex matrix-valued function is Hermitian. In addition, for signals $w(t)$ and $d(t)$ defined by Eq. (7.34), straightforward algebraic manipulations show that

$$\int_{\infty}^{\infty} \begin{bmatrix} w(j\omega) \\ d(j\omega) \end{bmatrix}^H \Pi(j\omega) \begin{bmatrix} w(j\omega) \\ d(j\omega) \end{bmatrix} d\omega \geq 0. \tag{7.39}$$

In this section, we discuss only situations where each operator $\Pi(\cdot, i, t)|_{i=1}^{N}$ is time invariant. The temporal variable t is therefore omitted to have a concise expression in the following investigations.

As in the previous section, the Laplace transform of a time series is denoted by the same symbol with temporal variable t replaced by s, the variable of the Laplace transform. Taking the Laplace transformation on both sides of Eqs. (7.31) and (7.32), we have that

$$\begin{bmatrix} sx(s,i) \\ w(s,i) \\ z(s,i) \\ y(s,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) & A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) & A_{\mathbf{wv}}(i) & B_{\mathbf{w}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) & A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{d}}(i) & C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} \begin{bmatrix} x(s,i) \\ d(s,i) \\ v(s,i) \\ u(s,i) \end{bmatrix}, \tag{7.40}$$

$$v(s) = \Phi z(s). \tag{7.41}$$

Substitute the expression for the vector $v(s)$ given by Eq. (7.41) into Eq. (7.40). Straightforward algebraic manipulations show that

$$\begin{bmatrix} w(s) \\ y(s) \end{bmatrix} = \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(s) & G_{\mathbf{wu}}^{[l]}(s) \\ G_{\mathbf{yd}}^{[l]}(s) & G_{\mathbf{yu}}^{[l]}(s) \end{bmatrix} \begin{bmatrix} d(s) \\ u(s) \end{bmatrix}, \tag{7.42}$$

where

$$\begin{aligned} \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(s) & G_{\mathbf{wu}}^{[l]}(s) \\ G_{\mathbf{yd}}^{[l]}(s) & G_{\mathbf{yu}}^{[l]}(s) \end{bmatrix} &= \begin{bmatrix} A_{\mathbf{wd}} & B_{\mathbf{w}} \\ C_{\mathbf{d}} & D_{\mathbf{u}} \end{bmatrix} + \begin{bmatrix} A_{\mathbf{wx}} & A_{\mathbf{wv}} \\ C_{\mathbf{x}} & C_{\mathbf{v}} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \Phi \end{bmatrix} \\ &\quad \times \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} A_{\mathbf{xx}} & A_{\mathbf{xv}} \\ A_{\mathbf{zx}} & A_{\mathbf{zv}} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \Phi \end{bmatrix} \right)^{-1} \begin{bmatrix} A_{\mathbf{xd}} & B_{\mathbf{x}} \\ A_{\mathbf{zd}} & B_{\mathbf{z}} \end{bmatrix}. \end{aligned}$$

Clearly, when the dynamics of the networked system $\mathbf{\Sigma}$ are described by Eqs. (7.34) and (7.42), connections between its nominal system and modeling errors are completely the same as in Lemma 7.3. This implies that the results of this lemma may be straightforwardly applicable to the aforementioned networked system. More precisely, by Lemma 7.3 we have that when the modeling errors are described by Eq. (7.30) for each subsystem of the networked system described by Eqs. (7.23) and (7.32), there exists a Hermitian complex matrix-valued function $\Pi(j\omega)$ satisfying

$$\begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}^{H} \Pi(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix} \leq -\varepsilon I \tag{7.43}$$

at each angular frequency ω, and then this networked system is robustly stable against these modeling errors.

However, it is worth mentioning that, completely as the difficulties encountered in dealing with other issues related to a networked system with a large number of subsystems, such as controllability and/or observability verifications, and so on, the dimension of the transfer function matrix $G_{\mathbf{wd}}^{[l]}(s)$ is generally high when the amount of subsystems is large. In addition to this, note that the inverse of a sparse matrix is in general dense. This means that in spite that the matrices $A_{*,\#}$ with $*, \# = \mathbf{w}, \mathbf{d}, \mathbf{v}, \mathbf{x}$ are block diagonal, both matrices $B_{\mathbf{w}}$ and $B_{\mathbf{x}}$ are also block diagonal, and that the subsystem connection matrix Φ is usually sparse, the transfer function matrix $G_{\mathbf{wd}}^{[l]}(s)$ is usually dense. Hence, for a large-scale networked system, direct applications of Lemma 7.3 may suffer from both computational cost and numerical stability problems.

To overcome these problems, the structure of the networked system is utilized in this section in its robust stability analysis, which has also been performed in the previous section when the modeling errors are described by Eqs. (7.24)–(7.26). For this purpose, the influences among the subsystems of the networked system, which is given by Eq. (7.2), is first expressed equivalently as an IQC.

Note that Eq. (7.2) is equivalent to that, for an arbitrary positive definite matrix X with an appropriate dimension, the following inequality is satisfied:

$$[v(t) - \Phi z(t)]^T X [v(t) - \Phi z(t)] \leq 0. \tag{7.44}$$

Hence, the subsystem interactions can be equivalently expressed as

$$\int_0^{\infty} \begin{bmatrix} v(t) \\ z(t) \end{bmatrix}^T \begin{bmatrix} -\Phi^T X \Phi & \Phi^T X \\ X\Phi & -X \end{bmatrix} \begin{bmatrix} v(t) \\ z(t) \end{bmatrix} dt \geq 0, \quad \forall z(t) \in \mathcal{L}_2^{M_{\mathbf{z}}}. \tag{7.45}$$

Define the operator $\bar{\Delta}(\cdot)$ as

$$\bar{\Delta}\left(\begin{bmatrix} w(t) \\ z(t) \end{bmatrix}\right) = \begin{bmatrix} \Delta(w(t)) \\ \Phi z(t) \end{bmatrix} = \begin{bmatrix} \Delta & \\ & \Phi \end{bmatrix}\left(\begin{bmatrix} w(t) \\ z(t) \end{bmatrix}\right).$$

Then, obviously from the definitions of the associated vectors we have the equality

$$\begin{bmatrix} d(t) \\ v(t) \end{bmatrix} = \bar{\Delta}\left(\begin{bmatrix} w(t) \\ z(t) \end{bmatrix}\right). \tag{7.46}$$

On the other hand, similarly to the derivation of Eq. (7.42), we can straightforwardly prove from Eq. (7.40) that

$$\begin{bmatrix} w(s) \\ z(s) \\ y(s) \end{bmatrix} = \begin{bmatrix} G_{\mathbf{wd}}(s) & G_{\mathbf{wv}}(s) & G_{\mathbf{wu}}(s) \\ G_{\mathbf{zd}}(s) & G_{\mathbf{zv}}(s) & G_{\mathbf{zu}}(s) \\ G_{\mathbf{yd}}(s) & G_{\mathbf{yv}}(s) & G_{\mathbf{yu}}(s) \end{bmatrix} \begin{bmatrix} d(s) \\ v(s) \\ u(s) \end{bmatrix}, \tag{7.47}$$

where

$$\begin{bmatrix} G_{\mathbf{wd}}(s) & G_{\mathbf{wv}}(s) & G_{\mathbf{wu}}(s) \\ G_{\mathbf{zd}}(s) & G_{\mathbf{zv}}(s) & G_{\mathbf{zu}}(s) \\ G_{\mathbf{yd}}(s) & G_{\mathbf{yv}}(s) & G_{\mathbf{yu}}(s) \end{bmatrix}$$
$$= \begin{bmatrix} A_{\mathbf{wd}} & A_{\mathbf{wv}} & B_{\mathbf{w}} \\ A_{\mathbf{zd}} & A_{\mathbf{zv}} & B_{\mathbf{z}} \\ C_{\mathbf{d}} & C_{\mathbf{v}} & D_{\mathbf{u}} \end{bmatrix} + \begin{bmatrix} A_{\mathbf{wx}} \\ A_{\mathbf{zx}} \\ C_{\mathbf{x}} \end{bmatrix} (sI - A_{\mathbf{xx}})^{-1} \begin{bmatrix} A_{\mathbf{xd}} & A_{\mathbf{xv}} & B_{\mathbf{x}} \end{bmatrix}.$$

Differently from the transfer function matrices $G_{\mathbf{pq}}^{[l]}(j\omega)$ with $p = w, d$ and $q = u, y$, which are defined in Eq. (7.42) and are usually dense, all the transfer function matrices $G_{\mathbf{pq}}(j\omega)$ with $p = w, z, y$ and $q = d, v, u$ are block diagonal.

Note that when $\delta = 0$, $\delta\Phi$ does not satisfy the IQC (7.44). This means that although Eqs. (7.46) and (7.47) construct a feedback connection as that of Lemma 7.3, its results cannot be directly applied to the verification of the robust stability of the networked system since its second condition is not satisfied. However, due to the specific structure of the transfer function matrix $G_{\mathbf{wd}}^{[l]}(j\omega)$, we can prove that satisfaction of the associated condition is equivalent to the feasibility of the matrix inequality of Eq. (7.43). As a matter of fact, we have the following results with their proof deferred to the appendix attached to this chapter.

Theorem 7.4. *There exists a complex matrix-valued Hermitian function $\Pi(j\omega)$ satisfying Eq. (7.43) if and only if there exist a complex matrix-valued function $\bar{\Pi}(j\omega)$, a positive scalar x, and a positive scalar $\bar{\varepsilon}$ such that*

$$\bar{\Pi}(j\omega) = \begin{bmatrix} \bar{\Pi}_{ww}(j\omega) & \bar{\Pi}_{wd}(j\omega) \\ \bar{\Pi}_{dw}(j\omega) & \bar{\Pi}_{dd}(j\omega) \end{bmatrix}, \quad \bar{\Pi}^H(j\omega) = \bar{\Pi}(j\omega), \tag{7.48}$$

and for each $s = j\omega$ with $\omega \in \mathcal{R}$, we have the matrix inequality

$$\begin{bmatrix} G_{\mathbf{wd}}(s) & G_{\mathbf{wv}}(s) \\ G_{\mathbf{zd}}(s) & G_{\mathbf{zv}}(s) \\ I & 0 \\ 0 & I \end{bmatrix}^H \begin{bmatrix} \bar{\Pi}_{ww}(s) & 0 & \bar{\Pi}_{wd}(s) & 0 \\ 0 & -x\Phi^T\Phi & 0 & x\Phi^T \\ \bar{\Pi}_{dw}(s) & 0 & \bar{\Pi}_{dd}(j\omega) & 0 \\ 0 & x\Phi & 0 & -xI \end{bmatrix} \begin{bmatrix} G_{\mathbf{wd}}(s) & G_{\mathbf{wv}}(s) \\ G_{\mathbf{zd}}(s) & G_{\mathbf{zv}}(s) \\ I & 0 \\ 0 & I \end{bmatrix} \le -\bar{\varepsilon}I. \tag{7.49}$$

Compared to Eq. (7.43), all the transfer function matrices in Eq. (7.49) are block diagonal. In addition, the subsystem connection matrix Φ is generally sparse for a large-scale networked system. This makes it possible to use results about sparse matrix computations in the feasibility verification of the associated matrix inequality, for which many efficient algorithms have been developed [21,27]. Numerical studies in [10] show that it really can significantly reduce computation costs for various types of large-scale networked systems.

7.4 Concluding Remarks

In summary, this chapter investigates both stability and robust stability for a networked system with nominal LTI subsystems and time-independent subsystem connections. The plant subsystems can have different nominal TFMs, and there do not exist any restrictions on their interactions. Some sufficient and necessary conditions have been derived. These conditions only depend on the subsystem connection matrix and parameter matrices of each plant subsystem, which make their verification easily implementable in general for a large-scale networked system. However, under some situations, for example, when plant subsystems are densely interconnected, both the subsystem connection matrix Φ and some of the transfer function matrices $G_i(\lambda)|_{i=1}^{N}$ may have a high dimension. In this case, results of this chapter usually do not significantly reduce computation costs, and further efforts are required to develop computationally more efficient methods.

In addition to these, influence of many important factors on the stability and robust stability of a networked system, such as signal transmission delays between subsystems, data droppings, and so on, have not been explicitly discussed here. Some of these factors, however, can be incorporated into a system model as modeling uncertainties [22,23], which may enable application of the results in this chapter to these situations.

7.5 Bibliographic Notes

Results of this chapter are on the basis of [10] and [11]. There are also various other researches on the stability and robust stability of a networked system, where its model has some special constraints, and different concepts of stability are utilized. For example, diagonal stability is investigated for a networked system in [3] with a cactus graph structure, which is basically based on system passivity analysis. In [2], the spatial $\mathcal{Z}$-transformation is utilized in the analysis and synthesis of a networked system that has identical subsystem dynamics, regular subsystem interactions, and infinite number of subsystems. Some sufficient conditions have been established there respectively for the stability and contractiveness of the networked system, and situations are clarified in [5], where this sufficient condition also becomes necessary. A less conservative stability condition is derived in [6] using the geometrical structure of the null space of a matrix polynomial and a necessary and sufficient condition based on the idea of parameter-dependent linear matrix inequalities. On the basis of dissipativity theory, some sufficient conditions are derived in [29] for the stability and contractiveness of a networked system in which each subsystem may have different dynamics and subsystem connections are arbitrary, provided that the number of signals transferred from the ith subsystem to the jth subsystem is equal to that from the jth subsystem to the ith subsystem. A sequentially semiseparable approach is adopted in [28] for the analysis and synthesis of a networked system with a string subsystem interconnection, in which each subsystem is permitted to have different dynamics, but the conditions for stability and contractiveness are only sufficient.

Appendix 7.A

7.A.1 Proof of Theorem 7.3

Denote the $\mathcal{Z}$-transform of a time series using the same symbol but replacing the temporal variable k by ζ. Taking the $\mathcal{Z}$-transformation on both sides of Eq. (7.23), we have that

$$\begin{bmatrix} \zeta x(\zeta,i) \\ w(\zeta,i) \\ z(\zeta,i) \\ y(\zeta,i) \end{bmatrix} = \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) & A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) & A_{\mathbf{wv}}(i) & B_{\mathbf{w}}(i) \\ A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) & A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{d}}(i) & C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} \begin{bmatrix} x(\zeta,i) \\ d(\zeta,i) \\ v(\zeta,i) \\ u(\zeta,i) \end{bmatrix}. \tag{7.A.1}$$

From this equation and Eq. (7.24) straightforward algebraic manipulations show that

$$\begin{bmatrix} z(\zeta,i) \\ y(\zeta,i) \end{bmatrix} = \begin{bmatrix} G_{\mathbf{zv}}(\zeta,i) & G_{\mathbf{zu}}(\zeta,i) \\ G_{\mathbf{yv}}(\zeta,i) & G_{\mathbf{yu}}(\zeta,i) \end{bmatrix} \begin{bmatrix} v(\zeta,i) \\ u(\zeta,i) \end{bmatrix}, \tag{7.A.2}$$

where

$$\begin{aligned} &\begin{bmatrix} G_{\mathbf{zv}}(\zeta,i) & G_{\mathbf{zu}}(\zeta,i) \\ G_{\mathbf{yv}}(\zeta,i) & G_{\mathbf{yu}}(\zeta,i) \end{bmatrix} \\ &= \begin{bmatrix} A_{\mathbf{zv}}(i) & B_{\mathbf{z}}(i) \\ C_{\mathbf{v}}(i) & D_{\mathbf{u}}(i) \end{bmatrix} + \begin{bmatrix} A_{\mathbf{zx}}(i) & A_{\mathbf{zd}}(i) \\ C_{\mathbf{x}}(i) & C_{\mathbf{d}}(i) \end{bmatrix} \begin{bmatrix} \zeta^{-1}I & \\ & \Delta_u(\zeta,i) \end{bmatrix} \\ &\quad \times \left(I - \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) \end{bmatrix} \begin{bmatrix} \zeta^{-1}I & \\ & \Delta_u(\zeta,i) \end{bmatrix} \right)^{-1} \begin{bmatrix} A_{\mathbf{xv}}(i) & B_{\mathbf{x}}(i) \\ A_{\mathbf{wv}}(i) & B_{\mathbf{w}}(i) \end{bmatrix}. \end{aligned}$$

Define $G_{*\#}(\zeta) = \mathbf{diag}\{G_{*\#}(\zeta,i)|_{i=1}^{N}\}$ with $*, \# = \mathbf{z}, \mathbf{y}, \mathbf{v}, \mathbf{u}$. Then

$$\begin{bmatrix} z(\zeta) \\ y(\zeta) \end{bmatrix} = \begin{bmatrix} G_{\mathbf{zv}}(\zeta) & G_{\mathbf{zu}}(\zeta) \\ G_{\mathbf{yv}}(\zeta) & G_{\mathbf{yu}}(\zeta) \end{bmatrix} \begin{bmatrix} v(\zeta) \\ u(\zeta) \end{bmatrix}. \tag{7.A.3}$$

On the other hand, from Eq. (7.2) we have that $v(\zeta) = \Phi z(\zeta)$. Substituting this relation into the last equation, we have that $y(\zeta) = \left[G_{\mathbf{yu}}(\zeta) + G_{\mathbf{yv}}(\zeta)\left(I - \Phi G_{\mathbf{zv}}(\zeta)\right)^{-1} \Phi G_{\mathbf{zu}}(\zeta)\right] u(\zeta)$. Then, by Lemma 7.1 we can declare that system $\mathbf{\Sigma}$ is stable if and only if

$$|I - \Phi G_{\mathbf{zv}}(\zeta)| \neq 0, \quad \forall |\zeta| \geq 1. \tag{7.A.4}$$

As in Theorem 7.2 we can prove that the above inequality is satisfied if $||\Theta_i G_{\mathbf{zv}}(\zeta, i)||_\infty < 1$ for every $i = 1, \cdots, N$, which is equivalent to

$$|I - \Theta_i G_{\mathbf{zv}}(\zeta, i)\Delta_a(\zeta, i)| \neq 0, \quad \forall \Delta_a(\zeta, i) \in \mathcal{BH}_\infty. \tag{7.A.5}$$

In addition, from the well-posedness assumption on each subsystem $\mathbf{\Sigma}_i$ we can declare that the matrix

$$I - \begin{bmatrix} A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) \\ A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) \end{bmatrix} \begin{bmatrix} \zeta^{-1} I_{m_{\mathbf{x}i}} & \\ & \Delta_u(\zeta, i) \end{bmatrix}$$

is of full normal rank. By Theorem 2.3 and the definition of the transfer function matrix $G_{\mathbf{zv}}(\zeta, i)$ straightforward algebraic manipulations show that Eq. (7.A.5) is equivalent to that for all $|\delta| \leq 1$, $\Delta_u(\zeta, i) \in \mathcal{B}\mathbf{\Delta}_u(i)$, and $\Delta_a(\zeta, i) \in \mathcal{BH}_\infty$,

$$\left| I - \begin{bmatrix} \Theta_i A_{\mathbf{zv}}(i) & \Theta_i A_{\mathbf{zx}}(i) & \Theta_i A_{\mathbf{zd}}(i) \\ A_{\mathbf{xv}}(i) & A_{\mathbf{xx}}(i) & A_{\mathbf{xd}}(i) \\ A_{\mathbf{wv}}(i) & A_{\mathbf{wx}}(i) & A_{\mathbf{wd}}(i) \end{bmatrix} \begin{bmatrix} \Delta_a(\zeta, i) & & \\ & \delta I_{m_{\mathbf{x}i}} & \\ & & \Delta_u(\zeta, i) \end{bmatrix} \right| \neq 0. \tag{7.A.6}$$

By Eq. (7.26) this inequality can be rewritten as

$$\left| I - \mathbf{diag}\{I_{m_{\mathbf{x}i}}, I_{m_{\mathbf{w}i}}, \Theta_i\} A(i) \mathbf{diag}\{\delta I_{m_{\mathbf{x}i}},\ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i)\} \right| \neq 0. \tag{7.A.7}$$

Substitute Eqs. (7.25) into this inequality. Note that to guarantee the satisfaction of Eq. (7.A.7) for each $\Delta_p(i) \in \mathcal{B}\mathbf{\Delta}_p(i)$, it is necessary that this inequality is satisfied for $\Delta_p(i) = 0$. This means that

$$\left| I - \mathbf{diag}\{I_{m_{\mathbf{x}i}}, I_{m_{\mathbf{w}i}}, \Theta_i\} A_0(i) \mathbf{diag}\{\delta I_{m_{\mathbf{x}i}},\ \ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i)\}\right| \neq 0.$$

Combining these two equations, we can claim from Theorem 2.3 that Eq. (7.A.7) is equivalent to

$$\left| I - \begin{bmatrix} \mathbf{diag}\{I_{m_{\mathbf{x}i}}, I_{m_{\mathbf{w}i}}, \Theta_i\}[A_0(i) \ \ E_i] \\ F_i \qquad 0 \end{bmatrix} \mathbf{diag}\{\delta I_{m_{\mathbf{x}i}},\ \ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i),\ \Delta_p(i)\} \right| \neq 0. \tag{7.A.8}$$

Note that

$$\begin{aligned} &\mathbf{diag}\{\delta I_{m_{\mathbf{x}i}},\ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i),\ \Delta_p(i)\} \\ &= \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \mathbf{diag}\{\Delta_p(i),\ \delta I_{m_{\mathbf{x}i}},\ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i)\} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}. \end{aligned}$$

Direct matrix operations show that the above inequality is further equivalent to

$$\left| I - G_i \mathbf{diag}\{\Delta_p(i),\ \delta I_{m_{\mathbf{x}i}},\ \Delta_u(\zeta, i),\ \Delta_a(\zeta, i)\} \right| \neq 0.$$

The proof can now be completed through a direct application of Definition 2.4 of the structured singular value. □

7.A.2 Proof of Theorem 7.4

To prove Theorem 7.4, we first construct the following matrix inequality:

$$\begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ I & 0 \\ 0 & I \end{bmatrix}^H \begin{bmatrix} \bar{\Pi}_{ww}(j\omega) & 0 & \bar{\Pi}_{wd}(j\omega) & 0 \\ 0 & -\Phi^T X \Phi & 0 & \Phi^T X \\ \bar{\Pi}_{dw}(j\omega) & 0 & \bar{\Pi}_{dd}(j\omega) & 0 \\ 0 & \Phi X & 0 & -X \end{bmatrix}$$

$$\times \begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ I & 0 \\ 0 & I \end{bmatrix} \leq -\bar{\varepsilon} I, \tag{7.A.9}$$

where X is a positive definite matrix with a compatible dimension. Assume that it is feasible.

Note that

$$\begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ 0 & I \end{bmatrix}.$$

Substitute this relation into Eq. (7.A.9). Straightforward matrix manipulations show that it is equivalent to

$$\begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ 0 & I \end{bmatrix}^H \begin{bmatrix} \bar{\Pi}_{ww}(j\omega) & \bar{\Pi}_{wd}(j\omega) & 0 & 0 \\ \bar{\Pi}_{dw}(j\omega) & \bar{\Pi}_{dd}(j\omega) & 0 & 0 \\ 0 & 0 & -\Phi^T X \Phi & \Phi^T X \\ 0 & 0 & \Phi X & -X \end{bmatrix}$$

$$\times \begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \\ G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ 0 & I \end{bmatrix} \leq -\bar{\varepsilon} I, \tag{7.A.10}$$

that is,

$$\begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix} + \begin{bmatrix} G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ 0 & I \end{bmatrix}^H$$

$$\times \left(- \begin{bmatrix} -\Phi^T \\ I \end{bmatrix} X \,[-\Phi \;\; I] \right) \begin{bmatrix} G_{\mathbf{zd}}(j\omega) & G_{\mathbf{zv}}(j\omega) \\ 0 & I \end{bmatrix} \leq -\bar{\varepsilon} I, \tag{7.A.11}$$

which can be reexpressed as

$$\begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix}$$
$$+ \left[-\Phi G_{\mathbf{zd}}(j\omega) \;\; I - \Phi G_{\mathbf{zv}}(j\omega)\right]^H X \left[-\Phi G_{\mathbf{zd}}(j\omega) \;\; I - \Phi G_{\mathbf{zv}}(j\omega)\right] \le -\bar{\varepsilon} I. \quad (7.A.12)$$

On the other hand, from Eqs. (7.42) and (7.47) direct but tedious matrix manipulations show that

$$G_{\mathbf{wd}}^{[l]}(s) = G_{\mathbf{wd}}(s) + G_{\mathbf{wv}}(s)\left[I - \Phi G_{\mathbf{zv}}(s)\right]^{-1} \Phi G_{\mathbf{zd}}(s). \quad (7.A.13)$$

Multiply both sides of Eq. (7.A.12) from left and right respectively by

$$\begin{bmatrix} I & 0 \\ \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]^{-1} \Phi G_{\mathbf{zd}}(j\omega) & I \end{bmatrix}^H \text{ and } \begin{bmatrix} I & 0 \\ \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]^{-1} \Phi G_{\mathbf{zd}}(j\omega) & I \end{bmatrix}.$$

Based on Eq. (7.A.13), we obtain the inequality

$$\begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) & G_{\mathbf{wv}}(j\omega) \\ I & 0 \end{bmatrix}$$
$$+ \left[0 \;\; I - \Phi G_{\mathbf{zv}}(j\omega)\right]^H X \left[0 \;\; I - \Phi G_{\mathbf{zv}}(j\omega)\right] \le -\tilde{\varepsilon} I, \quad (7.A.14)$$

where

$$\tilde{\varepsilon} = \bar{\varepsilon} \inf_{\omega \in \mathcal{R}} \sigma_{\min}^2 \left(\begin{bmatrix} I & 0 \\ \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]^{-1} \Phi G_{\mathbf{zd}}(j\omega) & I \end{bmatrix} \right).$$

Note that the matrix $\begin{bmatrix} I & 0 \\ \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]^{-1} \Phi G_{\mathbf{zd}}(j\omega) & I \end{bmatrix}$ is of full rank at each $\omega \in \mathcal{R}$. It is obvious that $\tilde{\varepsilon}$ is also a positive number.

According to the Schur complement theorem (Lemma 2.4), the satisfaction of inequality (7.A.14) is equivalent to the satisfaction of the following two inequalities:

$$\begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix} \le -\tilde{\varepsilon} I, \quad (7.A.15)$$

$$\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix} - \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]^H X \left[I - \Phi G_{\mathbf{zv}}(j\omega)\right]$$

$$
\begin{aligned}
&+\left(\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}\right)^H \\
&\times \left(\begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix} + (\tilde{\varepsilon}+\delta)I\right)^{-1} \\
&\times \left(\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}\right) \le -\tilde{\varepsilon} I,
\end{aligned}
\tag{7.A.16}
$$

where δ is a very small positive number.

From the well-posedness assumption about the networked system we can declare that the matrix $I - \Phi G_{\mathbf{zv}}(s)$ is of full rank at every complex value of the Laplace transform variable s. This implies that when inequality (7.A.15) is satisfied, there always exists a positive definite matrix X such that inequality (7.A.16) is also satisfied. As a matter of fact, we can even declaim that under such that a situation, there always exists a positive scalar x such that the positive definite matrix $X = xI$ satisfies inequality (7.A.16). To clarify this point, define the scalars

$$
\begin{aligned}
\kappa_l &= \inf_{\omega\in\mathcal{R}} \sigma_{\min}\left(I - \Phi G_{\mathbf{zv}}(j\omega)\right) \quad \text{and} \\
\kappa_h &= \sup_{\omega\in\mathcal{R}} \lambda_{\max}\left\{\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}\right. \\
&\quad + \left(\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}\right)^H \\
&\quad \times \left(\begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix} + (\tilde{\varepsilon}+\delta)I\right)^{-1} \\
&\quad \left.\times \left(\begin{bmatrix} G_{\mathbf{wv}}(j\omega) \\ 0 \end{bmatrix}^H \bar{\Pi}(j\omega) \begin{bmatrix} G_{\mathbf{wd}}^{[l]}(j\omega) \\ I \end{bmatrix}\right)\right\}.
\end{aligned}
$$

Then, from the well-posedness of the networked system and the satisfaction of inequality (7.A.15) we can directly claim that

$$
\kappa_l > 0 \quad \text{and} \quad \kappa_h < +\infty.
$$

Define the matrix

$$
X = \max\left\{\alpha,\ \frac{\kappa_h + \tilde{\varepsilon}}{\kappa_l^2}\right\} I,
\tag{7.A.17}
$$

where α is an arbitrary positive number. Then, X is positive definite and satisfies inequality (7.A.16).

Therefore, feasibility of inequality (7.43) is equivalent to inequality (7.A.9) and to inequality (7.49). This completes the proof. □

References

[1] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[2] R. D'Andrea, G.E. Dullerud, Distributed control design for spatially interconnected systems, IEEE Transactions on Automatic Control 48 (2003) 1478–1495.

[3] M. Arcak, Diagonal stability on cactus graphs and application to network stability analysis, IEEE Transactions on Automatic Control 56 (2011) 2766–2777.

[4] J. Schuppen, O. Boutin, P.L. Kempker, J. Komenda, T. Masopust, N. Pambakian, A.C.M. Ran, Control of distributed systems: tutorial and overview, European Journal of Control 17 (2011) 579–602.

[5] R.S. Chandra, R. D'Andrea, A scaled small gain theorem with applications to spatially interconnected systems, IEEE Transactions on Automatic Control 51 (2006) 465–469.

[6] T. Zhou, On the stability of spatially distributed systems, IEEE Transactions on Automatic Control 53 (2008) 2385–2391.

[7] T. Zhou, Stability and stability margin for a two-dimensional system, IEEE Transactions on Signal Processing 54 (2006) 3483–3488.

[8] J.K. Rice, M. Verhaegen, Distributed control in multiple dimensions: a structure preserving computational technique, IEEE Transactions on Automatic Control 56 (2011) 516–530.

[9] E. Summers, M. Arcak, A. Packard, Delay robustness of interconnected passive systems: an integral quadratic constraint approach, IEEE Transactions on Automatic Control 58 (2013) 712–724.

[10] M.S. Andersen, S.K. Pakazad, A. Hansson, A. Rantzer, Robust stability of sparsely interconnected uncertain systems, IEEE Transactions on Automatic Control 59 (2014) 2151–2156.

[11] T. Zhou, Y. Zhang, On the stability and robust stability of networked dynamic systems, IEEE Transactions on Automatic Control 61 (2016) 1595–1600.

[12] T. Zhou, Coordinated one-step optimal distributed state prediction for a networked dynamical system, IEEE Transactions on Automatic Control 58 (2013) 2756–2771.

[13] E.D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, 2009.

[14] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Review 51 (2009) 661–703.

[15] T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[16] D.D. Siljak, Large-Scale Dynamic Systems: Stability and Structure, North-Holland Books, New York, USA, 1978.

[17] E.D. Sontag, Mathematical Control Theory: Deterministic Finite Dimensional Systems, second edition, Springer-Verlag, New York, Inc., New York, USA, 1998.

[18] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[19] T. Zhou, On nonsingularity verification of uncertain matrices over a quadratically constrained set, IEEE Transactions on Automatic Control 56 (2011) 2206–2212.

[20] S.J. Benson, Y. Ye, DSDP5 user guide-software for semidefinite programming, Tech. Rep. ANL/MCS-TM-277, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA.

[21] A. George, J.R. Gilbert, J. Liu (Eds.), Graph Theory and Sparse Matrix Computation, Springer-Verlag, New York, USA, 1993.

[22] A. Packard, J.C. Doyle, The complex structured singular value, Automatica 29 (1993) 71–109.

[23] A. Megretski, Necessary and sufficient conditions of stability: a multi-loop generalization of the circle criterion, IEEE Transactions on Automatic Control 38 (1993) 753–756.

[24] G.E. Dullerud, F. Paganini, A Course in Robust Control Theory: A Convex Approach, Springer-Verlag, New York, USA, 2000.

[25] A. Megretski, A. Rantzer, System analysis via integral quadratic constraints, IEEE Transactions on Automatic Control 42 (1997) 819–830.

[26] A. Megretski, S. Treil, Power distribution inequalities in optimization and robustness of uncertain systems, Journal of Mathematical Systems, Estimation and Control 3 (1993) 301–319.

[27] G.H. Gloub, C. Loan, Matrix Computation, 2nd edition, The John Hopkins University Press, Baltimore, USA, 1989.

[28] J.K. Rice, M. Verhaegen, Distributed control: a sequentially semi-separable approach for spatially heterogeneous linear systems, IEEE Transactions on Automatic Control 54 (2009) 1270–1283.

[29] C. Langbort, R.S. Chandra, R. D'Andrea, Distributed control design for systems interconnected over an arbitrary graph, IEEE Transactions on Automatic Control 49 (9) (2004) 1502–1519.

CHAPTER 8

Control With Communication Constraints

8.1 Introduction

There has been a great interest in studying the impact of communication constraint on the networked control systems. The communication constraint is modeled as a discrete-time noiseless digital channel connecting the sensor to the controller. For each time step, this channel is only capable of transmitting a finite number of bit information, which involves quantized feedback control in networked control systems.

The idea of modeling the quantization error as an additive white Gaussian noise began to be challenged in the new environment where only very coarse information is allowed to propagate through the network. The change of view on quantization can be traced back to the paper [2], where the author treated quantization as partial information of the quantized entity rather than its approximation and demonstrated the significance of the historical values of the quantizer output. Since then, various methods for studying quantization effects on control and estimation have been developed.

Research on quantized feedback control can be categorized depending on whether the quantizer is static or dynamic. A static quantizer is a memoryless nonlinear function, whereas a dynamic quantizer uses memory and is more complicated and potentially more powerful. Following [2], the work [3] studied a dynamic finite-level uniform quantizer for stabilization and pointed out that there exist a dynamic adjustment policy for the quantizer sensitivity and a quantized state feedback controller to asymptotically stabilize an unstable linear system. This raised a fundamental question: how much information needs to be communicated between the quantizer and controller to stabilize an unstable linear system? Various authors have addressed this problem under different scenarios, see e.g. [4–7], and the appealing data rate theorem states that the minimum average data rate required for stabilization has to be strictly greater than a universal low bound.

The striking feature of these results is that the minimum data rate rely solely on the unstable eigenvalues of the open-loop system and is completely described by the so-called intrinsic entropy rate of the system. On the other hand, the problem of packet dropouts has been extensively studied in the literature as well. The packet dropout process is commonly modeled as an i.i.d. process [8] or a Markov chain [9,10].

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00008-7

The joint effects of limited data rate and packet dropout rate on the mean square stabilization of an unstable discrete LTI system have been studied under the assumption that the quantized input signal is to be transmitted through a lossy channel whose packet dropout process is modeled as an i.i.d. process [25] or a Markov process [27]. The main result shows that if the packet dropout rate is less than the threshold derived in [8], then the minimum data rate for the mean square stabilization of an unstable discrete LTI system is explicitly given by the intrinsic entropy rate [11] of the system, plus an additional nonnegative term, which is a function of the dropout rate. This term exactly quantifies the amount of the additional bit rate required to counter the effect of packet dropout on stabilization and monotonically converges to zero as the packet dropout rate decreases to zero, suggesting that our results naturally recover the well known-result mentioned before.

It should be noted that the almost sure stabilization of unstable systems over lossy channels has been investigated in [12,13]. Nonetheless, the stabilization problems under the mean square sense and almost sure sense are different, leading to different data rate requirements to achieve respective stabilization. The nominal moment stabilization has been studied in [30] using a parameterized notation of anytime capacity.

8.2 Entropies and Capacities of a Communication Channel

8.2.1 Entropy in Information Theory

For any probability distribution, *entropy* is a quantity to capture the uncertainty of information of a random variable, which agrees with the intuitive notion of a measure of information.

Definition 8.1. *([14]) The entropy of a discrete random variable X with distribution function $p(x)$ and sample space $\mathcal{X}$ is defined as*

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{8.1}$$

The log is to the base 2, and entropy is expressed in *bits* as it quantifies the number of bits needed to fully represent the associated random variable. The entropy can also be interpreted as the expectation of $-\log p(X)$, where X is drawn according to probability mass function $p(x)$. Then $H(X) = -\mathbb{E}[\log p(X)]$, where $\mathbb{E}[\cdot]$ is the mathematical expectation operator.

For two discrete random variables X and Y with joint probability mass function $p(x, y)$, the *joint entropy* is defined as $H(X, Y) = -\mathbb{E}[\log p(X, Y)]$. Similarly, the *conditional entropy* is defined by $H(X|Y) = -\mathbb{E}[\log p(X|Y)]$, where $p(x|y)$ is the conditional distribution function of X given Y.

Mutual information is a measure of the dependence between two random variables. For two discrete random variables X and Y with joint distribution $p(x, y)$, the mutual information is defined as

$$I(X; Y) = \mathbb{E}\left[\log \frac{p(X, Y)}{p(X)p(Y)}\right].$$

In particular, $I(X; Y) = 0$ if X and Y are independent, which essentially means that there is no mutual information between random variables X and Y, and $I(X; X) = H(X)$.

Remark 8.1. *If X is a continuous random variable, the differential entropy is defined accordingly; see [14, Chapter 9] for details.*

Example 8.1. *Let $X = 1$ with probability p and $X = 0$ with probability $1 - p$. Then, $H(X) = -p \log p - (1 - p) \log(1 - p)$. We can easily verify that $H(X) = 0$ when $p = 0$ or 1. This makes sense because if $p = 0$ or 1, then the variable X is essentially not random, and there is no uncertainty. Similarly, $H(x)$ is maximized at $p = 1/2$, which corresponds to the maximum uncertainty.*

If we have a sequence of n random variables, the *entropy rate* is defined as the growth rate of the entropy of the sequence with n.

Definition 8.2. *The entropy rate of a stochastic process $\{X_i\}$ is defined by*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{8.2}$$

when the limit exists.

If $\{X_i\}$ is an independent identically distributed (i.i.d.) process, then $H(\mathcal{X}) = H(X_1)$. If $\{X_i\}$ is a stationary stochastic process, it is easy to verify that the limit in (8.2) always exists [14]. It is well recognized that entropy plays an important role in the information and communication theories. More thorough discussions on entropy can be found in [14].

In the modern control theory, the *topological entropy* of a dynamical system is central to feedback control. Topological entropy measures the rate of generating information of a dynamical system by its initial state.

8.2.2 Topological Entropy in Feedback Theory

In information theory, the entropy rate is used to measure the rate at which a stochastic process generates information, whereas in feedback control theory, the rate at which a dynamical system with inputs generates information is quantified by *topological entropy* of Adler et al. [15]. It is expected that the topological entropy will be important to the data rate problem for stabilization of dynamical systems.

Definition 8.3. *The topological entropy of an LTI system with open-loop matrix A is defined as*

$$H_T(A) = \sum_i \max\{\log_2 |\lambda_i|, 0\},$$

where $\lambda_1, \ldots, \lambda_n$ denote all the eigenvalues of A.

This is equivalent to the Mahler measure [16] or the degree of instability [17] of the plant. The mathematician Kurt Mahler first introduced his measure to polynomials [16]. The Mahler measure of a monic polynomial $a(z) = \prod_{i=1}^{n}(z - a_i)$ is defined as

$$M(a) \triangleq \prod_{i=1}^{n} \max\{|a_i|, 1\}. \tag{8.3}$$

The Mahler measure of a square matrix $A \in \mathbb{R}^{n\times n}$ is given by that of its characteristic polynomial:

$$M(A) \triangleq M(\det(zI - A)) = \prod_i \max\{|\lambda_i|, 1\} = 2^{H_T(A)}. \tag{8.4}$$

Then the Mahler measure of an LTI plant with any detectable and stabilizable realization (A, B, C, D) can be defined as the Mahler measure of the system matrix A. The degree of instability of a square matrix A is defined in the same way as in (8.4) [17].

It is clear that the definition of topological entropy or Mahler measure makes no reference to *any* controller or feedback communication. This underlines its fundamental nature as an intrinsic property of dynamical system. In this book, we reveal the importance of the topological entropy or Mahler measure to NCSs.

8.2.3 Channel Capacities

In a communication system, source symbols from some finite samples are encoded into some sequence of channel input symbols, which then produces the sequence of channel output symbols. We attempt to recover the transmitted message from the output sequence. Since two different input sequences may result in the same output sequence, the input may be nonperfectly recovered.

A noisy communication channel is a system in which the output depends probabilistically on its input. It is characterized by a probability transition matrix that determines the conditional distribution of the output given the input. For a communication channel input X and output Y, the capacity $\mathcal{C}$ is defined as

$$\mathcal{C} = \max_{p(x)} I(X;Y), \tag{8.5}$$

where the maximum in (8.5) is taken over all possible input distributions $p(x)$.

Example 8.2 (Noiseless Binary Channel). *Suppose there is a channel whose binary input is reproduced exactly by the output, that is, any transmitted symbol is received without error. Then, the capacity of the channel is $\mathcal{C} = 1$ bit.*

Example 8.3 (Binary Erasure Channel). *The capacity of a binary erasure channel is*

$$\mathcal{C} = 1 - \alpha, \tag{8.6}$$

where α denotes the fraction of the erased bits.

A transmitted signal is usually corrupted by a channel additive noise. The additive noise channel model is one of the simplest yet typical models for a communication link. For an additive white Gaussian noise channel, the capacity can be computed simply from the noise characteristics of the channel as [14]

$$\mathcal{C} = \frac{1}{2}\log_2(1+\gamma), \tag{8.7}$$

where γ represents the signal-to-noise ratio (SNR) of the channel.

At each unit time, a symbol s_k from an elementary sample S_k of possibly time-varying size $\mu_k \geq 1$ is transmitted through a channel. For noiseless channels, s_k is received without error. The capacity $\mathcal{C}$ of a discrete noiseless channel is given by

$$\mathcal{C} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log_2(\mu_k) \tag{8.8}$$

when the limit exists.

The channel capacity can be used to characterize diverse communication constraints depending on the underlying channel model and information pattern. However, quantization and delay are unavoidable in every digital communication system. In information theory, quantizers are considered as information encoders and thus as an integral part of the whole system [14]. To achieve Shannon's capacity, the classic information theory allows for arbitrarily long sequences in coding, which results in significant time delays. Both quantization and delay are deemed to be necessary in standard information theory rather than undesirable.

8.3 Stabilization Over Communication Channel

Consider a linear time-invariant (LTI) system

$$\begin{cases} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k, \end{cases} \tag{8.9}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ are the system state and output measurement at time k, respectively, and $u_k \in \mathbb{R}^p$ is the control input. The initial state x_0 is unknown.

To make the problem well-posed, (A, B, C) are assumed to be stabilizable and detectable, with unstable A. The output sensor equipped with an encoder communicates with the controller over a digital channel that can only support information exchange with a finite bit rate. At each time instant, the sensor sends one encoded symbol s_k from a finite and possibly time-varying set $\mathcal{S}_k$ to the controller. On the other side of the channel, the controller decodes the received symbols and produces an input signal u_k to stabilize the system. Define the transmission data rate in the asymptotic average sense as

$$R = \liminf_{k\to\infty} \frac{1}{k} \sum_{t=0}^{k-1} \log_2 |\mathcal{S}_t|, \tag{8.10}$$

where $|\mathcal{S}_t|$ denotes the cardinality of the set $\mathcal{S}_t$. Clearly, the smaller the transmission rate R, the less information the controller obtains from the system and vice versa. A natural question is whether there exists a minimum data rate above which the controller is able to stabilize the LTI system.

8.3.1 Classical Approach for Quantized Control

Due to the data rate constraint, the information has to be quantized before being transmitted to the controller. The quantizer Q is a function whose range is discrete and usually finite, that is,

$$Q : \mathbb{R} \to \{q^1, \ldots, q^M\}.$$

The input to a quantizer is an analog value, and the output is from a finite set. The quantization noise is given by

$$w := Q(x) - x.$$

In the early development, it prevailed to model the quantization error w as an additive white Gaussian noise, that is,

$$Q(x) = x + w,$$

where w is assumed to be an additive white Gaussian noise uncorrelated with the random variable x. Then the well-developed tools from linear stochastic control theory can be used. Although this approach may be reasonable when the quantizer is of high resolution, it has at least one main shortcoming in control.

We use a simple example to elaborate it. Consider a scalar, fully observed, and unstable linear system, that is, (8.9) with $m = 1$, $A = a$ with $|a| > 1$, $B = C = 1$, and unknown x_0. By modeling the quantization error w_k as an additive white Gaussian noise the data available to the controller is expressed as the noisy measurement:

$$y_k' := Q(x_k) = x_k + w_k,$$

where the variance of the random noise w_k is constant, that is, $\mathbb{E}[w_k^2] = \sigma^2$, and w_k is uncorrelated with x_k. The shortcoming of this approach becomes obvious: a controller cannot asymptotically stabilize the system in the mean square sense as the noise cannot be eventually eliminated by a linear controller. Particularly, let $u_k = f y_k'$ where f is a control gain. Then, the closed-loop system is given as

$$x_{k+1} = a x_k + f(x_k + w_k) = (a + f)x_k + f w_k. \tag{8.11}$$

Since x_k and n_k are uncorrelated, it follows that

$$\mathbb{E}[x_{k+1}^2] = (a + f)^2 \mathbb{E}[x_k^2] + f^2\sigma^2. \tag{8.12}$$

Since $|a| > 1$, we have $f \neq 0$, which in turn implies that $\lim_{k\to\infty} \mathbb{E}[x_k^2] \neq 0$ for any feedback gain f. This implies that there does not exist any linear controller stabilizing the system.

To achieve the mean square stability of the closed-loop system, it actually requires the controller to estimate the initial state x_0 with mean square error diminishing strictly faster than a^{-2k}. This motivates us to study the quantized control with a more rigorous approach and obtain the following result.

Theorem 8.1. *Consider a networked control system (8.9), where the output sensor is connected to the controller via a noiseless digital channel. Then a necessary and sufficient condition for the asymptotic stabilization of the system is that*

$$R > \sum_{i=1}^{n} \max\{0, \log_2 |\lambda_i|\} := R_{\inf}, \tag{8.13}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

This result does not impose any assumption on the coder and control law except causality, which is reminiscent of the errorless Shannon source coding theorem [14]. It thus draws a fundamental line of demarcation between what is and is not achievable with linear systems when communication rates are limited. In this sense, R_{inf}, which is called *topological entropy*, plays a role similar to the source entropy in Shannon source coding theorem and can be quantified as a measure of the rate at which information is generated by an unstable linear plant. The communication between the sensor and controller is to reduce the plant uncertainty for the controller. The data rate quantifies how fast the reduction rate can be achieved. From this point of view, the channel must transport data as fast as it is produced, i.e., $R > R_{\mathrm{inf}}$.

A more physical insight can be gained by rewriting inequality (8.13) as

$$2^R > \prod_{|\lambda_i| \geq 1} |\lambda_i|.$$

The right-hand side is simply the factor by which a volume in the unstable subspace increases at each time step due to the plant dynamics, whereas the left-hand side is the asymptotic average number of disjoint regions into which the coder can partition the volume. In other words, the system is stabilizable if and only if the dynamical increase in “uncertainty volume” due to unstable dynamics is outweighed by the partitioning induced by the coder.

8.4 Universal Lower Bound

The lower bound in (8.13) is derived by using the argument of volume-partitioning. We first apply a coordinate transform to decouple the unstable and stable parts of the open-loop matrix A. Clearly, the state variables associated with the stable part automatically converge to zero for any initial state without using any control inputs. Thus, there is no need to communicate any information between the controller and the system for this subspace. This essentially implies that there is no loss of generality to assume that all the eigenvalues of A are unstable. We do this for the purpose of simplifying the presentation.

Let m_k be the Lebesgue measure of the set of values that x_k can take at time k. Considering the system dynamics and geometric interpretation of determinant, it follows that after k time steps, the plant dynamics expands the measure m_0 of the initial uncertainty set by the factor

$$\left(\prod_{|\lambda_i| \geq 1} |\lambda_i| \right)^k .$$

Under the data rate R, the channel can support kR bits of information transmitting from the coder to the decoder. Then, the coder can effectively divide this region into 2^{kR} disjoint and exhaustive pieces, each of which is shifted by the controller. As the Lebesgue measure is translation-invariant, it then follows that

$$m_k \geq \left(\frac{\prod_{|\lambda_i| \geq 1} |\lambda_i|}{2^R} \right)^k m_0.$$

To achieve the stability of the closed-loop system, we must require that $\lim_{k \to \infty} m_k = 0$. Thus, it follows that

$$\frac{\prod_{|\lambda_i| \geq 1} |\lambda_i|}{2^R} < 1. \tag{8.14}$$

Taking the logarithm of both sides of this inequality, we obtain the universal lower bound in (8.13), which is independent on the coder–decoder or controller.

8.5 Coder–Decoder Design

Since the classical quantized control approach does not consider the system dynamics, this results in several shortcomings. Here we need to design the coder–decoder with the consideration of the system dynamics. In fact, if the controller can obtain the exact initial state x_0, it can compute the system state at every time. Intuitively, we can design the coding–decoding scheme so that the controller is able to gradually learn the initial state x_0 by using the coded information s_k and the system dynamics.

To elaborate it, it is clear that

$$x_k = a^k x_0 - a^k z_k,$$

where $z_k = -a^{-k} \sum_{t=0}^{k-1} a^{k-t-1} u_t$ is regarded as the estimate of x_0 at time k. If the estimation error $n_k = x_0 - z_k$ is reduced at a rate strictly greater than $|a|$, that is, there exist $\eta > |a|$ and a constant α such that

$$|n_k| \leq \alpha / \eta^k,$$

then it follows that $|x_k| \leq \alpha \left(\frac{|a|}{\eta} \right)^k$, which in turn implies that $\lim_{k \to \infty} |x_k| = 0$.

In fact, the controller receives at most $\sum_{t=0}^{k-1} \log_2 |\mathcal{S}_t|$ bits of information at time k. In view of the information theory, the minimum estimation error on the initial state x_0 is given by

$$\eta_k = 2^{-\sum_{t=0}^{k-1} \log_2 |\mathcal{S}_t|}.$$

By (8.10) the estimation error is asymptotically reduced at the rate

$$2^{-\frac{1}{k}\sum_{t=0}^{k-1}\log_2|\mathcal{S}_t|} \approx 2^{-R},$$

so that the data rate R should satisfy that

$$2^R > |a|,$$

which leads to the data rate theorem in (8.13).

Now, we formally provide the coder–decoder to stabilize the linear system when the data rate satisfies the strict inequality in (8.13). Consider system (8.9) with $\|x_0\|_\infty \leqslant l_0$.[1] Given any $R > R_{\text{inf}}$, we only need to design a coder–decoder scheme that consumes a data rate less than R bits and the corresponding controller so that the closed-loop system is asymptotically stabilized.

Since system (8.9) is noiseless and the initial condition is bounded in a ball, we adopt a uniform quantizer. Specifically, if $-1 \leqslant x < 1$, then output of a uniform R-bit quantizer is given by

$$Q(x) = \frac{\lfloor 2^{R-1}x \rfloor + 0.5}{2^{R-1}},$$

and $Q(x) = 1 - 1/2^R$ for $x = 1$.

It is easy to verify that if $|x| \leqslant l$, then

$$|x - lQ(\frac{x}{l})| \leqslant \frac{l}{2^R}. \tag{8.15}$$

Clearly, there exists a transformation matrix $P \in \mathbb{R}^{n\times n}$ that satisfies the following expression:

$$PAP^{-1} = diag\{A_s, A_u\},$$

where all the eigenvalues of A_s are strictly in the unit circle, and any eigenvalue of A_u lies outside the unit circle. Obviously, the state variables corresponding to A_s asymptotically converge to 0 with zero inputs.

Thus we just need to consider the state variables corresponding to A_u. Without loss of generality, assume that all eigenvalues of A lie outside the unit circle, and thus the system (A, B, C) is controllable and observable. Then there exists a deadbeat observer such that the sensor and quantizer can obtain the complete state information of the system after time n.

[1] For the ease of presentation, we include this condition, which can be easily removed.

Hence we can assume that all the eigenvalues of A are out of the unit circle and $C = I$. Because the unstable poles of the system determine the uncertainty growth rate of their corresponding state variables, it makes sense to separate these state variables.

To this end, let all nonconjugate eigenvalues of A be $\lambda_1, \dots, \lambda_d$. If λ_i is complex, then its conjugate λ_i^* is not in this set. Denote the algebraic multiplicity of eigenvalue λ_i by m_i. Then we obtain that

$$\mu_i = \begin{cases} m_i & \text{if } \lambda_i \text{ is a real number,} \\ 2m_i & \text{if } \lambda_i \text{ is a complex number.} \end{cases} \tag{8.16}$$

Without loss of generality, let A be in the real Jordan form [18], that is,

$$A = diag\{J_1, \cdots, J_d\},$$

where $J_i \in \mathbb{R}^{\mu_i \times \mu_i}$ is the Jordan matrix corresponding to λ_i. Obviously, the eigenvalues of J_i are the same, and the corresponding state variables have the same growth rate as well. Moreover, we have the following lemma.

Lemma 8.1. *([19]) For any arbitrary natural number $q \in \mathbb{N}$, there exists a constant $\zeta > 0$ such that*

$$||J_i^q||_\infty \leq \zeta \sqrt{\mu_i} q^{\mu_i - 1} |\lambda_i|^q.$$

Remark 8.2. *According to Lemma 8.1, the growth rate of the Jordan matrix J_i is controlled by $|\lambda_i|$. In particular, when J_i is a diagonal matrix, then $||J_i^q||_\infty = |\lambda_i|^q$. This conclusion is critical to the subsequent design of the coder–decoder and the allocation of data rates.*

For an arbitrary date rate R, which satisfies

$$R > R_{\inf} = \sum_{i=1}^{d} \mu_i \log_2 |\lambda_i|,$$

there are positive integers α_i and β such that

$$\log_2 |\lambda_i| < \frac{\alpha_i}{\beta} \quad \text{and} \quad \frac{1}{\beta} \sum_{i=1}^{d} \mu_i \alpha_i \leq R. \tag{8.17}$$

In conformity with the structure of the Jordan matrix A, partition the state variables as

$$x_k = [(x_k^{(1)})^{\mathrm{T}}, \cdots, (x_k^{(d)})^{\mathrm{T}}]^{\mathrm{T}}.$$

Then d uniform quantizers, each of which has an average data rate α_i/β, are designed to quantize each state variables in $x_k^{(i)}$.

Next, we design the strategy of the coder/quantization, decoder, and control and provide the proof of the asymptotic stability of the closed-loop system.

Coder/quantizer. By Lemma 8.1 and (8.17), there is an integer q such that

$$\eta := \max_{i\in\{1,\ldots,d\}} \frac{\zeta\sqrt{\mu_i} q^{\mu_i-1}|\lambda_i|^q}{2^{q\alpha_i/\beta}} < 1. \tag{8.18}$$

Since the state estimate $\hat{x}_k$ is based on $s_0, \cdots, s_k$, the quantizer and decoder can obtain the same $\hat{x}_k$. For $i \in \{1, \cdots, d\}$, let $l_0^{(i)} = l_0$ and

$$l_{k+1}^{(i)} = \frac{l_k^{(i)}}{2^{q\alpha_i}}(\zeta\sqrt{\mu_i} q^{\mu_i-1}|\lambda_i|^q)^\beta. \tag{8.19}$$

By (8.18) and (8.19) we further obtain that

$$\lim_{k\to\infty} l_k^{(i)} \le l_0 \lim_{k\to\infty} \eta^{\beta k} = 0.$$

Let $\tau = q\beta$ and $\tilde{x}_{k\tau} = x_{k\tau} - \hat{x}_{k\tau}$. Consider the time intervals

$$\{k\tau, \ldots, (k+1)\tau - 1\},$$

that is, the length of each interval is τ. At the beginning of every time interval, $q\alpha_i$-bit uniform quantizers are utilized to quantize each component of $\tilde{x}_{k\tau}/l_k^{(i)}$ and get the quantized signal $s_k \in \mathbb{R}^n$. Therefore, the communication data rate of this protocol can be expressed as

$$\frac{1}{\tau}\sum_{i=1}^{d}(q\alpha_i)\mu_i = \frac{1}{\beta}\sum_{i=1}^{d}\alpha_i\mu_i.$$

By (8.17) the date rate is less than R.

Decoder/estimator. Based on the quantized signal s_k, the decoder utilizes the following algorithm to estimate the state x_k:

$$\begin{aligned}
\hat{x}_0 &= 0,\ L_0 = diag\{l_0^{(1)} I_{\mu_1}, \cdots, l_0^{(d)} I_{\mu_d}\},\\
\hat{x}_{k\tau+j} &= A\hat{x}_{k\tau+(j-1)} + Bu_{k\tau+(j-1)},\ 1 \le j \le \tau - 1,\\
L_k &= diag\{l_k^{(1)} I_{\mu_1}, \cdots, l_k^{(d)} I_{\mu_d}\},\\
\hat{x}_{(k+1)\tau} &= A^\tau(\hat{x}_{k\tau} + L_k s_k) + \sum_{j=k\tau}^{(k+1)\tau-1} A^{(k+1)\tau-j-1}Bu_j,
\end{aligned}$$

where $I_{\mu_i} \in \mathbb{R}^{\mu_i\times\mu_i}$ is the identity matrix.

Controller. Because the system (A, B) is controllable, there is a gain matrix K such that all the eigenvalues of the closed-loop matrix $A + BK$ are strictly in the unit circle. Let

$$u_k = K\hat{x}_k, \quad k \in \mathbb{N}.$$

Remark 8.3. *The quantized control method based on information-theoretic approach can stabilize the networked system (8.9) by linear feedback.*

Asymptotic stability. We first prove by induction that

$$\|\tilde{x}^{(i)}_{k\tau}\|_\infty \leqslant l^{(i)}_k.$$

Obviously, when $k = 0$, the inequality holds. Assume that for $t \le k$, we have $||\tilde{x}^{(i)}_{t\tau}||_\infty \le l^{(i)}_t$. By the estimation algorithm we obtain

$$\hat{x}_{(k+1)\tau} = A^\tau(\tilde{x}_{k\tau} - L_k s_k).$$

By inequality (8.15) it follows that

$$||\hat{x}^{(i)}_{(k+1)\tau}||_\infty \le \||\|_\infty|J^\tau_i||_\infty||x^{(i)}_{k\tau} - l^{(i)}_k s^{(i)}_k||_\infty \le \frac{||J^\tau_i||_\infty}{2^{q\alpha_i}} l^{(i)}_k. \tag{8.20}$$

Combining Lemma 8.1 and (8.19), we get that $||\tilde{x}^{(i)}_{(k+1)\tau}||_\infty \le l^{(i)}_{k+1}$. Therefore,

$$\lim_{k\to\infty} \tilde{x}_{k\tau} \le \lim_{k\to\infty} \max_{i\in\{1,\cdots,d\}} l^{(i)}_k = 0.$$

For $j \in \{1, \cdots, \tau - 1\}$, we obtain that $\tilde{x}_{k\tau+j} = A^j\tilde{x}_{k\tau}$. Furthermore, we know that $\lim_{k\to\infty} ||x_k||_\infty = 0$. Then the closed-loop system is

$$x_{k+1} = (A + BK)x_k - BK\tilde{x}_k.$$

Because all the eigenvalues of the closed-loop matrix $A + BK$ are in the unit circle, by the Toeplitz lemma [20] we obtain

$$\lim_{k\to\infty} x_k = 0.$$

8.6 Extension to Lossy Channels

The framework in the previous sections is generalized to noisy communication channels by many researchers. Due to the existence of channel uncertainties, the quantizer output s_k might not be exactly received by the decoder. This further induces information loss to the controller. To compensate this uncertainty, a greater data rate R is needed in comparison with the noiseless channels. We are particularly interested in the problem of how many additional bits are needed to achieve stabilization to counter the effects of channel uncertainties. Although the problem was initiated by Tatikonda and Mitter [12] in 2005, it is not fully understood for general vector linear systems to date. In [12], it is claimed that if the Shannon capacity of this channel is greater than R_{inf}, then the system with process disturbances can be almost surely stabilized with bounded error [12]. This is shown to be incorrect in [21], which shows that, on the contrary, any unstable linear system affected by arbitrarily and uniformly small external disturbances can never be almost surely stabilized via the erasure channel with nonzero erasure probability, irrespective of which algorithm of stabilization is employed. The almost sure stabilization is further investigated by the same authors [13,22,23].

There are a lot of uncertainties under the wireless network environment. For instance, the input signal s_k may drop out randomly because of the blocking and/or attenuation of channels, which means that it cannot be ensured that the channel sink can receive s_k, thus losing more information. We can imagine that we need a larger communication data rate to make up the loss due to uncertainties over noisy channels.

8.6.1 Erasure Channels

There exists random packet loss of s_k for erasure channels, which is supposed to be an independent and identically distributed Bernoulli process. Then, we have following results.

Theorem 8.2. *([11]) In the erasure channel environment, the networked linear scalar system (8.9), that is, $A = a$, is stabilizable via quantized feedback if and only if*

$$\mathbb{E}\left[\frac{|a|^2}{2^{2R\gamma_k}}\right] < 1. \tag{8.21}$$

Similarly to noiseless channels, the system dynamics renders the uncertainty in the form of mean square increases by $|a|^2$ every step. If there is no packet loss, that is. $\gamma_k = 1$, then the controller receives s_k and decreases the uncertainty by $1/2^{2R}$. If packet loss exists, that is, $\gamma_k = 0$, then the controller cannot receive s_k, and thereby it cannot decrease the uncertainty. The decreasing rate of uncertainty must be strictly greater than the increasing rate in the average sense to guarantee the stabilization of the system, meaning that inequality (8.21) holds.

Remark 8.4. *For systems with bounded noise such that* $\sup_{k\in\mathbb{N}} \max\{\|x_0\|, \|w_k\|, \|v_k\|\} < \infty$, *the authors in [24] designed an averaging quantizer to demonstrate the sufficiency of Theorem 8.2. In [25], the results are expanded to the single input vector systems. Further, in [11], Theorem 8.2 is proved for the random unbounded noise satisfying certain conditions.*

8.6.2 Gilbert–Elliott Channels

For Gilbert–Elliott channels γ_k, the packet loss process of s_k follows an ergodic Markov process [26]. Foregoing arguments cannot apply due to the correlation over time. To this end, in [27], the method of random oversampling is adopted.

Without loss of generality, let $\gamma_0 = 1$ and $t_0 = 0$. Define the random oversampling time point $\{t_k\}$ as the moment when the controller receive the channel input signal, that is, t_k satisfies

$$t_{k+1} = \inf\{j > t_k | \gamma_j = 1\}. \tag{8.22}$$

Denote the dwell time as $\tau_k = t_k - t_{k-1}$. Then it is clear that $\{\tau_k\}$ is an independent and identically distributed process according to the properties of a Markov process.

Theorem 8.3. *([27]) In the Gilbert–Elliott channel environment, networked linear scalar system (8.9), that is, $A = a$, is stabilizable via the quantized feedback if and only if*

$$\mathbb{E}\left[\frac{|a|^{2\tau_k}}{2^{2R}}\right] < 1. \tag{8.23}$$

Intuitively, the system dynamics makes the uncertainty in the form of mean square increase by $|a|^2$ during the period of τ_k. However, the controller receives the channel input signal only once within such a time interval and decreases the uncertainty by $1/2^{2R}$. The decreasing rate of uncertainty must be larger than the increasing rate in the average sense to guarantee the stabilization of the system, meaning that inequality (8.23) holds. Now we simply present the main idea of demonstration, the details of which are referred to [27].

Necessity. Via conditional entropy power in the information theory [14], we have $\mathbb{E}[\xi_k]$ as the lower bound of $\mathbb{E}[x_k^2]$, that is, $\mathbb{E}[x_k^2] \geqslant \mathbb{E}[\xi_k]$. Moreover, there exists a constant $\mu > 0$ satisfying

$$\xi_{k+1} = \frac{|a|^2}{2^{2R\gamma_k}}\xi_k + \mu.$$

Then, based on Markov jump linear system theory [28], it is easy to prove the necessity.

Sufficiency. The main difficulties are as follows.

- The correlation over time of γ_k. By [29] we observe that $\sup\limits_{k\in\mathbb{N}} \mathbb{E}[\|x_k\|^2] < \infty$ if and only if $\sup\limits_{k\in\mathbb{N}} \mathbb{E}[\|x_{t_k}\|^2] < \infty$. From the dynamics of the system we have

$$x_{t_k} = a^{\tau_k} x_{t_{k-1}} + p_k, \tag{8.24}$$

where p_k is the linear combination of control input u_k and disturbance w_k. Therefore, we only need to analyze the stabilization of the random oversampling system (8.24) driven by the independent and identically distributed process $\{\tau_k\}$.
- The unbounded random noise w_k and v_k of the system. Generally, the uniform quantizers are no longer applicable, where the adaptive quantizers in [11] are unutilized.
- The random loss of channel input signals. Divide the time axis into several parts, and the $(k+1)$th period is

$$\{t_{kq}, \ldots, t_{(k+1)q} - 1\},$$

where q is an integer to be determined. By the definition of t_k, the controller receives q data packets during every period.
Hence, a quantizer of qR bits is to be designed to quantify the state $x_{t_{kq}}$. Motivated by that idea, we obtain the following inequality:

$$\mathbb{E}[x^2_{t_{(k+1)q}}] \leqslant c_0(\mathbb{E}[\frac{|a|^{2\tau_1}}{2^{2R}}])^q \mathbb{E}[x^2_{t_{kq}}] + c_1, \tag{8.25}$$

where both c_0 and c_1 are positive constants. Select q large enough such that

$$c_0\left(\mathbb{E}[\frac{|a|^{2\tau_1}}{2^{2R}}]\right)^q < 1.$$

Then we have $\sup\limits_{k\in\mathbb{N}} \mathbb{E}[x^2_{t_{kq}}] < \infty$. Given that $q < \infty$, $\sup\limits_{k\in\mathbb{N}} \mathbb{E}[x^2_{t_k}] < \infty$. Finally, we obtain

$$\sup_{k\in\mathbb{N}} \mathbb{E}[x_k^2] < \infty.$$

8.7 Bibliographic Notes

The data rate theorem for stabilization of linear systems over perfect channels have been well established. Limited capacity channels in the classic communication theory are modeled in terms of not only quantization effects but also in terms of channel uncertainties and time delays. Many of the major results in this theory are developed on the ground of noisy channel models. So incorporating noisy digital channels into NCS problems seems to be unavoidable in the analysis and synthesis of NCSs. As an initial step, the issue of the minimum data rate for stabilizability of linear systems over *noisy* digital channels attracted good attention of researchers. The research results on this topic are not as fruitful as in the case with noiseless digital channels due to that the optimal data rate assignment among unstable system state variables intertwines with the channel uncertainty process and also depends on the sense of stabilization notion [11–13,21,24,25,27,30]. Nonetheless, some significant progress has been recently achieved toward this topic.

References

[1] A. Gersho, R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[2] D. Delchamps, Stabilizing a linear system with quantized state feedback, IEEE Transactions on Automatic Control 35 (1990) 916–924.

[3] R. Brockett, D. Liberzon, Quantized feedback stabilization of linear systems, IEEE Transactions on Automatic Control 45 (2000) 1279–1289.

[4] W. Wong, R. Brockett, Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback, IEEE Transactions on Automatic Control 44 (1999) 1049–1053.

[5] J. Baillieul, Feedback coding for information-based control: operating near the data-rate limit, in: Proc. 41st IEEE Conference on Decision and Control, 2002.

[6] G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates, SIAM Journal on Control and Optimization 43 (2004) 413–436.

[7] S. Tatikonda, S. Mitter, Control under communication constraints, IEEE Transactions on Automatic Control 49 (2004) 1056–1068.

[8] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M.I. Jordan, S.S. Sastry, Kalman filtering with intermittent observations, IEEE Transactions on Automatic Control 49 (2004) 1453–1464.

[9] M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses, Automatica 43 (2007) 598–607.

[10] L. Xie, L. Xie, Stability of a random Riccati equation with Markovian binary switching, IEEE Transactions on Automatic Control 53 (2008) 1759–1764.

[11] P. Minero, M. Franceschetti, S. Dey, G. Nair, Data rate theorem for stabilization over time-varying feedback channels, IEEE Transactions on Automatic Control 54 (2009) 243–255.

[12] S. Tatikonda, S. Mitter, Control over noisy channels, IEEE Transactions on Automatic Control 49 (2004) 1196–1201.

[13] A. Matveev, A. Savkin, An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels, SIAM Journal on Control and Optimization 46 (2007) 1323–1367.

[14] T. Cover, J. Thomas, Elements of Information Theory, Wiley-Interscience, 2006.

[15] R. Adler, A. Konheim, M. McAndrew, Topological entropy, Transactions of the American Mathematical Society 114 (1965) 309–319.

[16] K. Mahler, An application of Jensen's formula to polynomials, Mathematika 7 (1960) 98–100.

[17] N. Elia, When bode meets Shannon: control-oriented feedback communication schemes, IEEE Transactions on Automatic Control 49 (2004) 1477–1488.

[18] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2012.

[19] K. You, W. Su, M. Fu, L. Xie, Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization, Automatica 47 (2011) 170–176.

[20] R. Ash, C. Doléans-Dade, Probability and Measure Theory, Academic Press, 2000.

[21] A. Matveev, A. Savkin, Comments on "Control over noisy channels" and relevant negative results, IEEE Transactions on Automatic Control 50 (2005) 2105–2110.

[22] A. Matveev, A. Savkin, Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels, International Journal of Control 80 (2007) 241–255.

[23] A. Matveev, A. Savkin, Estimation and Control over Communication Networks, Springer, 2008.

[24] N. Martins, M. Dahleh, N. Elia, Feedback stabilization of uncertain systems in the presence of a direct link, IEEE Transactions on Automatic Control 51 (2006) 438–447.

[25] K. You, L. Xie, Minimum data rate for mean square stabilization of discrete LTI systems over lossy channels, IEEE Transactions on Automatic Control 55 (2010) 2373–2378.

[26] S. Meyn, R. Tweedie, J. Hibey, Markov Chains and Stochastic Stability, Springer-Verlag, London, 1996.

[27] K. You, L. Xie, Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses, IEEE Transactions on Automatic Control 56 (2011) 772–785.

[28] O. Costa, M. Fragoso, R. Marques, Discrete-Time Markov Jump Linear Systems, Springer, 2005.

[29] K. You, M. Fu, L. Xie, Mean square stability for Kalman filtering with Markovian packet losses, Automatica 47 (2011) 2647–2657.

[30] A. Sahai, S. Mitter, The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. Part I. Scalar systems, IEEE Transactions on Information Theory 52 (2006) 3369–3395.

CHAPTER 9

Distributed Control for Large-Scale NCSs

9.1 Introduction

An important research interest in LSSs today is the emphasis on distributed coordination due to the availability of ample processing power at low cost, which allows sensor data to be processed locally. Generally, the aim of distributed control is to coordinate a group of subsystems by implementing control policies locally. It is natural to suggest that distributed control can never be as good as centralized control. This holds only if there is no real-time limitation on the communication network, for example, delay-free and no packet losses. By accounting for the real network effect, distributed control may be better than centralized control in terms of robustness, scalability, security, and so on. A typical example of a distributed control system the focus of which is to coordinate a group of unmanned air vehicles (UAVs) to be a desired formation. Under this high-speed circumstance, the communication between UAVs becomes critical. The design procedure will impose a stringent requirement on simplifying the computational complexity on finding a controller with a distributed architecture.

Distributed coordination of multiple agents has broad applications in many areas including formation control [1,2], distributed sensor networks [3,4], flocking [5,6], distributed computation [7], and consensus of coupled chaotic oscillators [8,9]. Their common property is that each individual agent lacks global knowledge of the whole system and can only interact with its neighbors to achieve certain global behavior. Within this framework, communication graph (topology), which determines what information is available for each agent at each time instant, is an important aspect of information flow in distributed coordination. For example, to achieve an average consensus that requires the states of all agents to asymptotically converge to the average of their initial values, the communication graph must contain a spanning tree for a fixed topology [10,11], whereas for a switching topology, the union of the communication graphs should contain a spanning tree frequently enough as the system evolves [11–13]. In addition, the convergence rate to consensus directly relies on the second smallest eigenvalue of the graph Laplacian matrix [10,14].

A set of common and important research problems for multiagent systems focuses on how the agent dynamics and the interacting network topology affect their behavior. Recently, the emergence of NCSs has stimulated the research interest on multiagent systems. One of the interesting problems is the consensus of multiagent systems, which requires all networked agents to reach an agreement on quantity of common interest using the shared data through

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00009-9

local communications. Toward this objective, a key step is to design a network-based control protocol such that as time goes on, all the agents asymptotically reach consensus. We discuss this problem in this chapter.

9.2 Consensus of Multiagent Systems

9.2.1 Communication Graph

Let $\mathcal{V} = \{v_1, \dots, v_N\}$ be an index set of N agents with i representing the ith agent. A digraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{A}\}$ will be utilized to model the interactions among agents, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set of paired agents, and $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N\times N}$ with nonnegative elements is the weighted adjacency matrix of $\mathcal{G}$. Self-edges (i,i) are not allowed, that is, $(i,i) \notin \mathcal{E}$ for all $i \in \mathcal{N}$. An edge $(j,i) \in \mathcal{E}$ if and only if $a_{ij} > 0$, which means that agent j can send information to agent i.

A sequence of edges $(i_1, i_2), (i_2, i_3), \dots, (i_{k-1}, i_k)$ with $(i_{j-1}, i_j) \in \mathcal{E}$ for all $j \in \{2, \dots, k\}$ is called a directed path from agent i_1 to agent i_k. The digraph $\mathcal{G}$ contains a spinning tree if there is a root agent that can send information to all the other agents via directed paths. It is called a strongly connected digraph if for any two agents $i, j \in \mathcal{V}$, there exists a directed path from agent i to agent j. If $\mathcal{A}$ is a symmetric matrix, $\mathcal{G}$ is called an *undirected* graph. A strongly connected undirected graph is simply called a connected graph. For an *undirected* graph $\mathcal{G}$, it is clear that $\mathcal{G}$ contains a spanning tree if and only if $\mathcal{G}$ is connected. A digraph is called *complete* if each pair of agents can directly connect to each other, that is, $(i, j) \in \mathcal{E}$ for all $i \neq j$. The neighborhood of the ith agent is denoted by $\mathcal{N}_i \triangleq \{j | (j,i) \in \mathcal{E}\}$. The in-degree of agent i is represented by $\deg_i = \sum_{j=1}^{N} a_{ij}$. Denote $\mathcal{D} \triangleq diag(\deg_1, \dots, \deg_N)$ and the Laplacian matrix of $\mathcal{G}$ by $\mathcal{L}_{\mathcal{G}} = \mathcal{D} - \mathcal{A}$. The eigenvalues of $\mathcal{L}_{\mathcal{G}}$ are denoted by $\lambda_j \in \mathbb{C}$, $j \in \mathcal{N}$, and written in an ascending order in magnitude as $0 = |\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_N|$. Note that for an undirected graph $\mathcal{G}$, $\mathcal{L}_{\mathcal{G}}$ is a symmetric positive semidefinite matrix, and $\lambda_j \geq 0$ for all $j \in \mathcal{N}$ [15].

Lemma 9.1. *[16] Let the adjacency matrix $\mathcal{A} \in \mathbb{R}^{N\times N}$ of an undirected graph $\mathcal{G}$ be a symmetric $(0,1)$-matrix, that is, $a_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. Then $\mathcal{G}$ is complete if and only if $\mathcal{G}$ is connected and $\lambda_2 = \lambda_N$.*

Lemma 9.2. *[11,15] Let $\mathcal{G}$ be a digraph. Then all the nonzero eigenvalues of $\mathcal{L}_{\mathcal{G}}$ are in the open right half-plane. Moreover, $\mathcal{G}$ has a spanning tree if and only if $\mathcal{L}_{\mathcal{G}}$ contains exactly one zero eigenvalue.*

9.2.2 Consensus of Multiagent Systems

A multiagent system is a large-scale system consisting of multiple dynamical agents. Let the dynamics of agent i in discrete time take the following form:

$$\begin{cases} x_i(k+1) &= Ax_i(k) + Bu_i(k), \\ y_i(k) &= Cx_i(k), \end{cases} \quad \forall i \in \mathcal{V}, k \in \mathbb{N}, \tag{9.1}$$

where $x_i(k) \in \mathbb{R}^n$, $u_i(k) \in \mathbb{R}$, and $y_i(k) \in \mathbb{R}^m$ represent the state, control input, and output of agent i at the time step k, respectively, and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^n$ are the state and input matrices.

By adapting the available information for each agent that is subject to the graph information flow constraint, we say that a control protocol is *distributed* if each agent generates its control input signal by relying on relative outputs. Generally, *distributed* control protocols can be categorized depending on whether they are *static* or *dynamic*. We are interested in the design of the distributed control protocol to reach an agreement among agents in the following sense.

Definition 9.1. *The discrete-time multiagent systems (9.1) are said to reach* consensus *if for any finite $x_i(0)$, $i \in \mathcal{N}$, there exists a distributed control protocol such that*

$$\lim_{k\to\infty} \|x_i(k) - x_j(k)\| = 0, \ \forall i, j \in \mathcal{N}. \tag{9.2}$$

For a stable A, it is clear that the zero input $u_i(k) = 0$ can achieve consensus. To make the problem interesting, we focus on an unstable A.

By an appropriate coordinate transformation there is no loss of generality to assume that $A = diag\{A_s, A_u\}$, where $A_s \in \mathbb{R}^{n_1\times n_1}$ and $A_u \in \mathbb{R}^{n_2\times n_2}$ correspond to the stable and unstable (including marginally stable) parts of A, respectively. Since the stable part renders the associated state subspace of each agent converge to zero, in this chapter, we make the following assumption:

A1) All the eigenvalues of A lie on or outside the unit circle.

9.3 Consensus Control With Relative State Feedback

In this section, we consider the following distributed protocol with relative state feedback:

$$u_i(k) = K \sum_{j=1}^{N} a_{ij}(x_j(k) - x_i(k)), \ k = 0, 1, \ldots, \tag{9.3}$$

where $K \in \mathbb{R}^{1\times n}$ is a fixed control gain independent of the agent index i.

9.3.1 Design of Consensus Gain

We start with *undirected* graphs where the eigenvalues of the associated Laplacian matrix are nonnegative, for example, $\lambda_j \geq 0$ for all $j \in \mathcal{N}$. Extensions to *directed* graphs (digraphs) are delivered in the next subsection.

Theorem 9.1. *Given a fixed* undirected *graph $\mathcal{G}$, under* **A**1*), necessary and sufficient conditions for the discrete-time multiagent systems (9.1) to reach consensus under protocol (9.3) are the following:*

(a) *(A, B) is a controllable pair;*

(b) *Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the strict inequality*

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}, \tag{9.4}$$

where $\lambda_j^u(A)$ represents an unstable eigenvalue of A, and λ_2 and λ_N are respectively the second smallest and largest eigenvalues of $\mathcal{L}_{\mathcal{G}}$.

Moreover, if these conditions hold, let ζ be such that

$$\prod_j |\lambda_j^u(A)| < \zeta^{-1} \leq \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}.$$

Then, the control gain

$$K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B}$$

solves the consensus problem, where $P > 0$ is a positive solution to the modified algebraic Riccati inequality

$$P - A^T P A + (1 - \zeta^2) \frac{A^T P B B^T P A}{B^T P B} > 0. \tag{9.5}$$

Remark 9.1. *1. The existence of a positive solution P to (9.5) is proved in [17,18]. Here λ_2/λ_N is called the* eigenratio *of an* undirected *graph. By Lemmas A.1–A.2 in [19] we immediately obtain an upper bound of the eigenratio:*

$$\frac{\lambda_2}{\lambda_N} \leq \frac{\min_i \deg_i}{\max_i \deg_i}.$$

2. *For the average consensus problem in [10], the state of each agent is scalar, and* $A = B = K = 1$. *The condition in item (a) of Theorem 9.1 is automatically satisfied, whereas inequality (9.4) implies that* $\lambda_2 > 0$. *By Lemma 9.2 the communication graph has to be connected, which is consistent with the result in [10]. Thus, our result contains the classical average consensus as a particular case.*
3. *Inequality (9.4) implies that* $\lambda_2 > 0$. *Then the graph is connected. In contrast with the result on continuous-time systems in [20], the case of discrete-time systems has an additional constraint given in (9.4). The eigenratio* λ_2/λ_N *of a Laplacian matrix is an important factor [9]. A larger eigenratio corresponds to a better synchronizability of the underlying communication graph. Intuitively, a better network synchronizability allows a more unstable* A *to achieve consensus of the multiagent systems and vice versa, which are confirmed by our result.*

   For a continuous-time system under a sufficiently small sampling period, the unstable eigenvalues of the discretized system (9.1) can be made arbitrarily close to one, and thus inequality (9.4) will be eventually satisfied for any connected undirected graph. Thus, for the case of continuous-time agent dynamics, our result is consistent with that in [20]. In fact, for a continuous-time system, information can be transmitted arbitrarily fast so that the network synchronizability of the communication graph becomes less important for achieving consensus.
4. *If the adjacency matrix* $\mathcal{A}$ *of the graph* $\mathcal{G}$ *is selected as a symmetric* $(0, 1)$*-matrix, then the eigenratio* $\lambda_2/\lambda_N \to 1$ *means that the communication graph tends to be complete (cf. Lemma 9.1). In this case, the controller can be designed in an almost centralized fashion. Then consensus can be achieved for any stabilizable system.*
5. *The convergence rate of the average consensus over an undirected graph is determined by* λ_2 *[10,14]. By the Courant–Weyl interlacing inequalities [21], adding an undirected edge to an undirected incomplete graph* $\mathcal{G}$ *will never decrease* λ_2*, suggesting that the consensus performance will not deteriorate. However, adding an undirected edge to a graph may lead to a smaller eigenratio. For example, consider the following two graph Laplacian matrices:*

   $$\mathcal{L}_{\mathcal{G}_1} = \begin{bmatrix} 3 & -1 & 0 & 0 & -1 & -1 \\ -1 & 3 & -1 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & 0 & -1 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 0 & -1 & 3 \end{bmatrix} \text{ and } \mathcal{L}_{\mathcal{G}_2} = \begin{bmatrix} 4 & -1 & -1 & 0 & -1 & -1 \\ -1 & 3 & -1 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & 0 & -1 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 0 & -1 & 3 \end{bmatrix}.$$

   It is clear that $\mathcal{G}_2$ *with an eigenratio* 0.3970 *is formed by adding an undirected edge to* $\mathcal{G}_1$*, whose eigenratio is* 0.4. *Thus, it is possible to lose consensus of the multiagent systems (9.1) under protocol (9.3) by adding an edge. It appears to be counter-intuitive since* *the communication graph with a "better" connectivity may result in a worse consensus capability. Note that whether the eigenratio will increase or decrease by adding an edge is not conclusive; see [8] for more detail.*

6. *The importance of the intrinsic entropy rate of a linear dynamical system, quantified by $\sum_j \log_2 |\lambda_j^u(A)|$, has been widely recognized in networked control systems (see, e.g., [17, 22–26]) as it determines the minimum data rate for stabilization. Here the intrinsic entropy rate of the agent dynamics is first shown to pose a fundamental limitation on the eigenratio of an undirected graph for consensus.*

The proof of Theorem 9.1 depends on the following lemma, which gives a necessary and sufficient condition for a class of simultaneous stabilization problems for discrete-time systems.

Lemma 9.3. *Given $0 < \lambda_2 \leq \ldots \leq \lambda_N$, and under* **A***1), necessary and sufficient conditions for the existence of a common control gain $K \in \mathbb{R}^{1\times n}$ such that $\rho(A - \lambda_j BK) < 1$ for all $j \in \{2, \ldots, N\}$ is the following:*

(a) *(A, B) is controllable;*
(b) *The product of unstable eigenvalues of A is strictly upper bounded as follows:*

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}. \tag{9.6}$$

Proof. By the convention $\frac{2}{0} = \infty$ it is obvious that only the case with $\lambda_2/\lambda_N \neq 1$ needs to be elaborated.

Necessity: Under **A**1), it is straightforward that (A, B) is controllable. Without loss of generality (w.l.o.g.), assume that (A, B) is already in the controllable canonical form:

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots \\ \vdots & & \ddots & \ddots \\ 0 & \ldots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \ldots & -\alpha_{n_2-1} \end{bmatrix}; \ B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{9.7}$$

where $\alpha_0 = \prod_j \lambda_j^u(A)$. Let $K = [-k_0, -k_1, \ldots, -k_{n-1}]$ simultaneously stabilize $(A, \lambda_j B)$. It is obvious that

$$\det(zI_n - A + \lambda_j BK) = z^n + (\alpha_{n-1} - \lambda_j k_{n-1})z^{n-1} + \cdots + (\alpha_0 - \lambda_j k_0). \tag{9.8}$$

Since all the eigenvalues of $A - \lambda_j BK$ are within the unit disk, it follows from (9.8) that, for all $j \in \{2, \ldots, N\}$,

$$|\alpha_0 - \lambda_j k_0| < 1 \Rightarrow \frac{|\alpha_0| - 1}{\lambda_j} < |k_0| < \frac{|\alpha_0| + 1}{\lambda_j}. \tag{9.9}$$

Thus, we obtain that

$$\bigcap_{j=2}^{N}\left(\frac{|\alpha_0|-1}{\lambda_j},\frac{|\alpha_0|+1}{\lambda_j}\right)\neq\emptyset,$$

which further implies that $\frac{|\alpha_0|-1}{\lambda_2}<\frac{|\alpha_0|+1}{\lambda_N}$. Noting that $|\alpha_0|=\prod_j|\lambda_j^u(A)|$, the necessity follows directly.

Sufficiency: Select ζ such that

$$\prod_j|\lambda_j^u(A)|<\zeta^{-1}\leq\frac{1+\lambda_2/\lambda_N}{1-\lambda_2/\lambda_N}$$

and let $\zeta_j=1-\frac{2\lambda_j}{\lambda_2+\lambda_N}\leq\zeta$ for all $j\in\{2,\ldots,N\}$. Since (A,B) is controllable, there exists a positive definite solution P to the modified algebraic Riccati inequality (9.5) [17].

Letting the control gain be

$$K=\frac{2}{\lambda_2+\lambda_N}\frac{B^TPA}{B^TPB},$$

it follows that

$$\begin{aligned}(A-\lambda_jBK)^TP(A-\lambda_jBK)-P &= A^TPA-(1-\zeta_j^2)\frac{A^TPBB^TPA}{B^TPB}-P\\ &\leq A^TPA-(1-\zeta^2)\frac{A^TPBB^TPA}{B^TPB}-P\\ &< 0,\end{aligned}$$

which completes the proof. □

Proof of Theorem 9.9. Denote the average state of all agents by

$$\bar{x}(k)\triangleq\frac{1}{N}\sum_{i=1}^{N}x_i(k)=\frac{1}{N}(\mathbf{1}^T\otimes I_n)x(k)$$

and the deviation of each agent from the average state by $\delta_i(k)\triangleq x_i(k)-\bar{x}(k)$, where $\mathbf{1}$ is a vector of compatible dimension with all elements equal to one.

By the definition of consensus this yields that

$$\lim_{k\to\infty}\|\delta_i(k)\|\leq\frac{1}{N}\sum_{j=1}^{N}\lim_{k\to\infty}\|x_i(k)-x_j(k)\|=0.$$

Conversely, $\lim_{k\to\infty}\|\delta_i(k)\| = 0$ for all $i \in \mathcal{N}$ immediately implies consensus of the multiagent systems (9.1). Thus consensus is equivalent to $\lim_{k\to\infty}\|\delta_i(k)\| = 0$ for all $i \in \mathcal{N}$.

Stack x_j to get a new state vector $x(k) = [x_1^T(k), \dots, x_N^T(k)]^T$. By (9.3) the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)x(k). \tag{9.10}$$

Noting that $\mathbf{1}^T\mathcal{L}_{\mathcal{G}} = \mathbf{0}^T$, the following equalities are in force:

$$\begin{aligned} \bar{x}(k+1) &= \frac{1}{N}(\mathbf{1}^T \otimes A)x(k) - \frac{1}{N}(\mathbf{1}^T\mathcal{L}_{\mathcal{G}} \otimes BK)x(k) \\ &= A\bar{x}(k). \end{aligned} \tag{9.11}$$

Let $\delta(k) = [\delta_1^T(k), \dots, \delta_N^T(k)]^T$. Subtracting (9.10) from (9.11) immediately leads to

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k). \tag{9.12}$$

Select $\phi_i \in \mathbb{R}^N$ such that $\phi_i^T\mathcal{L}_{\mathcal{G}} = \lambda_i\phi_i^T$ and form the unitary matrix

$$\Phi = [\frac{\mathbf{1}}{\sqrt{N}}, \phi_2, \dots, \phi_N]$$

to transform $\mathcal{L}_{\mathcal{G}}$ into a diagonal form:

$$diag(0, \lambda_2, \dots, \lambda_N) = \Phi^T\mathcal{L}_{\mathcal{G}}\Phi. \tag{9.13}$$

Using the property of Kronecker product gives that

$$(\Phi \otimes I_n)^T(I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)(\Phi \otimes I_n) = diag(A, A - \lambda_2 BK, \dots, A - \lambda_N BK). \tag{9.14}$$

Denote $\widetilde{\delta}(k) = (\Phi \otimes I_n)^T\delta(k)$ and partition $\widetilde{\delta}(k) \in \mathbb{R}^{nN}$ into two parts, that is, $\widetilde{\delta}(k) = [\widetilde{\delta}_1^T(k), \widetilde{\delta}_2^T(k)]^T$, where $\widetilde{\delta}_1(k) \in \mathbb{R}^n$ is the vector consisting of the first n elements of $\widetilde{\delta}(k)$.

Then $\widetilde{\delta}_1(k) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \delta_i(k) = \mathbf{0}$. In view of (9.12) and (9.14), this yields that

$$\widetilde{\delta}_2(k+1) = diag(A - \lambda_2 BK, \dots, A - \lambda_N BK)\widetilde{\delta}_2(k). \tag{9.15}$$

Since $\Phi \otimes I_n$ is nonsingular, $\lim_{k\to\infty}\|\delta(k)\| = 0$ is equivalent to $\lim_{k\to\infty}\|\widetilde{\delta}_2(k)\| = 0$.

Necessity: By (9.15) it follows that $\rho(A - \lambda_i BK) < 1$ for $i \in \{2, \dots, N\}$, which in turn implies that $\lambda_2 > 0$ since if $\lambda_2 = 0$, then $\rho(A - \lambda_2 BK) \geq 1$ for all $K \in \mathbb{R}^{1\times n}$ by **A**1); that is, $(A, \lambda_j B)$ can be simultaneously stabilized by a common control gain $K \in \mathbb{R}^{1\times n}$. In light of Lemma 9.3, the necessity is established.

Sufficiency: Under **A**1), inequality (9.4) implies that $\lambda_2 > 0$. By Lemma 9.3 the common gain

$$K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B},$$

where P is a positive definite solution to the modified algebraic Riccati inequality (9.5), can simultaneously stabilize $(A, \lambda_i B)$, $i \in \{2, \dots, N\}$, thats is, $\rho(A - \lambda_i BK) < 1$. Together with (9.15), the proof of sufficiency is completed. □

In the proof of sufficiency, we have constructed a specific control gain K for the multiagent systems to reach a consensus, which is obtained by solving a modified algebraic Riccati inequality. It is interesting to note that the discrete-time consensus problem over an *undirected* graph is closely related to the robust stabilization problem with a bounded uncertainty in the input gain. Here the variation of the eigenvalues of the Laplacian matrix is interpreted as parameter uncertainty in the input. Differently from the classical robust stabilization problem, the uncertainty in this case takes only a finite number of positive values. What is particularly surprising is that the necessary and sufficient condition for the classical robust stabilization continues to hold.

9.3.2 Extensions to Digraphs

In the previous subsection, the consensus problem over *undirected* graphs is converted to a simultaneous stabilization problem. Similarly, we can easily show that the consensus problem over *directed* graphs is still equivalent to a simultaneous stabilization problem, that is, find a common gain such that $\rho(A - \lambda_j BK) < 1$ for all $j \in \{2, \dots, N\}$. The main difference is that, under *directed* graphs, the nonzero eigenvalues of the induced Laplacian matrix, λ_j, $j \in \{2, \dots, N\}$, are not real numbers in general. However, a simple necessary and sufficient condition for consensus of multiagent systems (9.1) under protocol (9.3) can be also established.

Theorem 9.2. *Given a fixed* directed *graph $\mathcal{G}$, under **A**1), necessary and sufficient conditions for the discrete-time multiagent systems (9.1) to reach consensus under protocol (9.3) are the following:*

(a) (A, B) is a controllable pair;

(b) Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the strict inequality

$$\prod_j |\lambda_j^u(A)| < \frac{1}{\min\limits_{\omega \in \mathbb{R}} \max\limits_{j \in \{2,\dots,N\}} |1 - \omega \lambda_j|}, \tag{9.16}$$

where λ_j is a complex eigenvalue of $\mathcal{L}_{\mathcal{G}}$.

Moreover, under these conditions, let ω^* be a solution to (9.16). Select ζ such that

$$\frac{1}{\prod_j |\lambda_j^u(A)|} > \zeta \geq \max_{j \in \{2,\ldots,N\}} |1 - \omega^* \lambda_j|.$$

Then, the control gain

$$K = \omega^* (B^T P B)^{-1} B^T P A$$

solves the consensus problem, where P is a positive definite solution to the modified algebraic Riccati inequality (9.5).

Proof. In view of the proof of Theorem 9.1, we only need to find a necessary and sufficient condition for the existence of a common control gain $K \in \mathbb{R}^{1\times n}$ such that $\rho(A - \lambda_j BK) < 1$ where $\lambda_j \in \mathbb{C}$ for all $j \in \{2, \ldots, N\}$.

Necessity: It is trivial that (A, B) is controllable. Next, it follows from (9.9) that

$$|\alpha_0 - \lambda_j k_0| < 1 \Rightarrow |1 - \lambda_j k_0'| < \frac{1}{|\alpha_0|}, \ j \in \{2, \ldots, N\},$$

where $k_0' = k_0/|\alpha_0| \in \mathbb{R}$. This implies that

$$\inf_{\omega \in \mathbb{R}} \max_{j \in \{2,\ldots,N\}} |1 - \omega \lambda_j| < \frac{1}{\prod_j |\lambda_j^u(A)|}.$$

Note that $|x|$ is continuous w.r.t. $x \in \mathbb{C}$, and thus the inf in the last inequality is achievable.

Sufficiency: Let ω^* be a solution to (9.16) and denote $\zeta_j = 1 - \omega^* \lambda_j$. Then, $|\zeta_j| \leq \zeta$ for all $j \in \{2, \ldots, N\}$. Using the proposed control gain K, we obtain that

$$\begin{aligned}(A - \lambda_j BK)^H P (A - \lambda_j BK) - P &= A^T P A - (1 - |\zeta_j|^2) \frac{A^T P B B^T P A}{B^T P B} - P \\ &\leq A^T P A - (1 - \zeta^2) \frac{A^T P B B^T P A}{B^T P B} - P \\ &< 0\end{aligned}$$

for all $j \in \{2, \ldots, N\}$. □

Condition (9.16) can be readily checked via the following lemma.

Lemma 9.4. *Let $\lambda_j = r_j \exp(\theta_j \iota)$ with $\iota^2 = -1$ and $\Delta_m = 1/(\prod |\lambda_i^u(A)|)$. Then inequality (9.16) holds if and only if the intersection*

$$\bigcap_{j=2}^{N}\left(\frac{\cos\theta_j-\sqrt{\Delta_m^2-\sin^2\theta_j}}{r_j},\frac{\cos\theta_j+\sqrt{\Delta_m^2-\sin^2\theta_j}}{r_j}\right) \tag{9.17}$$

is not empty, which is equivalent to that

$$\frac{1-\Delta_m^2}{\min_{j\in\{2,\ldots,N\}} r_j f(\theta_j)} < \min_{j\in\{2,\ldots,N\}}\frac{f(\theta_j)}{r_j}. \tag{9.18}$$

Here $f(\theta)=\cos\theta+\sqrt{\Delta_m^2-\sin^2\theta}$ *is a decreasing function w.r.t.* $\theta\in(0,\arcsin(\Delta_m))$*, where* $\arcsin(x)$ *is the inverse sine of* x.

Proof. We can easily verify that inequality (9.16) holds if and only if $\cap_{j=2}^{N}\{\omega\in\mathbb{R}||1-\omega\lambda_j|<\Delta_m\}\neq\emptyset$, which is equivalent to (9.17). The equivalence of (9.17) and (9.18) is trivial. □

Remark 9.2. *By Lemma 9.2 all the nonzero eigenvalues of* $\mathcal{L}_\mathcal{G}$ *lie in the open right half-plane, which implies that* $-\pi/2\le\theta_j\le\pi/2$*. Then it follows from (9.16) that*

$$|\theta_j|<\arcsin\left(\frac{1}{\prod_i|\lambda_i^u(A)|}\right),\ \forall j\in\{2,\ldots,N\}. \tag{9.19}$$

This means that the more unstable is the open-loop matrix, a stronger condition is required on the directed *graph to achieve consensus under protocol (9.3). Since* $|1-\omega\lambda_j|\ge|1-|\omega|\cdot|\lambda_j||$ *for all* $\omega\in\mathbb{R}$ *and* $j\in\{2,\ldots,N\}$*, it follows that*

$$\min_{\omega\in\mathbb{R}}\max_{j\in\{2,\ldots,N\}}|1-|\omega|\cdot|\lambda_j||\le\min_{\omega\in\mathbb{R}}\max_{j\in\{2,\ldots,N\}}|1-\omega\lambda_j|.$$

Thus, we can easily derive from (9.16) that

$$\prod_j|\lambda_j^u(A)|<\frac{|\lambda_N|+|\lambda_2|}{|\lambda_N|-|\lambda_2|}. \tag{9.20}$$

By Theorem 9.1 we know that, under undirected *graphs, the necessary condition of (9.20) on graphs is also sufficient for reaching a consensus. Unfortunately, this condition is no longer strong enough for achieving consensus of the multiagent systems under protocol (9.3) if the communication graph is* directed.

Remark 9.3. *We further only consider* undirected *graphs. The results in the rest of this section can be easily generalized to* directed *graphs.*

9.3.3 Performance Analysis

Under protocol (9.3), we can evaluate the corresponding *asymptotic convergence factor* [27] defined by

$$r_{\text{asym}} = \sup_{\delta(0)\neq \mathbf{0}} \lim_{k\to\infty} \left(\frac{\|\delta(k)\|}{\|\delta(0)\|} \right)^{\frac{1}{k}}. \tag{9.21}$$

Similarly, another measure of the speed of convergence is the *per-step convergence factor* [27] defined as

$$r_{\text{step}} = \sup_{\delta(k)\neq \mathbf{0}} \frac{\|\delta(k+1)\|}{\|\delta(k)\|}. \tag{9.22}$$

Denote the set of stabilizing gains of $(A, \lambda_j B)$ by

$$\Gamma_j = \{K \in \mathbb{R}^{1\times n} | \, \rho(A - \lambda_j BK) < 1\}. \tag{9.23}$$

Given any control gain $K \in \mathbb{R}^{1\times n}$, define

$$J(K) = diag(A - \lambda_2 BK, \ldots, A - \lambda_N BK). \tag{9.24}$$

The convergence speed is quantified below.

Corollary 9.1. *Given an* undirected *graph $\mathcal{G}$, under the conditions in Theorem 9.9, select $K \in \bigcap_{j=2}^{N} \Gamma_j$, where Γ_j is defined in (9.23). Then* the asymptotic convergence factor *and* per-step convergence factor *for consensus are respectively evaluated by*

$$r_{\text{asym}} = \rho(J(K)) \textit{ and } \; r_{\text{step}} = \|J(K)\|. \tag{9.25}$$

Proof. Since $\Phi \otimes I_n$ is an unitary matrix, it is obvious that $\|\delta(k)\| = \|\widetilde{\delta}(k)\|$ for $k \in \mathbb{N}$. Observe that if $\widetilde{\delta}_1(k) = \mathbf{0}$, then $\|\widetilde{\delta}(k)\| = \|\widetilde{\delta}_2(k)\|$ for all $k \in \mathbb{N}$. In view of (9.15), it follows that

$$\begin{aligned} r_{\text{asym}} &= \sup_{\widetilde{\delta}_2(0)\neq \mathbf{0}} \lim_{k\to\infty} \left(\frac{\|J^k(K)\widetilde{\delta}_2(0)\|}{\|\widetilde{\delta}_2(0)\|} \right)^{\frac{1}{k}} \\ &\leq \lim_{k\to\infty} \|J^k(K)\|^{\frac{1}{k}} \\ &= \rho(J(K)). \end{aligned} \tag{9.26}$$

On the other hand, select $\widetilde{\delta}_2(0)$ as an eigenvector of $J(K)$ corresponding to the largest eigenvalue, that is, $J(K)\widetilde{\delta}_2(0) = \lambda_M \widetilde{\delta}_2(0)$ and $|\lambda_M| = \rho(J(K))$. Then we have the following results:

$$r_{\text{asym}} \geq \lim_{k\to\infty} \left(\frac{\|J^k(K)\widetilde{\delta}_2(0)\|}{\|\widetilde{\delta}_2(0)\|} \right)^{\frac{1}{k}}$$

$$
\begin{aligned}
&= \lim_{k\to\infty} \left(\frac{\|\lambda_M^k \widetilde{\delta}_2(0)\|}{\|\widetilde{\delta}_2(0)\|} \right)^{\frac{1}{k}} \\
&= \rho(J(K)). \qquad (9.27)
\end{aligned}
$$

Hence this yields that $r_{\text{asym}} = \rho(J(K))$. The second part can be shown similarly. □

To make the convergence to consensus as fast as possible, a control gain $K \in \bigcap_{j=2}^{N} \Gamma_j$ should be selected to minimize the *asymptotic convergence factor* or *per-step convergence factor*. Since the spectral radius of a square matrix is not a convex function, not even Lipschitz continuous, the problem of finding an optimal K to minimize the *asymptotic convergence factor* is in general very difficult. However, we can derive a lower bound for the optimal *asymptotic convergence factor*.

Theorem 9.3. *Given an* undirected *graph $\mathcal{G}$, under the conditions in Theorem 9.1, the optimal* asymptotic convergence factor *is lower bounded as follows:*

$$
\begin{aligned}
r_{\text{asym}}^* &= \inf_{K\in\bigcap_{j=2}^{N}\Gamma_j} \rho(J(K)) \\
&\geq \Big(\prod_j |\lambda_j^u(A)|\Big)^{1/n} \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/n}. \qquad (9.28)
\end{aligned}
$$

Proof. It essentially follows from the necessity of Lemma 9.3 with some modifications. Without loss of generality, let (A, B) be given in the controllable canonical form. For any control gain $K \in \bigcap_{j=2}^{N} \Gamma_j$, it follows from the definition of $J(K)$ in (9.24) that $\rho(A - \lambda_j BK) \leq \rho(J(K)) < 1$ for all $j \in \{2, \dots, N\}$.

In consideration of (9.8) and letting $K = [-k_0, \dots, -k_{n-1}]$, we obtain that

$$
\begin{aligned}
&|\det(A) - \lambda_j k_0| \leq \rho(J(K))^n \\
&\Rightarrow \frac{|\det(A)| - \rho(J(K))^n}{\lambda_j} \leq |k_0| \leq \frac{|\det(A)| + \rho(J(K))^n}{\lambda_j}.
\end{aligned}
$$

Using the arguments in the necessity of Lemma 9.3, we can show that

$$
\frac{|\det(A)| - \rho(J(K))^n}{\lambda_2} \leq \frac{|\det(A)| + \rho(J(K))^n}{\lambda_N}.
$$

By simple algebraic manipulations we obtain that

$$
\rho(J(K)) \geq |\det(A)|^{1/n} \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/n}, \ \forall K \in \bigcap_{j=2}^{N} \Gamma_j.
$$

Taking the infinitum on both sides of this inequality completes the proof. □

Remark 9.4. *The lower bound in (9.28) is attainable for some particular cases:*

1. *The interaction graph is complete and the adjacency matrix $\mathcal{A}$ is a symmetric (0,1)-matrix. Lemma 9.2 implies that $\lambda_2 = \lambda_N > 0$. Since $(A, \lambda_2 B)$ is controllable, there exists a control gain K^* such that $\rho(A - \lambda_2 BK^*) = 0$. Thus $r_{\text{asym}} = \rho(J(K^*)) = 0$, and the consensus can be achieved in finite time.*
2. *The agent dynamic is an unstable scalar system, and the communication graph is* undirected, *that is, $x_i(k+1) = ax_i(k) + bu_i(k)$, where $a \geq 1$ and $b \neq 0$. Let $k^* = \frac{2a}{b(\lambda_2+\lambda_N)}$. Then*

$$a - \lambda_2 bk^* = \frac{a(1-\lambda_2/\lambda_N)}{1+\lambda_2/\lambda_N}$$

   and

$$|a - \lambda_j bk^*| \leq \frac{a(1-\lambda_2/\lambda_N)}{1+\lambda_2/\lambda_N}, \ \forall j \in \{2, \ldots, N\}.$$

   Thus, $r^*_{\text{asym}} = \rho(J(k^*)) = \frac{a(1-\lambda_2/\lambda_N)}{1+\lambda_2/\lambda_N}$.
3. *It can also be approached for the discrete-time second-order consensus with an* undirected *graph (see Section 9.3.4) for the design of an optimal control gain to reach this bound.*

9.3.4 Optimal Consensus Control for Second-Order Systems

In this subsection, an optimal control gain is designed to achieve the optimal *asymptotic convergence factor*, which is expressed by the *eigenratio* of an *undirected* graph.

Consider the sampled double-integrator systems with a sampling period $h > 0$ for each agent [28]:

$$x_i(k+1) = x_i(k) + hv_i(k) + \frac{1}{2}h^2 u_i(k), \tag{9.29}$$

$$v_i(k+1) = v_i(k) + hu_i(k), \forall i \in \mathcal{N}, \tag{9.30}$$

where $x_i(k) \in \mathbb{R}$ and $v_i(k) \in \mathbb{R}$ respectively correspond to the position and velocity of agent i at time kh, and $u_i(k) \in \mathbb{R}$ is the control input.

Denoting the configuration variable of agent i at time kh by $\xi_i(k) = [x_i(k), v_i(k)]^T$, the agent dynamic is written in a vector form as follows:

$$\xi_i(k+1) = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \xi_i(k) + \begin{bmatrix} \frac{1}{2}h^2 \\ h \end{bmatrix} u_i(k), \ \forall i \in \mathcal{N}. \tag{9.31}$$

The following control protocol is adopted:

$$u_i(k) = K \sum_{j=1}^{N} a_{ij}(\xi_j(k) - \xi_i(k)), \quad K \in \mathbb{R}^{1\times 2}. \tag{9.32}$$

By Lemma 9.2 and Theorem 9.9 we get a necessary and sufficient condition for reaching the second-order consensus.

Corollary 9.2. *Given an* undirected *graph $\mathcal{G}$, a necessary and sufficient condition for the second-order multiagent systems (9.31) to reach consensus under protocol (9.32) is that $\mathcal{G}$ is connected.*

In this occasion, a control gain K that solves the second-order consensus can be designed without resorting to the solution to the modified algebraic Riccati inequality (9.5).

Theorem 9.4. *Given an* undirected *graph $\mathcal{G}$, let*

$$\Omega = \{[\alpha, \beta] | \beta < \frac{2}{\lambda_N h}, 0 < \alpha < \frac{2\beta}{h}\}.$$

Under the conditions in Corollary 9.2, a control gain K in (9.32) solves the second-order consensus problem if and only if $K \in \Omega$.

Proof. By (9.15) and the definition of $J(K)$ in (9.24), a control gain $K \triangleq [\alpha, \beta]$ solves the second-order consensus problem under protocol (9.32) if and only if $\rho(J(K)) < 1$. It is easy to compute that

$$\begin{aligned} \det(zI_2 - (A - \lambda BK)) &= z^2 + (\frac{1}{2}\alpha\lambda h^2 + \beta\lambda h - 2)z + \frac{1}{2}\alpha\lambda h^2 - \beta\lambda h + 1 \quad (9.33) \\ &\triangleq (z - z_1)(z - z_2). \end{aligned}$$

Applying to the latter a bilinear transformation, that is, $z = \frac{s+1}{s-1}$, we obtain the new polynomial

$$\alpha\lambda h^2 s^2 + (2\beta\lambda h - \alpha\lambda h^2)s + 4 - 2\beta\lambda h \triangleq (\alpha\lambda h^2)(s - s_1)(s - s_2).$$

From the property of the bilinear transformation it follows that $|z_1| < 1$ and $|z_2| < 1$ if and only if $s_1 < 0$ and $s_2 < 0$, which are equivalent to that $\beta < \frac{2}{\lambda h}$ and $0 < \alpha < \frac{2\beta}{h}$ by the Routh stability criterion. The rest of the proof is trivial. □

The following result characterizes the optimal *asymptotic convergence factor* among all possible control gains that achieve a consensus under protocol (9.32).

Theorem 9.5. *Given an* undirected *graph $\mathcal{G}$, under the conditions in Corollary 9.2, the optimal* asymptotic convergence factor *for reaching consensus of the second-order multiagent systems (9.31) under protocol (9.32) is*

$$r_{\text{asym}}^* = \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/2}. \tag{9.34}$$

Moreover, the control gain

$$K^* = \left[\frac{1-(r_{\text{asym}}^*)^2}{h^2\lambda_N}, \ \frac{3+(r_{\text{asym}}^*)^2}{2h\lambda_N}\right]$$

leads to the optimal asymptotic convergence factor.

Proof. In light of Theorem 9.3, the optimal *asymptotic convergence factor* is lower bounded by

$$r_{\text{asym}}^* \geq \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/2}. \tag{9.35}$$

Next, a control gain $K = [\alpha, \beta]$ is to be constructed to show that the lower bound is tight. For notational simplicity, let $\alpha_0 = \frac{1}{2}\alpha h^2 + \beta h$, $\beta_0 = \frac{1}{2}\alpha h^2 - \beta h$, and $\sigma = \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/2}$. Then we obtain that

$$\begin{aligned} \det(zI_2 - (A - \lambda BK)) &= z^2 + (\alpha_0\lambda - 2)z + \beta_0\lambda + 1 \\ &\triangleq (z - z_+(\lambda))(z - z_-(\lambda)). \end{aligned}$$

It is clear that for all $\lambda \geq \frac{4(\alpha_0+\beta_0)}{\alpha_0^2}$, $z_+(\lambda)$ and $z_-(\lambda)$ are real numbers and can be expressed by

$$z_+(\lambda) = 1 - \frac{\alpha_0+\beta_0}{\frac{\alpha_0}{2} + \sqrt{\frac{\alpha_0^2}{4} - \frac{\alpha_0+\beta_0}{\lambda}}}, \tag{9.36}$$

$$z_-(\lambda) = 1 - \frac{\alpha_0+\beta_0}{\frac{\alpha_0}{2} - \sqrt{\frac{\alpha_0^2}{4} - \frac{\alpha_0+\beta_0}{\lambda}}}. \tag{9.37}$$

Setting $z_+(\lambda_N) = \sigma$ and $z_-(\lambda_N) = -\sigma$, we get a solution of (α_0, β_0) as follows:

$$\begin{cases} \alpha_0^* &= \frac{2}{\lambda_N}, \\ \beta_0^* &= -\frac{\sigma^2+1}{\lambda_N}. \end{cases}$$

Noting that

$$\frac{4(\alpha_0^* + \beta_0^*)}{(\alpha_0^*)^2} = (1 - \sigma^2)\lambda_N \in [\lambda_2, \lambda_N),$$

there exists $k \in \{2, \dots, N-1\}$ such that $\frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} \leq \lambda_{k+1}$ and $\frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} \geq \lambda_k$. Thus, $z_+(\lambda_j)$ and $z_-(\lambda_j)$, $j \in \{k+1, \dots, N\}$ are real numbers. For any control gain K solving the second-order consensus, Theorem 9.4 assures that $\alpha_0^* + \beta_0^* = \frac{1-\sigma^2}{\lambda_N} > 0$. Then, $z_+(\lambda)$ and $z_-(\lambda)$ are respectively increasing and decreasing functions w.r.t. $\lambda > \frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2}$. This implies that

$$z_+(\lambda_N) \geq \cdots \geq z_+(\lambda_{k+1}) \geq z_-(\lambda_{k+1}) \geq \cdots \geq z_-(\lambda_N)$$

and

$$\max_{j\in\{k+1,\dots,N\}} \{|z_+(\lambda_j)|, |z_-(\lambda_j)|\} = \sigma. \tag{9.38}$$

On the other hand, for any $\lambda \leq \frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2}$, $z_+(\lambda)$ and $z_-(\lambda)$ are a pair of conjugate complex numbers, and

$$\begin{aligned} |z_+(\lambda)|^2 &= |z_-(\lambda)|^2 = 1 + \frac{1}{2}\lambda^2(\alpha_0^*)^2 - \lambda(\beta_0^* + 2\alpha_0^*) \\ &= 2\left(\frac{\lambda}{\lambda_N}\right)^2 + (\sigma^2 - 3)\left(\frac{\lambda}{\lambda_N}\right) + 1. \end{aligned}$$

In particular, for all $j \in \{2, \dots, k\}$, $z_+(\lambda_j)$ and $z_-(\lambda_j)$ are complex numbers. Moreover,

$$|z_+((1-\sigma^2)\lambda_N)|^2 - \sigma^2 = \sigma^2(\sigma^2 - 1) \leq 0$$

and

$$\begin{aligned} |z_+(\lambda_2)|^2 - \sigma^2 &= 2\left(\frac{1-\sigma^2}{1+\sigma^2}\right)^2 + (\sigma^2 - 3)\left(\frac{1-\sigma^2}{1+\sigma^2}\right) + 1 - \sigma^2 \\ &= -\frac{2\sigma^2(\sigma^2-1)^2}{(1+\sigma^2)^2} \leq 0. \end{aligned}$$

Together with

$$\lambda_2 \leq \cdots \leq \lambda_k \leq \frac{4(\alpha_0^* + \beta_0^*)}{(\alpha_0^*)^2} = (1-\sigma^2)\lambda_N,$$

it follows that

$$\max_{j\in\{2,\dots,k\}} |z_+(\lambda_j)|^2 \leq \max\{|z_+(\lambda_2)|^2, |z_+((1-\sigma^2)\lambda_N)|^2\} \leq \sigma^2.$$

Combing this with (9.38), we conclude that

$$\max_{j\in\{2,\ldots,N\}}\{|z_+(\lambda_j)|, z_-(\lambda_j)|\}=\sigma.$$

Thus, the lower bound of (9.35) is attainable. Solving the equations

$$\begin{cases} \frac{2}{\lambda_N} &= \frac{1}{2}\alpha h^2+\beta h, \\ -\frac{\sigma^2+1}{\lambda_N} &= \frac{1}{2}\alpha h^2-\beta h, \end{cases}$$

we obtain that the control gain

$$K^*=\left[\frac{1-\sigma^2}{h^2\lambda_N}, \frac{3+\sigma^2}{2h\lambda_N}\right]$$

leads to the optimal *asymptotic convergence factor* $r^*_{\text{asym}}=\sigma$. □

9.4 Consensus Control With Relative Output Feedback

In this subsection, we consider the situation where each agent does not know its exact output but can measure the output relative to those of his neighboring agents. For instance, in vehicle coordination, the vision-based sensor on a vehicle cannot directly locate the position of the vehicle in a global coordinate system but can measure the relative position to its neighbors. While in networked clock consensus, we are more concerned with the time difference between each pair of clocks.

9.4.1 Distributed Observer-Based Protocol

We propose two *admissible* control protocols. Precisely, we first adopt a *static* control protocol:

$$u_i(k)=F\sum_{j\in\mathcal{N}_i}a_{ij}(y_j(k)-y_i(k))\triangleq F\zeta_i(k), F\in\mathbb{R}^{1\times m}. \tag{9.39}$$

The second *admissible* control protocol is an observer-based *dynamic* protocol that depends on an internal controller state. Let $\sum_{j\in\mathcal{N}_i}a_{ij}(x_j(k)-x_i(k))\triangleq\xi_i(k)$. Since $\xi_i(k)$ is no longer available, a very natural thing is to design an observer to estimate $\xi_i(k)$ for the control design. In view of the agent dynamics, we will study the following observer-based control protocol for agent i:

$$\begin{cases} \widehat{\xi}_i(k+1) &= A\widehat{\xi}_i(k)+B\sum_{j\in\mathcal{N}_i}a_{ij}(u_j(k)-u_i(k))+L(\zeta_i(k)-C\widehat{\xi}_i(k)), \\ u_i(k) &= K\widehat{\xi}_i(k), L\in\mathbb{R}^{n\times m}, K\in\mathbb{R}^{1\times n}. \end{cases} \tag{9.40}$$

At time k, agent i computes the aggregate relative measurements $\zeta_i(k)$ to those of its neighbors. Together with control inputs from its neighbors $u_j(k)$, $j \in \mathcal{N}_i$, which will be received before time $k+1$, the agent updates its internal controller state to obtain $\widehat{\xi}_i(k+1)$ and produces the control input $u_i(k+1)$. It is clear that the *dynamic* control protocol in (9.40) is *admissible*. Compared to the *static* protocol in (9.39), this *dynamic* protocol requires each agent to broadcast its control input to his neighboring agents.

Observe the special case that the initial estimate is perfect, that is, $\xi_i(0) = \widehat{\xi}_i(0)$, it can be easily shown that $\xi_i(k) = \widehat{\xi}_i(k)$ for all $k \in \mathbb{N}$. When the consensus is reached, the internal controller state $\widehat{\xi}_i(k)$ of this case becomes zero. Thus it is reasonable to impose an additional condition on the definition of consensus that all controller internal states $\widehat{\xi}_i(k),\ i \in \mathcal{V}$, should asymptotically converge to zero.

Definition 9.2. *Given an* undirected *communication graph $\mathcal{G}$, the discrete-time multiagent systems (9.1) are said to reach* consensus *under the* dynamic *protocol (9.40) if for any finite $x_i(0), \forall i \in \mathcal{V}$, the control protocol can asymptotically drive the states of all agents close to each other and all the controller internal states to zero, that is,*

$$\lim_{k\to\infty} \|x_i(k) - x_j(k)\| = 0 \ \&\ \lim_{k\to\infty} \|\widehat{\xi}_i(k)\| = 0, \forall i, j \in \mathcal{V}. \tag{9.41}$$

To elucidate the role of the graph, we focus ourselves on *undirected* graphs.

9.4.2 Consensus Under Static Protocol

In this subsection, we first provide a necessary and sufficient condition under the *static* control protocol (9.39). Noting that the verification of this condition is nontrivial, we proceed to seek a necessary and sufficient condition for consensus under the *dynamic* control protocol (9.40). The roles of the *undirected* graph and agent dynamics on consensus are exactly quantified.

Theorem 9.6. *Given an* undirected *communication graph $\mathcal{G}$, the discrete-time multiagent systems (9.1) reach consensus under the* static *control protocol (9.39) if and only if there exists a common gain $F \in \mathbb{R}^{1\times m}$ such that $\rho(A - \lambda_j BFC) < 1,\ \forall j \in \{2, \dots, N\}$.*

Proof. Denote the average state of all agents by

$$\bar{x}(k) \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i(k) = \frac{1}{N}(\mathbf{1}^T \otimes I_n)x(k),$$

where $x(k) \triangleq [x_1^T(k), \dots, x_N^T(k)]^T$, and the deviation of each state from the average state by $\delta_i(k) \triangleq x_i(k) - \bar{x}(k)$, where $\mathbf{1}$ is a compatible dimension vector with each element of one and similarly for $\mathbf{0}$. Then, this yields that

$$\lim_{k\to\infty} \|\delta_i(k)\| \leq \frac{1}{N}\sum_{j=1}^{N} \lim_{k\to\infty} \|x_i(k) - x_j(k)\| = 0.$$

Conversely, $\lim_{k\to\infty} \|\delta_i(k)\| = 0,\ \forall i \in \mathcal{V}$, immediately implies the consensus of the multiagent systems (9.1). Thus, it is equivalent to finding a necessary and sufficient condition such that $\lim_{k\to\infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{V}$. Inserting the control protocol (9.39) into each agent dynamics, the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BFC)x(k). \tag{9.42}$$

Since $\mathbf{1}^T \mathcal{L}_{\mathcal{G}} = \mathbf{0}^T$, we obtain

$$\begin{aligned} \bar{x}(k+1) &= \frac{1}{N}(\mathbf{1}^T \otimes A)x(k) - \frac{1}{N}(\mathbf{1}^T \mathcal{L}_{\mathcal{G}} \otimes BFC)x(k) \\ &= A\bar{x}(k). \end{aligned} \tag{9.43}$$

Letting $\delta(k) = [\delta_1^T(k), \ldots, \delta_N^T(k)]^T$ and subtracting (9.42) from (9.43) lead to that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BFC)\delta(k). \tag{9.44}$$

Select $\phi_i \in \mathbb{R}^N$ such that $\phi_i^T \mathcal{L}_{\mathcal{G}} = \lambda_i \phi_i^T$ and form the unitary matrix $\Phi = [\frac{\mathbf{1}}{\sqrt{N}}, \phi_2, \ldots, \phi_N]$ to transform $\mathcal{L}_{\mathcal{G}}$ into a diagonal form: $diag(0, \lambda_2, \ldots, \lambda_N) = \Phi^T \mathcal{L}_{\mathcal{G}} \Phi$. Further, using the property of Kronecker product gives that

$$\begin{aligned} (\Phi \otimes I_n)^T (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BFC)(\Phi \otimes I_n) &= I_N \otimes A - \Phi^T \mathcal{L}_{\mathcal{G}} \Phi \otimes BFC \\ &= diag(A, A - \lambda_2 BFC, \ldots, A - \lambda_N BFC). \end{aligned} \tag{9.45}$$

Denote $\widetilde{\delta}(k) = (\Phi \otimes I_n)^T \delta(k)$ and partition $\widetilde{\delta}(k) \in \mathbb{R}^{nN}$ into two parts, that is, $\widetilde{\delta}(k) = [\widetilde{\delta}_1^T(k), \widetilde{\delta}_2^T(k)]^T$, where $\widetilde{\delta}_1(k) \in \mathbb{R}^n$ is a vector consisting of the first n elements of $\widetilde{\delta}(k)$. Then, $\widetilde{\delta}_1(k) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \delta_i(k) = \mathbf{0}$. In view of (9.44) and (9.45), this yields that

$$\widetilde{\delta}_2(k+1) = diag(A - \lambda_2 BFC, \ldots, A - \lambda_N BFC)\widetilde{\delta}_2(k). \tag{9.46}$$

The rest of the proof is straightforward. □

This result reveals how the eigenvalues of the Laplacian matrix affect consensus. However, the verification of the condition in Theorem 9.6 is nontrivial although some conservative sufficient conditions can be given in terms of linear matrix inequalities [29]. A more explicit characterization of the effect of these eigenvalues on the consensus is given in the following subsection.

9.4.3 Consensus Under Dynamic Protocol

Theorem 9.7. *Given an* undirected *communication graph $\mathcal{G}$, the discrete-time multiagent systems (9.1) reach consensus under the* dynamic *control protocol (9.40)* if and only if *the following conditions hold:*

(a) (A, B, C) *are stabilizable and detectable;*

(b) *Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the strict inequality*

$$\prod_j |\lambda_j^u(A)| < \frac{1+\lambda_2/\lambda_N}{1-\lambda_2/\lambda_N}, \tag{9.47}$$

where $\lambda_j^u(A)$ represents an unstable eigenvalue of A, and λ_2 and λ_N are respectively the second smallest and largest eigenvalues of the Laplacian matrix associated with $\mathcal{G}$.

Moreover, if the stated conditions hold, a control gain K that solves the consensus problem can be selected as

$$K = \frac{2}{\lambda_2 + \lambda_N}\frac{B^T P A}{B^T P B},$$

where P is a positive definite solution to the following discrete-time algebraic Riccati inequality:

$$P - A^T P A + \frac{A^T P B B^T P A}{B^T P B} > 0. \tag{9.48}$$

The observer gain L is chosen to make $\rho(A - LC) < 1$.

The proof depends on the following lemmas.

Lemma 9.5. *([22]) Suppose that the sequence $\{z_k\} \subset \mathbb{R}$ is recursively computed by the formula $z_{k+1} = (1 - a_k)z_k + b_k$, $\forall k \in \mathbb{N}$, and $a_k \in [0, 1)$, $\sum_{k=0}^{\infty} a_k = \infty$, $|z_0| < \infty$. Then if $\lim_{k\to\infty} \frac{b_k}{a_k}$ exists, we have $\lim_{k\to\infty} z_k = \lim_{k\to\infty} \frac{b_k}{a_k}$.*

Lemma 9.6. *([30]) For any $A \in \mathbb{R}^{n\times n}$ and $\epsilon > 0$, we have*

$$\|A^k\| \le M\eta^k, \forall k \ge 0, \tag{9.49}$$

where $M = \sqrt{n}\left(1 + \frac{2}{\epsilon}\right)^{n-1}$ and $\eta = \rho(A) + \epsilon\|A\|$.

Proof of Theorem 9.7. Define $\widetilde{\xi}_i(k)$ as the estimation error of $\xi_i(k)$, that is, $\widetilde{\xi}_i(k) = \widehat{\xi}_i(k) - \xi_i(k)$. Inserting the control protocol (9.40) into each agent dynamics, the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)x(k) + (I_N \otimes BK)\widetilde{\xi}(k). \tag{9.50}$$

Similarly to the proof of Theorem 9.6, it is easy to show that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k) + (I_N \otimes BK)\widetilde{\xi}(k). \tag{9.51}$$

Let $E(k) = (P \otimes I_n)^T (I_N \otimes BK)\widetilde{\xi}(k)$ and partition it into two parts $E(k) = [E_1^T(k), E_2^T(k)]^T$, where $E_1(k) \in \mathbb{R}^n$ is the vector consisting of the first n elements of $E(k)$. Then, following similar arguments of the proof of Theorem 9.6, we have that

$$\left\{ \begin{array}{rcl} \widetilde{\delta}_1(k) & = & 0, \forall k \in \mathbb{N}. \\ \widetilde{\delta}_2(k+1) & = & diag(A - \lambda_2 BK, \ldots, A - \lambda_N BK)\widetilde{\delta}_2(k) + E_2(k). \end{array} \right. \tag{9.52}$$

Necessity: By (9.1) it follows that

$$\begin{aligned} \xi_i(k+1) &= \sum_{j \in \mathcal{N}_i} a_{ij}(x_j(k+1) - x_i(k+1)) \\ &= A\xi_i(k) + BK \sum_{j \in \mathcal{N}_i} a_{ij}(\widehat{\xi}_j(k) - \widehat{\xi}_i(k)). \end{aligned} \tag{9.53}$$

Together with (9.40), the error dynamic of $\widetilde{\xi}_i(k)$ is written by $\widetilde{\xi}_i(k+1) = (A - LC)\widetilde{\xi}_i(k)$. Assume that the multiagent systems (9.1) reach a consensus under the *dynamic* protocol (9.40), it follows that

$$\begin{aligned} &\lim_{k\to\infty} \|\widetilde{\xi}_i(k)\| = \lim_{k\to\infty} \|\widehat{\xi}_i(k) - \xi_i(k)\| \\ &\leq \lim_{k\to\infty} \|\widehat{\xi}_i(k)\| + \lim_{k\to\infty} \|\xi_i(k)\| \\ &\leq \|K\| \sum_{j \in \mathcal{N}_i} a_{ij} \lim_{k\to\infty} \|x_j(k) - x_i(k)\| = 0, \forall i \in \mathcal{V}. \end{aligned} \tag{9.54}$$

Thus we get that $\rho(A - LC) < 1$. This implies that (C, A) is detectable.

Now, we consider a particular case where the initial estimate of $\xi_i(0), \forall i \in \mathcal{V}$, is perfect. By the error dynamics of $\widetilde{\xi}_i(k)$ it is easy to see that $\widetilde{\xi}_i(k) = 0, \forall i \in \mathcal{V}$, which further implies that $E_2(k) = 0, \forall k \in \mathbb{N}$. In light of (9.52), it immediately follows that $\rho(A - \lambda_j BK) < 1, \forall j \in \{2, \ldots, N\}$. The rest of the proof of the necessity follows from Lemma 9.3.

Sufficiency: Since (A, B) is stabilizable, there exists a positive definite solution P to the algebraic Riccati inequality (9.48). In view of Lemma 9.3, the proposed control gain K can simultaneously stabilize the stabilizable pairs $(A, \lambda_j B), \forall j \in \{2, \ldots, N\}$, that is,

$$\varrho \triangleq \max_{j\in\{2,\ldots,N\}} \rho(A - \lambda_j BK) < 1. \tag{9.55}$$

In addition, the observer gain L makes the estimation error asymptotically converge to zero, that is,

$$\lim_{k\to\infty} \widetilde{\xi}_i(k) = 0,$$

which further implies that $\lim_{k\to\infty} \|E_2(k)\| = 0$. Denoting $J(K) = diag(A - \lambda_2 BK, \ldots, A - \lambda_N BK)$, it follows from (9.52) that

$$\widetilde{\delta}_2(k+1) = J(K)^{k+1}\widetilde{\delta}_2(0) + \sum_{i=0}^{k} J(K)^{k-i} E_2(i). \tag{9.56}$$

Select a positive ϵ such that $\epsilon < \frac{1-\varrho}{\|J(K)\|}$ and $\eta = \varrho + \epsilon\|J(K)\| < 1$. By Lemma 9.6 it follows that $\|J(K)^k\| \le M\eta^k$. Thus we obtain that

$$\|\widetilde{\delta}_2(k+1)\| \le M\left(\eta^{k+1} + \sum_{i=0}^{k} \eta^{k-i}\|E_2(i)\|\right). \tag{9.57}$$

Consider the following auxiliary system: $z_{k+1} = \eta z_k + \|E_2(k)\|, z_0 = 1$. In view of Lemma 9.5, we have that

$$\lim_{k\to\infty} z_k = \frac{\lim_{k\to\infty}\|E_2(k)\|}{1-\eta} = 0. \tag{9.58}$$

By iteration it is clear that $z_{k+1} = \eta^{k+1} + \sum_{i=0}^{k} \eta^{k-i}\|E_2(i)\|$. Hence, we have proved that

$$\lim_{k\to\infty} \|\widetilde{\delta}_2(k)\| = 0. \tag{9.59}$$

Together with the fact that $\widetilde{\delta}_1(k) = 0, \forall k \in \mathbb{N}$, it follows that $\lim_{k\to\infty}\|\delta(k)\| = 0$. Thus, we get that $\lim_{k\to\infty}\|x_i(k) - x_j(k)\| = 0, \forall i, j \in \mathcal{V}$, which further implies that $\lim_{k\to\infty}\|\xi_i(k)\| = 0, \forall i \in \mathcal{V}$. Moreover, the following statement is straightforward:

$$\lim_{k\to\infty}\|\widehat{\xi}_i(k)\| = \lim_{k\to\infty}\|\xi_i(k)\| + \lim_{k\to\infty}\|\widetilde{\xi}_i(k)\| = 0. \tag{9.60}$$

By Definition 9.2 the proof is completed. □

9.4.4 Multiagent Systems With Double Integrators

Consider the following discrete-time double-integrator systems for each agent:

$$\left\{\begin{array}{lcl} x_i(k+1) & = & x_i(k)+hv_i(k), \\ v_i(k+1) & = & v_i(k)+hu_i(k), \end{array}\right. \forall i \in \mathcal{V}, k \in \mathbb{N}, \tag{9.61}$$

where h is the sampling interval, $x_i(k) \in \mathbb{R}$ and $v_i(k) \in \mathbb{R}$ respectively correspond to the position and velocity of agent i at time kh, and $u_i(k) \in \mathbb{R}$ is the control input. Under this setting, A, B, C are respectively written as

$$A = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ h \end{bmatrix}, C = [1\ 0].$$

Consider the situation that each agent does not know its position in a global coordinate system but can measure the position relative to those of its neighboring agents. We may attempt to reach a consensus by adopting a control protocol of the form

$$u_i(k) = \gamma \sum_{j=1}^{N} a_{ij}(x_j(k) - x_i(k)), \quad \gamma \in \mathbb{R}. \tag{9.62}$$

Intuitively, this control protocol only uses a relative position information, and it may be not able to drive the multiagent system to reach a consensus. We note that the protocol is adopted for first-order multiagent systems to reach an average consensus [14].

Theorem 9.8. *The second-order multiagent systems (9.61) cannot reach a consensus under the control protocol (9.62) for any undirected communication graph.*

Proof. In view of (9.52), it can be similarly established that

$$\widetilde{\delta}_j(k+1) = (A - \lambda_j \gamma BC)\widetilde{\delta}_j(k), \forall j \in \{2, \ldots, N\}. \tag{9.63}$$

It is straightforward that

$$\det(zI_2 - (A - \lambda_j \gamma BC)) = z^2 - 2z + 1 + \lambda_j h^2 \gamma. \tag{9.64}$$

Together with (9.63), we cannot guarantee that for any finite initial state $\xi_0(k)$, $\lim_{k\to\infty} \|\widetilde{\delta}(k)\| \neq 0$ since (9.64) contains at least one unstable root. This completes the proof. □

Due to the distinct feature of the double-integrator system, the relative velocity can be accessed by using the relative position information with one-step delay. For example, by (9.61) it follows that

$$v_j(k-1) = \frac{x_j(k) - x_j(k-1)}{h}.$$

Thus, we study the following control protocol. Let $x_j(k) = 0, \forall k < 0$, and

$$\begin{aligned} u_i(k) &= \sum_{j=1}^{N} a_{ij} \left[\gamma_0(x_j(k) - x_i(k)) + \gamma_1(v_j(k-1) - v_i(k-1))\right] \\ &\triangleq \sum_{j=1}^{N} a_{ij} \left[\alpha(x_j(k) - x_i(k)) + \beta(x_j(k-1) - x_i(k-1))\right]. \end{aligned} \tag{9.65}$$

Without incurring any additional communication cost, this protocol requires each agent to store the relative position feedback at the previous step. Under such a simple protocol of (9.65), a connected graph is also necessary and sufficient for reaching a consensus.

Theorem 9.9. *Given an* undirected *communication graph, the second-order multiagent systems (9.61) reach consensus under the control protocol (9.65)* if and only if *the communication graph is connected. Moreover, if this condition holds, then* (α, β) *in the protocol (9.65) can be selected from the set*

$$\Omega_c \triangleq \left\{(\alpha, \beta) | \max\{-\frac{1}{h^2}, -\frac{1}{\lambda_N h^2}\} < \beta < 0, \alpha = -\frac{\lambda_N h^2 \beta^2 + 3\beta}{2}\right\}. \tag{9.66}$$

Proof. Similarly to the proof of Theorem 9.6, we obtain that, for all $j \in \{2, \ldots, N\}$,

$$\widetilde{\delta}_j(k+1) = (A - \alpha\lambda_j BC)\widetilde{\delta}_j(k) - \beta\lambda_j BC\widetilde{\delta}_j(k-1). \tag{9.67}$$

Let $\Delta_j(k) = [\widetilde{\delta}_j^T(k-1), \widetilde{\delta}_j^T(k)]^T$, where $\widetilde{\delta}_j(k) = 0, \forall k < 0$. In view of (9.67), the dynamical equation of $\Delta_j(k)$ is expressed by

$$\begin{aligned} \Delta_j(k+1) &= \begin{bmatrix} 0 & I_2 \\ -\beta\lambda_j BC & A - \alpha\lambda_j BC \end{bmatrix} \Delta_j(k) \\ &\triangleq M_j(\alpha, \beta)\Delta_j(k), \forall j \in \{2, \ldots, N\}. \end{aligned} \tag{9.68}$$

Thus, a necessary and sufficient condition for the multiagent system (9.61) to reach a consensus is that $\rho\left(M_j(\alpha, \beta)\right) < 1, \forall j \in \{2, \ldots, N\}$.

Necessity: If the communication graph is not connected, it immediately follows that $\lambda_2 = 0$, which implies that $\rho(M_2(\alpha, \beta)) = 1, \forall \alpha, \beta \in \mathbb{R}$. In view of (9.68), we cannot guarantee that $\lim_{k\to\infty} \|\Delta_j(k)\| = 0, \forall \Delta_j(0) \in \mathbb{R}^4$. This contradicts Definition 9.4.

Sufficiency: We show that, for any connected graph and any $(\alpha, \beta) \in \Omega_c$, we have $\rho\left(M_j(\alpha, \beta)\right) < 1, \forall j \in \{2, \ldots, N\}$. It is easy to compute that

$$\det(zI_4 - M_j(\alpha, \beta)) = z\left(z^3 - 2z^2 + (1 + \lambda_j h^2\alpha)z + \lambda_j h^2\beta\right). \tag{9.69}$$

Let the polynomial be $f(z) = z^3 - 2z^2 + (1 + x)z + y$.

By using the Jury stability test [31] we ca show that all roots of $f(z)$ are inside the unit circle *if and only if* $(x, y) \in \Omega$, where

$$\Omega \triangleq \{(x, y)| - y < x < -y^2 - 2y\}. \tag{9.70}$$

Finally, we can verify that, for any $(\alpha, \beta) \in \Omega_c$, $(\lambda_j \alpha h^2, \lambda_j \beta h^2) \in \Omega, \forall j \in \{2, \dots, N\}$. Together with (9.69), the proof is completed. □

9.5 Formation Control for Multiagent Systems

As an important application, the result on consensus is extended to study formation of the discrete-time multiagent systems (9.1). Specifically, given a formation vector $H = [h_1^T, h_2^T, \dots, h_N^T]^T$, the following control protocol is adopted to study the formation problem of the discrete-time multiagent systems (9.1):

$$u_i(k) = K \sum_{j=1}^{N} a_{ij} \left[(x_j(k) - h_j) - (x_i(k) - h_i)\right], \tag{9.71}$$

where $h_i - h_j$ is the desired distance vector between agents i and j. In the context of formation control, the protocol (9.71) has been widely adopted for continuous-time systems [1,32, 33]. As in those works, the common knowledge of the directions of reference axes is required for all the agents.

Definition 9.3. *The discrete-time multiagent systems (9.1) with a fixed graph $\mathcal{G}$ are said to reach formation under protocol (9.71) if for any finite $x_i(0), \forall i \in \mathcal{N}$, there exists a control gain $K \in \mathbb{R}^{1\times n}$ in (9.71) such that*

$$\lim_{k\to\infty} \|(x_i(k) - h_i) - (x_j(k) - h_j)\| = 0, \forall i, j \in \mathcal{N}. \tag{9.72}$$

Based on Theorem 9.9, a necessary and sufficient condition for reaching formation of the discrete-time multiagent systems is stated as follows.

Theorem 9.10. *Given a set of desired formation vectors h_i, $i \in \mathcal{N}$, and an* undirected *graph $\mathcal{G}$, assume that* **A1**) *holds. Then the discrete-time multiagent systems (9.1) reach formation under protocol (9.71) if and only if the following conditions hold:*

(a) *(A, B) is a controllable pair, and $A(h_i - h_j) = h_i - h_j, \forall i, j \in \mathcal{N}$;*
(b) *Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the strict inequality*

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}. \tag{9.73}$$

Proof. Denote the average formation vector $\bar{h} \triangleq \frac{1}{N}\sum_{i=1}^{N} h_i$ and $\delta_i(k) \triangleq x_i(k) - h_i - (\bar{x}(k) - \bar{h})$. Similarly, it is easy to verify that reaching formation is equivalent to that $\lim_{k\to\infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{N}$. The following dynamical equation can be easily derived:

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k) + (I_N \otimes (A - I_n)) \begin{bmatrix} h_1 - \bar{h} \\ \vdots \\ h_N - \bar{h} \end{bmatrix}. \tag{9.74}$$

Thus, to reach the desired formation, we have that $(A - I_n)(h_i - \bar{h}) = 0, \forall i \in \mathcal{N}$, which implies that $A(h_j - h_i) = h_j - h_i, \forall i, j \in \mathcal{N}$. The rest follows from the proof of the necessity of Theorem 9.9.

Conversely, using the condition that $A(h_j - h_i) = h_j - h_i, \forall i, j \in \mathcal{N}$, (9.74) is reduced to the following form:

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k). \tag{9.75}$$

Again, the remainder of the proof follows from the sufficiency proof of Theorem 9.9. □

Remark 9.5. *For the continuous-time case, the formation condition is modified as* $A(h_i - h_j) = \mathbf{0}, \forall i, j \in \mathcal{N}$ *[20,33]. The physical meaning of the constraint* $A(h_i - h_j) = h_i - h_j, \forall i, j \in \mathcal{N}$*, will become clear for the second-order consensus problem in Section 9.4.4. For example, to maintain a fixed formation, the velocities of all the agents should be the same.*

9.5.1 Vehicle Formation With Double Integrators

As an important application, we study the vehicle formation problem with the relative position feedback. The vehicle dynamical equation is described by the discrete-time double-integrator system (9.61).

Given an arbitrary formation vector $f = [f_1, \cdots, f_N]^T \in \mathbb{R}^N$, where f_i represents the desired separation of agent i from the centroid of all agents, the objective is to design a simple control protocol such that the vehicles reach the desired formation. Motivated by (9.65), we will investigate the formation of the following distributed controller:

$$u_i(k) = \sum_{j=1}^{N} a_{ij}[\alpha(x_j(k) - x_i(k) - f_j + f_i) + \beta(x_j(k-1) - x_i(k-1) - f_j + f_i)]. \tag{9.76}$$

Definition 9.4. *Given an* undirected *communication graph* $\mathcal{G}$*, the second-order multiagent systems (9.61) are said to reach formation under the protocol (9.76) if for any finite initial position* $x_i(0)$ *and velocity* $v_i(0), i \in \mathcal{V}$*, there exists a pair of* $(\alpha, \beta) \in \mathbb{R}^2$ *such that*

$$\lim_{k\to\infty} \|(x_i(k) - f_i) - (x_j(k) - f_j)\| = 0, \forall i, j \in \mathcal{V}. \tag{9.77}$$

Theorem 9.11. *Given an* undirected *communication graph $\mathcal{G}$, the second-order multiagent systems (9.61) reach formation under the control protocol (9.76)* if and only if *the communication graph is connected. Moreover, if this condition holds, then (α, β) in the protocol of (9.76) can be chosen from the set Ω_c of (9.66).*

Proof. Let $\bar{f} = \frac{1}{N}\sum_{j=1}^{N} f_j$ be the average of the formation vector. Denote the displacement vector by $\delta_i(k) = [x_i(k) - f_i - \bar{x}(k) + \bar{f}, v_i(k) - \bar{v}(k)]^T$ and $\delta(k) = [\delta_1(k)^T, \dots, \delta_N(k)^T]^T$. Inserting the controller (9.76) into (9.61) leads to that

$$\delta(k+1) = (I_N \otimes A - \alpha\mathcal{L}_{\mathcal{G}} \otimes BC)\delta(k) - (\beta\mathcal{L}_{\mathcal{G}} \otimes BC)\delta(k-1). \tag{9.78}$$

The rest of the proof follows from that of Theorem 9.9. □

9.5.2 Formation-Based Tracking Problem

The dynamics of the uncooperative leader is described by the following constant velocity model:

$$\begin{cases} x_0(k+1) & = & x_0(k) + hv_0(k), \\ v_0(k+1) & = & v_0(k) + hu_0(k), k \in \mathcal{V}, \end{cases} \tag{9.79}$$

where the control input $u_0(k)$ is an independent and identically distributed (i.i.d.) random process with zero mean and $\mathbb{E}[|u_0(k)|^2] = \sigma^2$.

The connection weights vector between followers and the leader is denoted by $b = [b_1, \dots, b_N]^T$, where b_i is positive if and only if the leader is a neighbor of agent i, and otherwise $b_i = 0$. The leader's neighboring agent can measure its position relative to the leader. The goal is to design a simple distributed controller such that the center of all vehicles (except the leader), denoted by $\bar{x}(k) = \frac{1}{N}\sum_{i=1}^{N} x_i(k)$, asymptotically tracks the leader while keeping a given formation vector f.

To this purpose, assume that the average of the formation vector $\bar{f}$ is accessible to the vehicles that are connected to the leader. We propose the following control protocol:

$$\begin{aligned} u_i(k) \quad = \quad & \sum_{j=1}^{N} a_{ij}[\alpha(x_j(k) - x_i(k) - f_j + f_i) + \beta(x_j(k-1) - x_i(k-1) - f_j + f_i)] \\ & - b_i[\alpha(x_i(k) - x_0(k) - f_i + \bar{f}) + \beta(x_i(k-1) - x_0(k-1) - f_0 + \bar{f})]. \end{aligned} \tag{9.80}$$

The summation of the first square bracket is to make the vehicles to maintain the given formation vector f, whereas the rest is used to drive the center of the vehicles to asymptotically track the leader.

Denote by $\mathbb{R}_{\geq 0}$ the set of nonnegative real numbers. A function $g : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ is said to be of class $\mathcal{K}_\infty$ if it is continuous, strictly increasing, unbounded, and crosses the origin.

Definition 9.5. *Given an* undirected *communication graph $\mathcal{G}$ and the formation-based tracking problem associated with the double-integrator multiagent systems (9.61), the leader (9.79) is said to be* solvable *under the protocol in (9.80) if for any finite initial position $x_i(0)$ and velocity $v_i(0), i \in \mathcal{V} \bigcup \{0\}$, there exist a pair $(\alpha, \beta) \in \mathbb{R}^2$ and $g \in \mathcal{K}_\infty$ such that*

$$\begin{cases} \limsup_{k\to\infty} \mathbb{E}[\|(x_i(k) - f_i) - (x_j(k) - f_j)\|^2] \leq g(\sigma^2), \\ \limsup_{k\to\infty} \mathbb{E}[\|\bar{x}(k) - x_0(k)\|^2] \leq g(\sigma^2), \forall i, j \in \mathcal{V}, \end{cases} \tag{9.81}$$

where the mathematical expectation is taken w.r.t. the process $\{u_0(k)\}_{k\in\mathbb{N}}$.

Denote the index of the leader by 0 and $\mathcal{V}' = \mathcal{V} \bigcup \{0\}$. Similarly, denote the new adjacency matrix

$$\mathcal{A}' = \begin{bmatrix} 0 & b \\ b^T & \mathcal{A} \end{bmatrix}$$

and the corresponding edge set $\mathcal{E}'$. Let $\mathcal{L}' = \mathcal{L}_\mathcal{G} + diag(b_1, \ldots, b_N)$ and write the ascending order of the eigenvalues of $\mathcal{L}'$ by $\lambda'_1 \leq \lambda'_2 \leq \cdots \leq \lambda'_N$. It is clear that the new graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}', \mathcal{A}'\}$ is generated by adding an *undirected* edge from the agent i to the leader if $b_i \neq 0$. Note that in fact only the follower can take relative measurements to its neighbors. The following result solves the formation-based leader–follower consensus problem.

Theorem 9.12. *Given an* undirected *communication graph $\mathcal{G}$, the control protocol (9.80) solves the formation-based tracking problem associated with the second-order multiagent systems (9.61) and the leader (9.79)* if and only if

(a) The communication graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}', \mathcal{A}'\}$ is connected;

(b) At least one agent connects to the leader, that is, $\sum_{i=1}^N b_i \neq 0$.

Moreover, if these conditions hold, $(\alpha, \beta) \in \Omega'_c$ solves the formation-based tracking problem, where Ω'_c is given by

$$\Omega'_c \triangleq \left\{ (\alpha, \beta) | \max\{-\frac{1}{h^2}, -\frac{1}{\lambda'_N h^2}\} < \beta < 0, \alpha = -\frac{\lambda'_N h^2 \beta^2 + 3\beta}{2} \right\}. \tag{9.82}$$

The function $g \in \mathcal{K}_\infty$ can be chosen to be linear, that is, $g(\sigma^2) = c\sigma^2$, where the positive number c is a constant depending on h^2, (α, β), and λ'_j, $j \in \mathcal{V}$.

Proof. Define the "tracking error" of agent i by $\delta_i(k) = [x_i(k) - f_i - x_0(k) + \bar{f},\ v_i(k) - v_0(k)]^T$. Since if $g_1, g_2 \in \mathcal{K}_\infty$, then $g_1 + g_2 \in \mathcal{K}_\infty$, it is easy to verify that the solvability of the formation-based leader–follower consensus problem is equivalent to that there exists $g_i^\delta \in \mathcal{K}_\infty$ such that $\lim_{k\to\infty} \mathbb{E}[\|\delta_i(k)\|^2] \le g_i^\delta(\sigma^2), \forall i \in \mathcal{V}$. Collect $\delta_i(k)$ to get the new vector $\delta(k) = [\delta_1^T(k), \ldots, \delta_N^T(k)]^T$. Inserting the control protocol (9.80) into (9.61) leads to that

$$\delta(k+1) = \big(I_N \otimes A - \alpha\mathcal{L}' \otimes BC\big)\,\delta(k) - (\beta\mathcal{L}' \otimes BC)\delta(k-1) - (\mathbf{1} \otimes \begin{bmatrix} 0 \\ h \end{bmatrix})u_0(k). \quad (9.83)$$

Select $\psi_i \in \mathbb{R}^N$ such that $\psi_i^T \mathcal{L}' = \lambda_i' \psi_i^T, \forall i \in \mathcal{V}$. Form the unitary matrix $\Psi = [\psi_1, \psi_2, \ldots, \psi_N]$ to transform $\mathcal{L}'$ into a diagonal form

$$diag(\lambda_1', \lambda_2', \ldots, \lambda_N') = \Psi^T \mathcal{L}' \Psi. \quad (9.84)$$

Define $\widetilde{\delta}(k) = (\Psi \otimes I_2)^T \delta(k)$ and partition it in conformity with $\delta(k)$. Apply the same partition pattern to $\Psi\mathbf{1} \otimes \begin{bmatrix} 0 \\ h \end{bmatrix} \triangleq [q_1^T, \ldots, q_N^T]^T$. It follows that

$$\widetilde{\delta}_j(k+1) = (A - \alpha\lambda_j' BC)\widetilde{\delta}_j(k) - \beta\lambda_j' BC\widetilde{\delta}_j(k-1) - q_j u_0(k). \quad (9.85)$$

Letting $M_j'(\alpha, \beta) \triangleq \begin{bmatrix} 0 & I_2 \\ -\beta\lambda_j' BC & A - \alpha\lambda_j' BC \end{bmatrix}$ and $\Delta_j(k) = [\widetilde{\delta}_j^T(k), \widetilde{\delta}_j^T(k-1)]^T$, we obtain that

$$\Delta_j(k+1) = M_j'(\alpha, \beta)\Delta_j(k) + [\mathbf{0}^T, q_j^T]^T u_0(k). \quad (9.86)$$

Denote $P_j(k) \triangleq \mathbb{E}[\Delta_j(k)\Delta_j(k)^T]$ and the zero matrix $0_2 \in \mathbb{R}^{2\times 2}$. It is easy to derive that

$$P_j(k+1) = M_j'(\alpha, \beta)P_j(k)M_j'(\alpha, \beta)^T + \sigma^2 diag(0_2, q_j q_j^T). \quad (9.87)$$

Necessity: By Definition 9.5 it is clearly seen that there exists $g' \in \mathcal{K}_\infty$ such that $\limsup_{k\to\infty} E\|P_j(k)\|^2 \le g'(\sigma^2)$. Jointly with (9.87), this implies that $\rho(M_j'(\alpha, \beta)) < 1$, $\forall j \in \mathcal{V}$. Similarly to the proof of Theorem 9.9, it is obvious that a necessary condition should be $\lambda_j' > 0, \forall j \in \mathcal{V}$. Thus, we obtain that $\sum_{i=1}^N b_i \neq 0$ since otherwise $\mathcal{L}' = \mathcal{L}_\mathcal{G}$, which contains at least one zero eigenvalue, say $\lambda_1' = 0$. In addition, the Laplacian matrix corresponding to the new graph $\mathcal{G}'$ can be written as

$$\mathcal{L}_{\mathcal{G}'} = \begin{bmatrix} \sum_{i=1}^N b_i & -b \\ -b^T & \mathcal{L}' \end{bmatrix}.$$

Since $\lambda_1' > 0$, this means that $\mathcal{L}'$ is nonsingular. Then we have

$$\begin{bmatrix} 1 & b(\mathcal{L}')^{-1} \\ \mathbf{0} & I_N \end{bmatrix}\begin{bmatrix} \sum_{i=1}^N b_i & -b \\ -b^T & \mathcal{L}' \end{bmatrix}\begin{bmatrix} 1 & \mathbf{0}^T \\ (\mathcal{L}')^{-1}b^T & I_N \end{bmatrix} = \begin{bmatrix} b\mathbf{1} - b(\mathcal{L}')^{-1}b^T & \mathbf{0}^T \\ \mathbf{0} & \mathcal{L}' \end{bmatrix}.$$

Since $\mathcal{L}'\mathbf{1} = b^T$, it follows that $b(\mathbf{1} - (\mathcal{L}')^{-1}b^T) = 0$. This implies that $\mathcal{L}_{\mathcal{G}'}$ contains one simple zero eigenvalue. In particular, the second smallest eigenvalue of $\mathcal{L}_{\mathcal{G}'}$ is positive. This implies that the graph $\mathcal{G}'$ is connected [10].

Sufficiency: In view of the sufficiency part of Theorem 9.7, we obtain that $\varrho \triangleq \max_{j\in\mathcal{V}} \rho(M_j'(\alpha,\beta)) < 1, \forall(\alpha,\beta) \in \Omega_c'$. It follows from (9.87) that

$$\begin{aligned}\lim_{k\to\infty} P_j(k) &= \sigma^2 \sum_{k=0}^{\infty} M_j'(\alpha,\beta)^k diag(0_2, q_j q_j^T)(M_j'(\alpha,\beta)^k)^T \\ &\le 2h^2\sigma^2 \sum_{k=0}^{\infty} (M_j'(\alpha,\beta)M_j'(\alpha,\beta)^T)^k,\end{aligned}$$

where the first inequality is due to that $\|diag(0_2, q_j q_j^T)\| \le 2h^2$. By Lemma 9.6 and $\varrho < 1$ it easy to establish that there exists a finite positive number $\varsigma = \varsigma(\varrho)$ such that $\max_{j\in\mathcal{V}} \| \sum_{k=0}^{\infty} (M_j'(\alpha,\beta)M_j'(\alpha,\beta)^T)^k \| \le \varsigma < \infty$. Because of the unitary matrix Ψ, it is trivial that $\limsup_{k\to\infty} \mathbb{E}[\|\delta_j(k)\|^2] = \limsup_{k\to\infty} \mathbb{E}[\|\widetilde{\delta}_j(k)\|^2] = \frac{1}{2}\limsup_{k\to\infty} \mathbb{E}[\|\Delta_j(k)\|^2] = \frac{1}{2}\lim_{k\to\infty} tr(P_j(k)) \le 4h^2\varsigma\sigma^2$. Thus, the function $g \in \mathcal{K}_\infty$ can be selected as $g(\sigma^2) = 8h^2\varsigma\sigma^2$. Noting that ϱ depends on (α,β) and λ_j', $j \in \mathcal{V}$, the proof is completed. □

9.6 Simulations and Experiments

9.6.1 Modeling

The cooperative control of a multirobot system is a typical consensus problem. Based on Fig. 9.1, the kinematic equation for the ith robot, which is the differentially driven wheeled mobile robot, as follows:

$$\dot{r}_{xi}(t) = v_i(t)\cos(\theta_i(t)), \quad \dot{r}_{yi}(t) = v_i(t)\sin(\theta_i(t)), \quad \dot{\theta}_i(t) = \omega_i(t), \tag{9.88}$$

where (r_x, r_y) is the center position, $v_i(t)$ is the linear velocity, $\omega_i(t)$ is the rotation angle velocity, and $\theta_i(t)$ is the rotation angle. Denoting by (h_x, h_y) the head position of a robot, we can obtain the head position as

$$\begin{bmatrix} h_{xi}(t) \\ h_{yi}(t) \end{bmatrix} = \begin{bmatrix} r_{xi}(t) \\ r_{yi}(t) \end{bmatrix} + L_i \begin{bmatrix} \cos(\theta_i(t)) \\ \sin(\theta_i(t)) \end{bmatrix}. \tag{9.89}$$

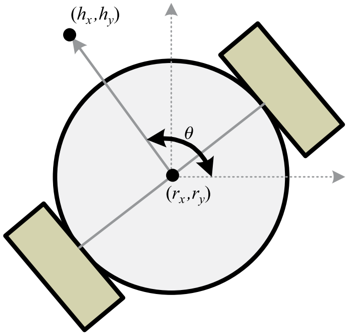

Figure 9.1: Differentially driven wheeled mobile robot.

From (9.89) the kinematic model of the robot system can be expressed as

$$\begin{bmatrix}\dot{h}_{xi}(t)\\ \dot{h}_{yi}(t)\end{bmatrix} = \begin{bmatrix}\cos(\theta_i(t)) & -l_i\sin(\theta_i(t))\\ \sin(\theta_i(t)) & +l_i\cos(\theta_i(t))\end{bmatrix}+\begin{bmatrix}v_i(t)\\ \omega_i(t)\end{bmatrix}, \tag{9.90}$$

$$\begin{bmatrix}\ddot{h}_{xi}(t)\\ \ddot{h}_{yi}(t)\end{bmatrix} = \begin{bmatrix}\cos(\theta_i(t)) & -l_i\sin(\theta_i(t))\\ \sin(\theta_i(t)) & +l_i\cos(\theta_i(t))\end{bmatrix}+\begin{bmatrix}\dot{v}_i(t)\\ \dot{\omega}_i(t)\end{bmatrix}+\begin{bmatrix}g_1\\ g_2\end{bmatrix}, \tag{9.91}$$

where

$$\begin{bmatrix}g_1\\ g_2\end{bmatrix}=\begin{bmatrix}-\sin(\theta_i(t))v_i(t)\omega_i(t)-l_i\cos(\theta_i(t))\omega_i^2(t)\\ \cos(\theta_i(t))v_i(t)\omega_i(t)-l_i\sin(\theta_i(t))\omega_i^2(t)\end{bmatrix}.$$

From (9.90) and (9.91) we can observe that the kinematic model of a robot system is nonlinear. It could be more difficult to perform the consensus control when the agent model is nonlinear. To transform (9.90) and (9.91) to a general double-integrator system, the kinematic equation (9.90) and (9.91) can be rewritten as

$$\begin{bmatrix}\dot{h}_{xi}(t)\\ \dot{h}_{yi}(t)\\ \ddot{h}_{xi}(t)\\ \ddot{h}_{yi}(t)\end{bmatrix}=\begin{bmatrix}0\\ 0\\ g_1\\ g_2\end{bmatrix}+\begin{bmatrix}\cos(\theta_i(t)) & -l_i\sin(\theta_i(t)) & 0 & 0\\ \sin(\theta_i(t)) & l_i\cos(\theta_i(t)) & 0 & 0\\ 0 & 0 & \cos(\theta_i(t)) & -l_i\sin(\theta_i(t))\\ 0 & 0 & \sin(\theta_i(t)) & l_i\cos(\theta_i(t))\end{bmatrix}\begin{bmatrix}v_i(t)\\ \omega_i(t)\\ \dot{v}_i(t)\\ \dot{\omega}_i(t)\end{bmatrix}. \tag{9.92}$$

Let

$$\begin{bmatrix}v_i(t)\\ \omega_i(t)\\ \dot{v}_i(t)\\ \dot{\omega}_i(t)\end{bmatrix}=E^{-1}\left(\begin{bmatrix}\dot{h}_{xi}(t)\\ \dot{h}_{yi}(t)\\ u_{xi}(t)\\ u_{yi}(t)\end{bmatrix}-\begin{bmatrix}0\\ 0\\ g_1\\ g_2\end{bmatrix}\right), \tag{9.93}$$

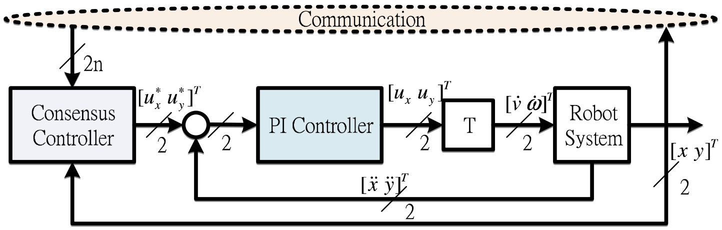

Figure 9.2: The scheme of the formation control of a multirobot system.

where

$$E = \begin{bmatrix} \cos(\theta_i(t)) & -l_i\sin(\theta_i(t)) & 0 & 0 \\ \sin(\theta_i(t)) & l_i\cos(\theta_i(t)) & 0 & 0 \\ 0 & 0 & \cos(\theta_i(t)) & -l_i\sin(\theta_i(t)) \\ 0 & 0 & \sin(\theta_i(t)) & l_i\cos(\theta_i(t)) \end{bmatrix}.$$

Substituting (9.93) into (9.92), we can obtain the equivalent double-integrator linear model as

$$\dot{X}_i(t) = V_i(t), \quad \dot{V}_i(t) = U_i(t),$$

where $x_i\dot{(}t) = \begin{bmatrix} h_{xi}(t) & h_{yi}(t) \end{bmatrix}^T$, $V_i(t) = \dot{X}_i(t)$, and $U_i(t) = \begin{bmatrix} u_{xi}(t) & u_{yi}(t) \end{bmatrix}^T$. The scheme of the formation control of a multirobot system is shown in Fig. 9.2. The control scheme contains the consensus control loop and the acceleration control loop. Based on the wheeled mobile robot, the consensus control action does not approach easily since the mobile robot has some limits, such as actuator ability and robot moving direction. A proportional-integral (PI) acceleration controller can increase the bandwidth in the inner loop. Note that the consensus control action can be guaranteed on the wheeled mobile robot.

9.6.2 Simulation Results

To validate the feasibility of proposed consensus protocol, we consider a multirobot system with one leader and four followers. The initial positions of the followers are set to $([0,5]\ [0,-5]\ [0,-15]\ [0,0])$, the initial velocities are given as $([3,1]\ [1.5,0.5]\ [0.75,0.25]\ [-0.25,0])$, formation shapes are specified as $([0,0]\ [-5,-5]\ [-10,0]\ [-5,5])$, and the initial position and velocity of the leader are $[0,15]$, $[0.5,0.2]$. The communication graph with $(0,1)$-weights for modeling the interactions among robots is illustrated in Fig. 9.3, where the information of the leader can be only transmitted to the first follower. In this case, the associated graph Laplacian matrix $\mathcal{L}_{G'}$ can be obtained as

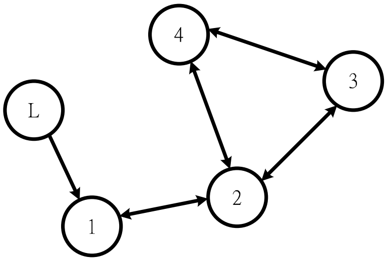

Figure 9.3: Scheme diagram of the ball and beam system.

$$\mathcal{L}_{\mathcal{G}'} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}.$$

From the proposed protocol the stability range of the control gains can be similarly derived as in Lemma 9.8:

$$K = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}, 0 < k_1 < 15.55, 0 < k_2 < 1.94,$$

and the optimal control gain vector can be obtained as $K^* = \begin{bmatrix} 0.25 \\ 1.91 \end{bmatrix}$ by Theorem 9.5.

The simulation results are shown in Figs. 9.4–9.6. Considering the ideal double-integral model, the responses of the following control corresponding to different control gains are shown in Fig. 9.4. With the optimal control gains $K = \begin{bmatrix} 0.25 & 1.91 \end{bmatrix}^T$, the responses of followers are shown in Fig. 9.4A. By observation, Fig. 9.4A has better performance compared to the responses with other designated control gains. Interestingly, choosing the control gains $K = \begin{bmatrix} 0.25 & 2.01 \end{bmatrix}^T$, the stability condition is not satisfied. From Fig. 9.4D we can see that the required formation task cannot be fulfilled. In practice, the kinematic model of a wheel robot is not the same as the ideal double-integral model. In fact, a certain input–output linearization technique is required to obtain the equivalent double-integral formulation. In this chapter, we propose an inner-loop control scheme. To verify the essential requirement of the proposed inner-loop control, the corresponding simulation results are shown in Figs. 9.5 and 9.6. In Fig. 9.5, the input–output linearization is performed without the inner-loop control. It can be seen that the desired formation control is failed even if the control gains are inside the stability region. The main reason about the instability in Fig. 9.5 is that the desired commands u_x^* and u_y^* for the coordinate transformation are not tracked well without the inner-loop loop control. On the other hand, with the inner-loop control scheme, the wheel robots can perform the desired formation task quite well. Compared with the responses of ideal double-integral model,

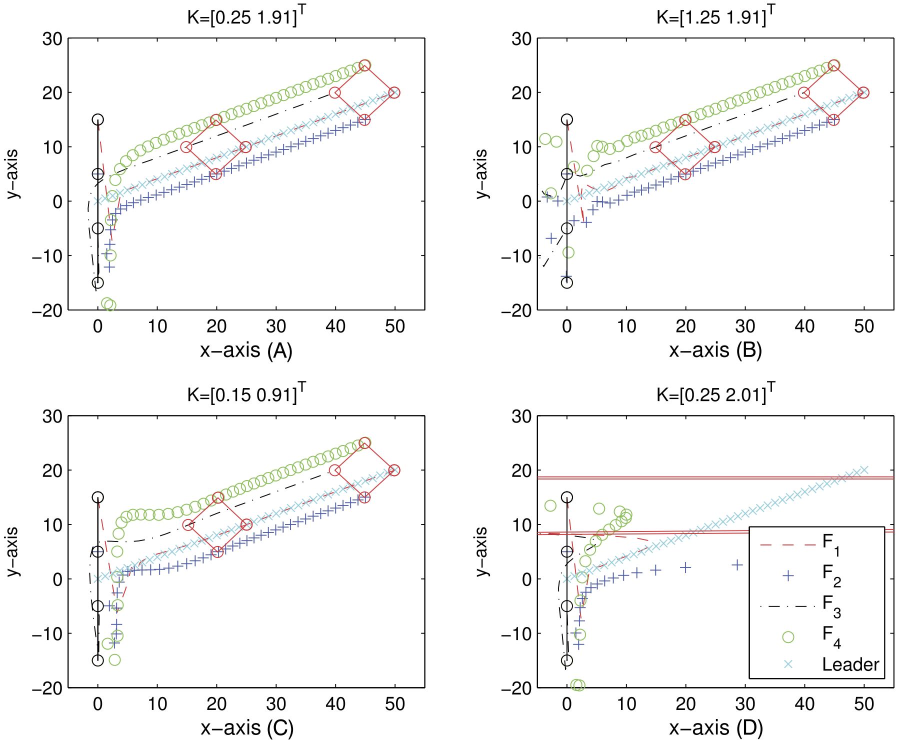

Figure 9.4: Responses of the following control with the ideal double-integrator linear system. (A) $k = [0.25,\ 1.91]^T$**, (B)** $K = [1.25,\ 1.91]^T$**, (C)** $K = [0.15,\ 0.91]^T$**, (D)** $K = [0.25,\ 2.01]^T$**.**

the formation responses of the wheel robots with inner-loop control are quite similar in the steady state.

9.7 Bibliographic Notes

There are certain limitations in existing studies of multiagent systems. First, the assumption that the communication link is perfect and an agent has some global knowledge on network topology is somehow restrictive. Note that changes in the operating environment, such as the random presence of large metal objects between agents, will inevitably affect the propagation properties of the channels. Thus, it is more interesting to consider the scenario that the communication channel is time varying and unreliable. The investigation of consensus over time-varying graphs may have far reaching consequences on the understanding and engineering of networked multiagent systems. With fixed graphs, consensus of multiagent systems under a common control protocol is converted into a simultaneous stabilization problem.

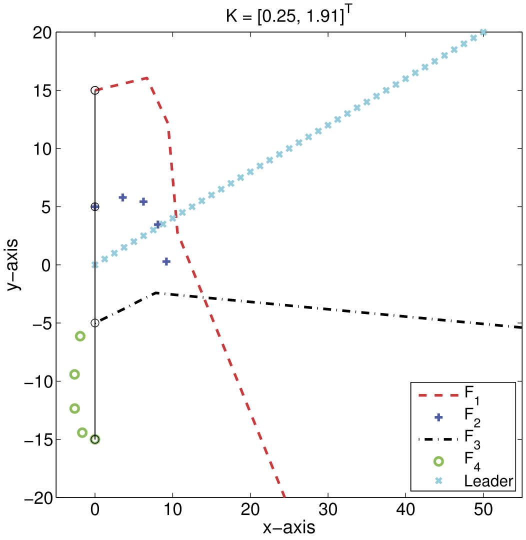

Figure 9.5: Responses of following control of the wheeled mobile robot system without inner loop controller, $K = [0.25,\ 1.91]^T$**.**

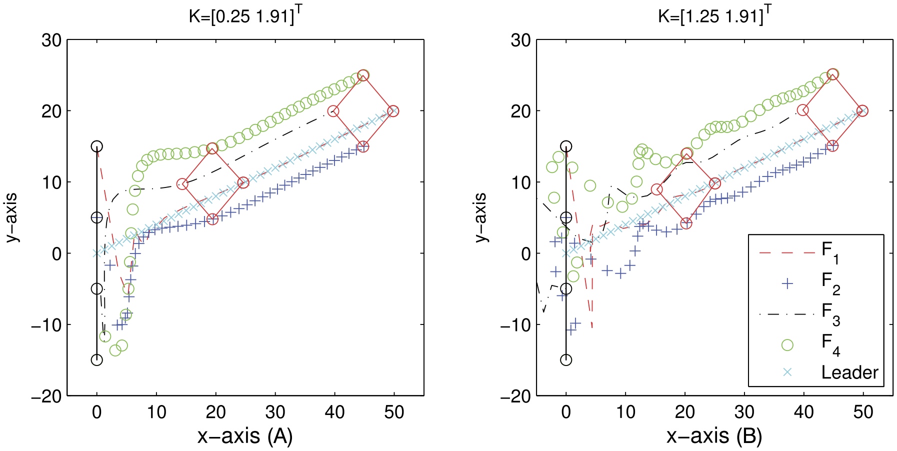

Figure 9.6: Responses of the following control of the wheeled mobile robot system with inner-loop control: (A) $K = [0.25,\ 1.91]^T$**, (B)** $K = [1.25,\ 1.91]^T$**.**

However, this key property does not hold in the case of time-varying graphs. Perhaps, a completely new method needs to be developed. On the other hand, in networked systems, there may be the case that different kinds of agents join and leave the network from time to time, and it is unrealistic to assume that an agent has perfect knowledge of other agents' dynamics.

How each agent will optimize its utility while minimizing its interferences to others requires some new thinking.

References

[1] J. Fax, R. Murray, Information flow and cooperative control of vehicle formations, IEEE Transactions on Automatic Control 49 (2004) 1465–1476.

[2] P. Yang, R. Freeman, K. Lynch, Multi-agent coordination by decentralized estimation and control, IEEE Transactions on Automatic Control 53 (2008) 2480–2496.

[3] J. Cortés, S. Martinez, T. Karatas, F. Bullo, et al., Coverage control for mobile sensing networks, IEEE Transactions on Robotics and Automation 20 (2004) 243–255.

[4] J. Cortés, F. Bullo, Coordination and geometric optimization via distributed dynamical systems, SIAM Journal on Control and Optimization 44 (2006) 1543–1574.

[5] A. Okubo, Dynamical aspects of animal grouping: Swarms, schools, flocks, and herds, Advances in Biophysics 22 (1986) 1.

[6] R. Olfati-Saber, R. Murray, Flocking with obstacle avoidance: cooperation with limited communication in mobile networks, in: Proc. 42nd IEEE Conference on Decision and Control, 2003.

[7] N. Lynch, Distributed Algorithms, Morgan Kaufmann, 1996.

[8] Z. Duan, G. Chen, L. Huang, Complex network synchronizability: analysis and control, Physical Review E 76 (2007) 56103.

[9] F. Sorrentino, M. Di Bernardo, F. Garofalo, Synchronizability and synchronization dynamics of weighed and unweighed scale free networks with degree mixing, International Journal of Bifurcation and Chaos 17 (2007) 2419–2434.

[10] R. Olfati-Saber, R. Murray, Consensus problems in networks of agents with switching topology and time-delays, IEEE Transactions on Automatic Control 49 (2004) 1520–1533.

[11] W. Ren, R. Beard, Consensus seeking in multiagent systems under dynamically changing interaction topologies, IEEE Transactions on Automatic Control 50 (2005) 655–661.

[12] A. Jadbabaie, J. Lin, A. Morse, Coordination of groups of mobile autonomous agents using nearest neighbor rules, IEEE Transactions on Automatic Control 48 (2003) 988–1001.

[13] T. Li, J. Zhang, Consensus conditions of multi-agent systems with time-varying topologies and stochastic communication noises, IEEE Transactions on Automatic Control 55 (2010) 2043–2057.

[14] R. Olfati-Saber, J. Fax, R. Murray, Consensus and cooperation in networked multi-agent systems, Proceedings of the IEEE 95 (2007).

[15] C. Godsil, G. Royle, Algebraic Graph Theory, Springer, New York, 2001.

[16] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak Mathematical Journal 23 (1973) 298–305.

[17] M. Fu, L. Xie, The sector bound approach to quantized feedback control, IEEE Transactions on Automatic Control 50 (2005) 1698–1711.

[18] N. Elia, Remote stabilization over fading channels, Systems & Control Letters 54 (2005) 237–249.

[19] T. Li, M. Fu, L. Xie, J. Zhang, Distributed consensus with limited communication data rate, IEEE Transactions on Automatic Control 56 (2011) 279–292.

[20] C. Ma, J. Zhang, Necessary and sufficient conditions for consensusability of linear multi-agent systems, IEEE Transactions on Automatic Control 55 (2010) 1263–1268.

[21] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2012.

[22] K. You, L. Xie, Minimum data rate for mean square stabilization of discrete LTI systems over lossy channels, IEEE Transactions on Automatic Control 55 (2010) 2373–2378.

[23] K. You, W. Su, M. Fu, L. Xie, Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization, Automatica 47 (2011) 170–176.

[24] K. You, L. Xie, Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses, IEEE Transactions on Automatic Control 56 (2011) 772–785.

[25] G. Nair, F. Fagnani, S. Zampieri, R. Evans, Feedback control under data rate constraints: an overview, Proceedings of the IEEE 95 (2007) 108–137.

[26] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks, Proceedings of the IEEE 95 (2007) 163–187.

[27] L. Xiao, S. Boyd, Fast linear iterations for distributed averaging, Systems & Control Letters 53 (2004) 65–78.

[28] W. Ren, E. Atkins, Distributed multi-vehicle coordinated control via local information exchange, International Journal of Robust and Nonlinear Control 17 (2007) 1002–1033.

[29] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, Society for Industrial Mathematics, 1994.

[30] V. Solo, One Step Ahead Adaptive Controller With Slowly Time-Varying Parameters, Technical report, Dept. ECE, John Hopkins University, Baltimore, 1991.

[31] P. Antsaklis, A. Michel, Linear Systems, Birkhäuser, 2006.

[32] C. Ma, System Analysis and Control Synthesis of Linear Multi-Agent Systems, Ph.D. dissertation, Institute of Systems Science, Chinese Academy of Sciences, China, May 2009.

[33] G. Lafferriere, A. Williams, J. Caughman, J. Veerman, Decentralized control of vehicle formations, Systems & Control Letters 54 (2005) 899–910.

CHAPTER 10

Structure Identification for Networked Systems

10.1 Introduction

In the previous chapter, we analyzed and synthesized networked systems under the condition that their models are available. In many practical applications, however, these models are not known. An unavoidable task for an efficient utilization of the results developed in this chapter is therefore establishing a model on the basis of existing information, such as collected experimental data, related field knowledge given by physics, chemistry, biology, finance, and so on. This problem is of particular importance in many fields, such as industrial systems, biological systems, financial systems, and so on, in which thousands of subsystems interact with each other, but the way of their interactions is not clear. For example, a cellular network to accomplish a biological task usually consists of numerous chemical species, such as DNA, RNA, proteins, small molecules, and so on. Different biological tasks are generally performed by complex interactions of these species. These interactions can rarely be directly measured, and/or their measurements are too economically expensive [1–4]. It is now widely recognized that causal relationship identification is essential in understanding biological behavior of a cellular network. Challenging issues here include not only a large number of interactions to be estimated, but also many restrictions on probing signals and limitations on the data length in a biological experiment. Similar phenomena happen to industrial systems, financial systems, and so on.

On the other hand, industries have accumulated a great amount of operation data in their long production process. Same things happen to financial markets in which various stock data have been recorded for more than a century. Recently, with significant developments of high-throughout technologies and proteomics analysis methods, economic costs of performing biological experiments in a cellular level have been greatly reduced, and various experimental data have began to accumulate. All these advancements in data acquisition technologies greatly increase possibilities of estimating direct influences among their subsystems.

Causality inference has been attracting attention from various fields for a long time. Among the numerous models suggested for causality description, the most popular three appear to be the so-called potential outcome model proposed by Neyman and Rubin, the causality network model proposed by Spirtes and Pearl, and the Granger causality model proposed by

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00010-5

Granger [5,6]. Causality is closely related to structure identification of a networked system, and estimation of the existence of a direct influence from one subsystem to another subsystem is in fact determining whether or not the output of a subsystem can directly affect the behavior of another subsystem. In other words, to determine weather or not the behavior variations in one subsystem are directly caused by the output of another subsystem. In this chapter, however, rather than investigating relations among these models, our attention is focused on the estimation of direct interactions among different subsystems in a networked system from experiment data.

Currently, several approaches have already been proposed for unraveling direct interactions in a networked system. These attempts include Boolean network methods [7–9], Bayesian network methods [10,11], partial correlation analysis [12], differential-equation-based time series analysis [13,14], and so on. However, when these methods are applied to a large-scale networked system, several difficulties usually arise [4,15]. One difficulty is that with the increment of the subsystem number in a networked system, computational costs of most of the methods increase exponentially. To overcome this difficulty, the maximum direct regulation number is limited in some methods, but this may significantly restrict application ranges of the method itself and leads to another problem of selecting principal subsystems, which is also mathematically difficult. On the other hand, statistical methods such as partial correlation analysis rely on a variety of pairwise correlation metrics, but this treatment usually recognizes an indirect effect wrongly as a direct one, which may lead to a high rate of the so-called false-positive errors.

In addition to these attempts, a so-called "top-down" approach has been proposed in [3,16] for causal regulation inference from steady-state concentration changes of chemical species in a cellular network, which is based on the total differential formula and total least squares (TLS) estimations. The results have been extended afterward to time series data, quasi-steady-state data, and so on [14], and the so-called "nondirect effect" condition on experiment designs has been significantly weakened [17]. It is reported in [18], however, that when the data length of an experiment is in a moderate size from a biological view of point, it is very rare that an identified regulation coefficient is near zero in a statistical sense. This means that the identified connections among the subsystems of a networked system are in general dense. Recalling that a large-scale system usually has a sparse structure [19–21], this means that the aforementioned TLS-based estimates usually give incorrect information in case that a subsystem has no direct effect on another subsystem. In other words, there may exist significant false-positive errors in these TLS-based estimates.

On the other hand, it is widely recognized that in the identification of a networked system, distinguishing direct and indirect regulations is of great engineering/financial/biological significance [4,15,22]. In actual identifications, however, experimental data alone are rarely sufficient to obtain a statistically sound model. To make things worse, it may even be possible that the identification problem is underdetermined if direct regulations exist between every two chemical elements. These imply that in this identification problem, compared to false-negative errors, it may be much more difficult to reduce false-positive errors. To overcome this difficulty, various interesting methods have been suggested, such as incorporating qualitative knowledge, restriction of the maximum number of nonzero regulatory inputs, penalizing the sum of direct regulation strengths, and so on [2,4,15,23]. But these methods are still far from being satisfactory.

In this chapter, we investigate how to estimate direct effects among subsystems of a networked system from experimental data, which summarizes the results obtained in [18] and [24]. As general conclusions on structure identification for a networked system are still far from mature, particular attention is given to gene regulatory networks. In this investigation, sparsity of a large-scale system is also taken into account. More precisely, the so-called power law is incorporated into the structure identification of a networked system to increase estimation accuracy on causal regulations, especially to reduce false-positive errors. Two types of experimental data are considered. One is steady-state experimental data, and the other is time series data, which reflect dynamic responses of a gene regulation network to external disturbances. Hopefully, the reported results reveal important factors and characteristics in structure identification of a large-scale networked system.

10.2 Steady-State Data-Based Identification

In this section, we study how to estimate the structure of a gene regulatory network using corrupted steady-state experimental data. We develop an identification algorithm, which explicitly incorporates the power law into estimations, which is widely adopted in the description of the sparsity of a large-scale network. Under the condition that parameters of the power law are known and measurement errors are Gaussian, we adopt the likelihood maximization approach. The developed estimation algorithms consist of three major steps. First, an angle minimization between subspaces is utilized to identify chemical elements that have direct influences on a prescribed chemical element under the assumption that the number of direct regulations is known. Second, interference coefficients from prescribed chemical elements are estimated through likelihood maximization with respect to measurement errors. Finally, direct regulation numbers are identified through maximizing a lower bound of an overall likelihood function. Application results of these methods are briefly summarized to an artificially constructed linear system with 100 elements, a MAPK pathway model with 103 chemical elements, some DREAM initiative *in silico* data, and some *in vivo* data. These results show that, compared with the widely adopted total least squares (TLS) method, parametric estimation accuracy can be significantly increased, and false-positive errors can be greatly reduced.

10.2.1 Description of the Inference Procedure

In this section, we infer causal relations from steady-state concentration changes of chemical species in a cellular network. This strategy is essentially based on the total differential of a nonlinear function, which has also been used in [3,16].

To derive a structure inference algorithm, we first establish relations between direct causal influences and variations of the steady-state concentrations in a cellular network. Assume that the dynamics of species concentrations in a cellular network with n chemical elements can be described by the following nonlinear differential equations:

$$\frac{dx_i}{dt} = f_i(x_k|_{k=1}^{n},\ p_k|_{k=1}^{q}), \quad i = 1, 2, \ldots, n, \tag{10.1}$$

where x_i stands for the concentration of the ith chemical element, whereas p_k is a kinetic parameter that can be changed or controlled through some external perturbations.

For notational simplicity, denote vectors $[x_1\ x_2\ \cdots\ x_n]^T$ and $[p_1\ p_2\ \cdots\ p_q]^T$ respectively by x and p, and let $x^{[s]}$ represent an equilibrium of the cellular network at which $\frac{dx_i}{dt} = 0$, $i = 1, 2, \cdots, n$. As argued in [3], at any steady-state $x^{[s]}$, a direct effect of the jth chemical species on the ith chemical species ($i \neq j$) can be measured by r_{ij}, which is defined as follows:

$$r_{ij} = \left.\frac{\partial ln(x_i)}{\partial ln(x_j)}\right|_{x=x^{[s]}}. \tag{10.2}$$

More specifically, a positive r_{ik} means that there is an activation effect from the kth chemical species to the ith chemical species, whereas a negative r_{ik} means that there is a repression effect from the kth chemical species to the ith chemical species. If $r_{ik} = 0$, then it is regarded that there are no direct influences from the kth chemical species to the ith chemical species. Based on this definition, direct algebraic operations show that

$$r_{ij} = -\left.\left(\frac{\partial f_i(x,\ p)}{\partial ln(x_j)} \bigg/ \frac{\partial f_i(x,\ p)}{\partial ln(x_i)}\right)\right|_{x=x^{[s]}}. \tag{10.3}$$

Let $\delta_x^{[s]}$ represent variations of the steady-state $x^{[s]}$ of the cellular network when the kinetic parameter changes from p to $p + \delta_p$. Then, from the definition of a steady state it is clear that

$$f_i(x^{[s]},\ p) = 0, \quad f_i(x^{[s]} + \delta_x^{[s]},\ p + \delta_p) = 0, \quad i = 1, 2, \ldots, n. \tag{10.4}$$

Assume that in an experiment, k of the kinetic parameters $p_j|_{j=1}^{q}$, $k \in \{1, 2, \cdots, q\}$, have been perturbed/changed by external perturbations. Denote the subscripts of the perturbed kinetic parameters by $j_1, j_2, \cdots, j_k$. Then, according to the dynamic equation of the cellular network given by Eq. (10.1), it is clear that $k \leq q$ and $j_\alpha \in \{1, 2, \cdots, q\}$ for each $\alpha = 1, 2, \cdots, k$. Moreover, assume that external perturbations are chosen so that the perturbed kinetic parameter p_{j_α} does not *directly* affect variations of x_i, $\alpha = 1, 2, \cdots, k$. This means that

$$\frac{\partial f_i(x, p)}{\partial p_{j_\alpha}} \equiv 0, \quad \alpha = 1, 2, \cdots, k.$$

Then, taking the Taylor expansions at $x^{[s]}$ and p of the difference between the left-hand sides of the two equations in Eq. (10.3), we have that

$$\sum_{j=1}^{n} \left.\frac{\partial f_i(x, p)}{\partial x_j}\right|_{x=x^{[s]}} \delta_{x_j}^{[s]} + O(||\delta_x^{[s]}||_2^2, \ ||\delta_p||_2^2) = 0, \tag{10.5}$$

where $\delta_{x_j}^{[s]}$ is the jth element of $\delta_x^{[s]}$. When both $||\delta_x^{[s]}||_2$ and $||\delta_p||_2$ are sufficiently small, this relation can be approximately expressed as

$$\sum_{j=1}^{n} \left.\frac{\partial f_i(x, p)}{\partial x_j}\right|_{x=x^{[s]}} \delta_{x_j}^{[s]} \approx 0, \tag{10.6}$$

which is equivalent to

$$\sum_{j=1}^{n} \left.\frac{\partial (x_j^{[s]} f_i(x, p))}{\partial x_j}\right|_{x=x^{[s]}} \times \frac{\delta_{x_j}^{[s]}}{x_j^{[s]}} \approx 0. \tag{10.7}$$

Here, $x_j^{[s]}$ represents the jth element of the vector $x^{[s]}$.

Dividing both sides of the last equation by $-\left.\frac{\partial (x_i^{[s]} f_i(x,p))}{\partial x_i}\right|_{x=x^{[s]}}$, we obtain the following relation:

$$\sum_{j=1}^{n} \left.\frac{\partial f_i(x, p)/\partial \ln(x_j)}{\partial f_i(x, p)/\partial \ln(x_i)}\right|_{x=x^{[s]}} \times \frac{\delta_{x_j}^{[s]}}{x_j^{[s]}} \approx 0. \tag{10.8}$$

Assume that m experiments have been performed. Denote $\frac{\delta_{x_j}^{[s]}}{x_j^{[s]}}$ of the lth experiment by R_{jl}, which is the relative variation of the steady concentrations of the jth chemical species in the lth experiment. Then, from the definition of r_{ij} and the last equation we further establish the following approximation:

$$\sum_{k=1,\ k\neq i}^{n} r_{ik} R_{kl} \approx R_{il}, \quad l = 1, 2, \ldots, m. \tag{10.9}$$

These derivations extend the results of [3,16] on a single perturbation to multiple perturbations. Under the assumption that there are n chemical elements in a cellular network and in the jth experiment, some external perturbations are added on a chemical specifies, which do not directly change the concentration of this species, and the dynamics of its species concentrations can be described by a set of ordinary nonlinear differential equations, Eq. (10.9) gives an approximate relation between steady-state concentration variations and direct causal effects for a cellular network. In this equation, r_{ik}, which is defined by Eq. (10.2), stands for the direct influences of the kth chemical element on the ith chemical element around the perturbed equilibrium, and $R_{kj}|_{k=1}^{n}$ defined just before Eq. (10.9) can be obtained from measured species concentrations.

Assume that m experiments are performed. Denote the vectors $[r_{i1}\ r_{i2}\ \cdots\ r_{i,i-1}\ r_{i,i+1}\ \cdots\ r_{in}]^T$ and $[R_{1j}\ R_{2j}\ \cdots\ R_{i-1,j}\ R_{i+1,j}\ \cdots\ R_{nj}]$ respectively by x and R_j. Moreover, define the matrix A and vector b respectively as

$$A = col(R_j|_{j=1}^{m}), \quad b = col(R_{ij}|_{j=1}^{m}).$$

Then relation (10.9) can be compactly expressed as

$$Ax \approx b. \tag{10.10}$$

The problem discussed in this section is identifying the vector x under the condition that both the matrix A and the vector b are provided. A distinctive characteristic of this problem is that measurement errors exist in both the matrix A and the vector b. On the other hand, when n is large, it is now well known that the distribution of the number of nonzero elements of vector x obeys approximately the so-called power law [6,19]. More precisely, let n_i represent the number of chemical elements that have direct influences on a randomly chosen chemical element in a cellular network, and let $\mathbf{Pr}\{\cdot\}$ be the probability of the occurrence of a random event. Then, there exist a positive number γ and a positive integer k_{min} such that[1]

$$\mathbf{Pr}\{n_i = k\} = \begin{cases} ck_{\min}^{-\gamma}, & 1 \le k \le k_{\min}, \\ ck^{-\gamma}, & k_{\min} < k \le n, \end{cases} \tag{10.11}$$

where $c = \left[k_{\min}^{1-\gamma} + \sum_{k=k_{\min}+1}^{n} k^{-\gamma}\right]^{-1}$. This structural information is incorporated into the algorithm for cellular network identification, which is proven to be helpful in estimation accuracy improvements.

[1] It is worth pointing out that in structural analysis for a large-scale network, investigations are usually focused on large n_is [6,19]. As there is generally no statistical information about the distribution of small n_is, it is reasonable to assume that all they have an equal probability to occur. This treatment makes the methods given in this section also applicable to identification of small or moderate-size networks and consistent with the method proposed in [2], which restricts the maximum of the number of nonzero direct regulations.

10.2.2 Identification Algorithm

As a first step toward incorporating the so-called power law into cellular network identification, the following two assumptions are adopted.

- Represent measurement errors of the matrix A and vector b respectively by ε_A and ε_b and assume that elements of $[\varepsilon_A\ \varepsilon_b]$ are independent of each other and have an identical normal distribution $\mathbf{N}(0,\ \sigma^2)$ with known σ.
- The parameters $k_{\min}$ and γ in the description of the power law are known.

When these pieces of information are available, a natural approach for causal regulation identification from experiment data is likelihood maximization. More specifically, let $\#(\cdot)$ represent the number of nonzero elements of a matrix or vector. Then, the likelihood function of measurement errors and the direct regulation number, denoted $L(\varepsilon_A,\ \varepsilon_b,\ k)$, can be written as follows:

$$L(\varepsilon_A,\ \varepsilon_b,\ k) = (2\pi\sigma^2)^{-mn/2}\mathbf{Pr}\{n_i = k\}e^{-\frac{\mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\}}{2\sigma^2}} \tag{10.12}$$

$$\text{subject to : } (A-\varepsilon_A)x = b-\varepsilon_b \text{ and } \#(x) = k, \tag{10.13}$$

where $\mathbf{tr}(\cdot)$ denotes the trace of a square matrix.

Note that the parameters m, n, $k_{\min}$, γ, and σ are assumed to be known. Recalling that $\mathbf{ln}(\cdot)$ is an increasing function over $(0,\ \infty)$, it is obvious that the above maximization problem is equivalent to minimizing the cost function $l(\varepsilon_A,\ \varepsilon_b,\ k)$ under the conditions of Eq. (10.13), where

$$l(\varepsilon_A,\ \varepsilon_b,\ k) = \frac{\mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\}}{2\sigma^2} + \begin{cases} \gamma\mathbf{ln}(k_{\min}), & 1 \le k \le k_{\min}, \\ \gamma\mathbf{ln}(k), & k_{\min} < k \le n. \end{cases} \tag{10.14}$$

This cost function is significant from a biological view of point, and both discrete and continuous variables are included in its optimization. Its minimization, however, currently is mathematically challenging. More precisely, it appears difficult to derive an analytic expression for the optimal k, ε_A, and ε_b and to develop a globally/locally convergent optimization algorithm. To obtain an estimate about the interactions in a cellular network, the following three major steps are adopted:

- For a fixed k, angle minimization between two subspaces is utilized to determine positions of the nonzero elements of the vector x.
- Under the condition that positions of nonzero elements are prescribed, the optimal value of the vector x is obtained through minimizing the first term of the cost function $l(\varepsilon_A, \varepsilon_b, k)$ with respect to measurement errors ε_A and ε_b.

- The number of nonzero elements in the vector x is obtained through a numerical search that minimizes an upper bound of the cost function $l(\varepsilon_A, \varepsilon_b, k)$.

These three steps are investigated respectively in the following subsections.

Position Determination for Direct Regulations

When the number of nonzero elements of the vector x is given, denoted by k, there are in principle C_{n-1}^{k} possibilities to locate these nonzero elements. Here, C_{n-1}^{k} denotes the combinatorial number of selecting k elements from the set $\{1, 2, \cdots, n-1\}$. Note that for fixed k, C_{n-1}^{k} increases exponentially as n increases. This implies that when a large-scale system is under investigation, in other words, when n is large, it is usually computationally intractable to consider all these combinations in order to find the optimal locations of the nonzero elements. As a matter of fact, according to our experience, the computation time is currently prohibitive if four direct regulations should be searched for a chemical element within a network having more than 100 species. On the other hand, it is clear from Eq. (10.13) that when there are only k nonzero elements in the vector x, then only k columns of the matrix A are used to fit the experiment data of the vector b. Denote the ith column of the matrix A from the left by a_i and the ith element of the vector x from the ceiling by x_i, $i = 1, 2, \cdots, n-1$. Moreover, assume that the j_αth element of the vector x has been determined to be nonzero, $\alpha = 1, 2, \cdots, k$. Define the matrix $\tilde{A} = [a_{j_1}\ a_{j_2}\ \cdots\ a_{j_k}]$ and the vector $\tilde{x} = [x_{j_1}\ x_{j_2}\ \cdots\ x_{j_k}]^T$. Then, the first constraint of Eq. (10.13) can be rewritten as

$$[\tilde{A} - \varepsilon_{\tilde{A}}]\tilde{x} = b - \varepsilon_b, \tag{10.15}$$

where $\varepsilon_{\tilde{A}} = [\varepsilon_{a_{j_1}}\ \varepsilon_{a_{j_2}}\ \cdots\ \varepsilon_{a_{j_k}}]$, and ε_{a_j} represents the measurement errors contained in the vector a_j, $j = 1, 2, \cdots, n-1$.

Note that under this situation, there are no longer any restrictions on the vector ε_{a_j} whenever $j \neq j_\alpha$, $\alpha = 1, 2, \cdots, k$. According to the definition of the cost function $l(\varepsilon_A,\ \varepsilon_b,\ k)$, the corresponding optimal value of the vector ε_{a_j} is the zero vector. We can therefore declare that

$$\min_{s.t.\ (A-\varepsilon_A)x=b-\varepsilon_b\ \text{ and }\ \#(x)=k} \mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\} = \min_{s.t.\ (\tilde{A}-\varepsilon_{\tilde{A}})\tilde{x}=b-\varepsilon_b} \mathbf{tr}\{[\varepsilon_{\tilde{A}}\ \varepsilon_b]^T[\varepsilon_{\tilde{A}}\ \varepsilon_b]\}. \tag{10.16}$$

The minimization problem on the right-hand side of Eq. (10.15) has been settled very well and is widely known as the total least squares (TLS). In particular, the following results have been established for a long time [25,26]:

$$\min_{s.t.\ (\tilde{A}-\varepsilon_{\tilde{A}})\tilde{x}=b-\varepsilon_b} \mathbf{tr}\{[\varepsilon_{\tilde{A}}\ \varepsilon_b]^T[\varepsilon_{\tilde{A}}\ \varepsilon_b]\} = \underline{\sigma}^2([\tilde{A}\ b]), \tag{10.17}$$

where $\underline{\sigma}(\cdot)$ is the minimal singular value of a matrix. Note that every singular value of a matrix is nonnegative. Therefore, the optimal nonzero element position determination problem can be mathematically expressed as

$$\min_{j_\alpha \in \{1,2,\cdots,n-1\},\ \alpha=1,2,\cdots,k} \underline{\sigma}([\tilde{A}\ b]). \tag{10.18}$$

These results are elegant. However, they cannot be directly applied to the optimal position determination of the nonzero elements of the vector x, noting that the corresponding minimization problem is still a combinatorial optimization problem, for which it is generally hard to find a globally or locally convergent algorithm with polynomial computational complexities. On the other hand, let $\mathcal{S}pan(b)$ and $\mathcal{S}pan(\tilde{A})$ denote respectively the subspaces spanned by the vector b and the column vectors of the matrix $\tilde{A}$. Then, an upper bound can be derived for $\underline{\sigma}([\tilde{A}\ b])$, which is proportional to the sine of the half of the angle between $\mathcal{S}pan(\tilde{A})$ and $\mathcal{S}pan(b)$. More precisely, we have the next theorem.

Theorem 10.1. *Let $\tilde{\theta}_k$ represent the angle between $\mathcal{S}pan(\tilde{A})$ and $\mathcal{S}pan(b)$. Moreover, let $||b||_2$ denote the Euclidean norm of the vector b. Then*

$$\underline{\sigma}([\tilde{A}\ b]) \leq 2||b||_2 \sin\frac{\tilde{\theta}_k}{2}. \tag{10.19}$$

Proof. For brevity, denote the matrix $\tilde{A}(\tilde{A}^T\tilde{A})^{-1/2}$ and the vector $b(b^Tb)^{-1/2}$ respectively by $\hat{\tilde{A}}$ and $\hat{b}$. Then, by the definition of the angle between two linear subspaces [26], we have that

$$\tilde{\theta}_k = \arccos(||\hat{\tilde{A}}^T\hat{b}||_2). \tag{10.20}$$

On the other hand, from the definitions of the matrix $\hat{\tilde{A}}$ and the vector $\hat{b}$ we can directly prove that

$$[\tilde{A}\ \ b]^T[\tilde{A}\ \ b] = \begin{bmatrix} (\tilde{A}^T\tilde{A})^{1/2} & 0 \\ 0 & (b^Tb)^{1/2} \end{bmatrix} \begin{bmatrix} I_k & \hat{\tilde{A}}^T\hat{b} \\ \hat{b}^T\hat{\tilde{A}} & 1 \end{bmatrix} \begin{bmatrix} (\tilde{A}^T\tilde{A})^{1/2} & 0 \\ 0 & (b^Tb)^{1/2} \end{bmatrix}. \tag{10.21}$$

Moreover,

$$\left|\lambda I_{k+1} - \begin{bmatrix} I_k & \hat{\tilde{A}}^T\hat{b} \\ \hat{b}^T\hat{\tilde{A}} & 1 \end{bmatrix}\right| = (\lambda-1)^{k-1}(\lambda-1-||\hat{\tilde{A}}^T\hat{b}||_2)(\lambda-1+||\hat{\tilde{A}}^T\hat{b}||_2). \tag{10.22}$$

Let $\mathcal{N}ull(\hat{b}^T\hat{\tilde{A}})$ denote the null space of $\hat{b}^T\hat{\tilde{A}}$. As $\hat{b}$ is a nonzero column vector, it is obvious that there exists a $k \times (k-1)$-dimensional real matrix T_0 such that

$$T_0^T T_0 = I_{k-1}, \quad \mathcal{N}ull(\hat{b}^T\hat{\tilde{A}}) = \mathcal{S}pan(T_0). \tag{10.23}$$

Define the matrix

$$T_1 = \begin{bmatrix} T_0 & \frac{\sqrt{2}\hat{\tilde{A}}^T\hat{b}}{2\|\hat{\tilde{A}}^T\hat{b}\|_2} & -\frac{\sqrt{2}\hat{\tilde{A}}^T\hat{b}}{2\|\hat{\tilde{A}}^T\hat{b}\|_2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

Then by Eq. (10.22) we can directly prove that $T_1T_1^T = T_1^TT_1 = I_{k+1}$ and

$$T_1^T \begin{bmatrix} I_k & \hat{\tilde{A}}^T\hat{b} \\ \hat{b}^T\hat{\tilde{A}} & 1 \end{bmatrix} T_1 = \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0 & 1+\|\hat{\tilde{A}}^T\hat{b}\|_2 & 0 \\ 0 & 0 & 1-\|\hat{\tilde{A}}^T\hat{b}\|_2 \end{bmatrix}. \tag{10.24}$$

For notational simplicity, define the scalar κ and vector ξ respectively as

$$\kappa = \frac{\|b\|_2}{\sqrt{\hat{b}^T\hat{\tilde{A}}[\|b\|_2^2(\tilde{A}^T\tilde{A})^{-1} + I_k]\hat{\tilde{A}}^T\hat{b}}}, \qquad \xi = \kappa \begin{bmatrix} (\tilde{A}^T\tilde{A})^{-1/2}\hat{\tilde{A}}^T\hat{b} \\ -\|b\|_2^{-1}\|\hat{\tilde{A}}^T\hat{b}\|_2 \end{bmatrix}.$$

Then, straightforward matrix manipulations show that $\xi^T\xi = 1$ and

$$T_1^T \begin{bmatrix} (\tilde{A}^T\tilde{A})^{1/2} & 0 \\ 0 & (b^Tb)^{1/2} \end{bmatrix} \xi = -\sqrt{2}\kappa\|\hat{\tilde{A}}^T\hat{b}\|_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{10.25}$$

From this relation and from Eqs. (10.21) and (10.24) we obtain the equality

$$\xi^T([\tilde{A}\ \ b]^T[\tilde{A}\ \ b])\xi = 2\kappa^2\|\hat{\tilde{A}}^T\hat{b}\|_2^2(1-\|\hat{\tilde{A}}^T\hat{b}\|_2). \tag{10.26}$$

Note that from the definition of κ it is clear that $0 \le \kappa\|\hat{\tilde{A}}^T\hat{b}\|_2 \le \|b\|_2$. We can therefore declare from Eqs. (10.20) and (10.26) that

$$\begin{aligned} \underline{\sigma}([\tilde{A}\ \ b]) &= \sqrt{\inf_{\xi^T\xi=1} \xi^T([\tilde{A}\ \ b]^T[\tilde{A}\ \ b])\xi} \\ &\le \sqrt{2\kappa^2\|\hat{\tilde{A}}^T\hat{b}\|_2^2(1-\|\hat{\tilde{A}}^T\hat{b}\|_2)} \end{aligned}$$

$$\begin{aligned}&\leq \sqrt{2||b||_2^2(1-\cos\tilde{\theta}_k)}\\ &\leq 2||b||_2 \sin\frac{\tilde{\theta}_k}{2}. \qquad (10.27)\end{aligned}$$

This completes the proof. □

In addition to these results, we have the following results on subspace angle minimization.

Define the matrix $\hat{A} = A(A^TA)^{-1/2}$ and the vector $\hat{b} = b(b^Tb)^{-1/2}$ and denote the jth column vector of the matrix $\hat{A}$ from the left by $\hat{a}_j$, $j = 1, 2, \cdots, n-1$. Then, we can establish the following theorem.

Theorem 10.2. *Let θ_k represent the angle between the subspace spanned by vectors $\hat{a}_{j_\alpha}|_{\alpha=1}^k$ and the subspace spanned by $\hat{b}$. Moreover, let $j_\alpha^{[opt]}$, $\alpha = 1, 2, \cdots, k$, be the positions of the elements of the vector $\hat{A}^Tb$ with the first k greatest magnitudes. Then*

$$\{j_\alpha^{[\mathrm{opt}]}|_{\alpha=1}^k\} = \arg\min_{j_1,j_2,\ldots,j_k} \theta_k. \qquad (10.28)$$

Proof. Assume that the vectors $a_i|_{i=1}^{n-1}$ are linearly independent and the vector b is not a zero vector. In this case, both the matrix $\hat{A}$ and the vector $\hat{b}$ are well defined. Moreover, we can straightforwardly prove that $\hat{a}_i|_{i=1}^n$ and $\hat{b}$ constitute respectively an orthonormal basis for $\mathcal{S}pan(A)$ and $\mathcal{S}pan(b)$.

Assume now that k vectors $\hat{a}_{j_\alpha}|_{\alpha=1}^k$ have been chosen from the vectors $\hat{a}_j|_{j=1}^{n-1}$, in which $j_\alpha \neq j_\beta$ whenever $\alpha \neq \beta$. Denote $[\hat{a}_{j_1}\ \hat{a}_{j_2}\ \cdots\ \hat{a}_{j_k}]$ by $\hat{A}^{[\mathrm{opt}]}$. As $\hat{a}_{j_\alpha}|_{\alpha=1}^k$ are orthogonal to each other and have a unit Euclidean length, it is a direct result of matrix analysis [26] that the angle between $\mathcal{S}pan(\hat{A}^{[opt]})$ and $\mathcal{S}pan(b)$ is

$$\theta_k^{[\mathrm{opt}]} = \arccos(\bar{\sigma}(\hat{A}^{[opt]T}\hat{b})), \qquad (10.29)$$

where $\bar{\sigma}(\cdot)$ denotes the maximum singular value of a matrix.

Note that $\hat{b}$ is a vector. Then from the definition of maximum singular value we have that

$$\begin{aligned}\sin^2\theta_k^{[\mathrm{opt}]} &= 1-\bar{\sigma}^2(\hat{A}^{[\mathrm{opt}]T}\hat{b})\\ &= 1-tr\{(\hat{A}^{[\mathrm{opt}]T}\hat{b})^T(\hat{A}^{[\mathrm{opt}]T}\hat{b})\}\\ &= 1-tr\{\hat{A}^{[\mathrm{opt}]}\hat{A}^{[\mathrm{opt}]T}\hat{b}\hat{b}^T\}. \qquad (10.30)\end{aligned}$$

Define the vectors $t_\alpha = [t_{1\alpha}\ t_{2\alpha}\ \cdots\ t_{n-1,\alpha}]^T$, $\alpha = 1, 2, \cdots, k$, where $t_{\beta\alpha} = 1$ when $\beta = j_\alpha$ and $t_{\beta\alpha} = 0$ when $\beta \neq j_\alpha$. Denote $[t_1\ t_2\ \cdots\ t_k]$ by T. Then from the definition of the matrix $\hat{A}^{[opt]}$ we have that

$$\hat{A}^{[opt]} = \hat{A}T. \tag{10.31}$$

Substitute Eq. (10.31) into Eq. (10.30). Direct algebraic manipulations show that

$$\sin^2\theta_k^{[\mathrm{opt}]} = 1 - tr\{(TT^T)(\hat{A}^T\hat{b})(\hat{A}^T\hat{b})^T\}. \tag{10.32}$$

On the other hand, from the definition of matrix T we can directly prove that $T^TT = I_k$ and

$$TT^T = [\tau_{ij}]_{i,j=1}^{n-1}, \quad \tau_{ij} = \begin{cases} 1, & i = j = j_\alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{10.33}$$

Assume that $\hat{A}^T\hat{b} = [\phi_1\ \phi_2\ \cdots\ \phi_{n-1}]^T$. Then, $(\hat{A}^T\hat{b})(\hat{A}^T\hat{b})^T = [\phi_i\phi_j]_{i,j=1}^{n-1}$. Based on this relation and Eq. (10.33), we can straightforwardly show that

$$tr\{(TT^T)(\hat{A}^T\hat{b})(\hat{A}^T\hat{b})^T\} = \sum_{\alpha=1}^{k}\phi_{j_\alpha}^2. \tag{10.34}$$

Hence

$$\sin^2\theta_k^{[\mathrm{opt}]} = 1 - \sum_{\alpha=1}^{k}\phi_{j_\alpha}^2. \tag{10.35}$$

Note that $\sin^2\theta$ is an increasing function as θ varies over the interval $\left[0,\ \frac{\pi}{2}\right]$. It is clear from the last relation that when k is fixed, to minimize θ_k, it is desirable to select $\phi_{j_\alpha}|_{\alpha=1}^{k}$ having the first k greatest magnitudes. The proof can now be completed by noting that $\hat{A}^Tb$ is proportional to $\hat{A}^T\hat{b}$. □

From these two theorems we can see that if the minimization of the minimal singular value of $[\tilde{A}\ b]$ is replaced by the minimization of θ_k, then the optimal $j_\alpha|_{\alpha=1}^{k}$ has a closed-form solution, and therefore the problem of the exponential increment of computation complexities is successfully avoided.

On the other hand, let ψ_i, $1 \le i \le n-1$, denote the ith row element of the vector A^Tb. Using the matrix T defined just before Eq. (10.31), we have that $\tilde{A} = AT$. Therefore, by Eq. (10.20) we can directly prove that

$$\begin{aligned} \cos^2\tilde{\theta}_k &= \hat{b}^T\hat{\tilde{A}}\hat{\tilde{A}}^T\hat{b} \\ &= \frac{1}{||b||_2^2}b^TAT(T^TA^TAT)^{-1}T^TA^Tb. \end{aligned} \tag{10.36}$$

Recall that $T^T T = I_k$. From the definition of the maximal singular value we can directly prove that $T^T A^T A T \leq \bar{\sigma}^2(A) I_k$. Hence,

$$\begin{aligned}\cos^2 \tilde{\theta}_k &\geq \frac{1}{\bar{\sigma}^2(A)||b||_2^2} b^T A T T^T A^T b \\ &= \frac{1}{\bar{\sigma}^2(A)||b||_2^2} tr\{(T T^T)(A^T b)(A^T b)^T\}.\end{aligned} \tag{10.37}$$

Note that the right-hand side of the above equation takes a form very similar to that of the second term of the right-hand side of Eq. (10.32). As in the proof of Theorem 10.2, we can straightforwardly show that

$$\sin^2 \tilde{\theta}_k \leq 1 - \frac{1}{\bar{\sigma}^2(A)||b||_2^2} \sum_{\alpha=1}^{k} \psi_{j_\alpha}^2 \tag{10.38}$$

Therefore, through selecting $j_\alpha|_{\alpha=1}^k$ that maximizes $\sum_{\alpha=1}^k \psi_{j_\alpha}^2$, an upper bound of $\tilde{\theta}_k$, and therefore an upper bound of $\underline{\sigma}([\tilde{A}\ \ b])$, has been minimized.

It is worth pointing out that although relations between θ_k and $\tilde{\theta}_k$ are still not very clear, computation experience reported in [18] shows that compared with maximization of $\sum_{\alpha=1}^k \psi_{j_\alpha}^2$, minimization of θ_k generally leads to better estimation performances in cellular network identifications. However, theoretical reasons for this phenomenon are still under investigation.

Estimation of Regulation Coefficients

When locations of nonzero elements of the vector x have been determined, their values can be directly obtained through singular value decomposition [25,26]. More specifically, we have the following results.

Theorem 10.3. *Assume that $\underline{\sigma}(\tilde{A}) > \underline{\sigma}([\tilde{A}\ b])$. Let $v = [v_1\ v_2\ \cdots\ v_{k+1}]^T$ denote the right singular vector of the matrix $[\tilde{A}\ b]$ with respect to its minimal singular value, and let $\tilde{x}^{[\text{opt}]}$ be the optimal vector $\tilde{x}$ corresponding to the left-hand side minimization problem of Eq. (10.17). Then, $\tilde{x}^{[\text{opt}]} = -\frac{1}{v_{k+1}}[v_1\ v_2\ \cdots\ v_k]^T$.*

When the condition $\underline{\sigma}(\tilde{A}) > \underline{\sigma}([\tilde{A}\ b])$ is not satisfied, analytical forms are still available for the optimal solution of the aforementioned minimization problem, but the expressions are more complicated. We refer the interested reader to [25,26] for details.

Determination of the Number of Direct Regulations

In the previous estimations, it was assumed that the number of direct regulations is known for a species in the cellular network. In actual applications, this is generally not the case. To have an estimate of this number, the cost function (10.14) should be minimized. This minimization problem, however, currently is not mathematically tractable. To overcome this difficulty, from Eq. (10.17) we obtain an estimate of the direct regulation number through minimizing the cost function $J(k)$ defined as follows:

$$J(k) = \frac{\underline{\sigma}^2\left(\left[a_{j_1^{[\mathrm{opt}]}}\ a_{j_2^{[\mathrm{opt}]}}\ \cdots\ a_{j_k^{[\mathrm{opt}]}}\ b\right]\right)}{2\sigma^2} + \begin{cases} \gamma\,\mathbf{ln}(k_{\min}) & k \in [1,\ k_{\min}], \\ \gamma\,\mathbf{ln}(k) & k \in (k_{\min},\ n]. \end{cases} \tag{10.39}$$

Obviously,

$$J(k) \geq \min_{\substack{\varepsilon_A,\ \varepsilon_b \\ \text{subject to Eq. (10.13)}}} .l(\varepsilon_A,\ \varepsilon_b,\ k) \tag{10.40}$$

Therefore, minimization of $J(k)$ has an explanation of minimizing an upper bound of the cost function $l(\varepsilon_A,\ \varepsilon_b,\ k)$, which is equivalent to maximizing a lower bound of the likelihood function $L(\varepsilon_A,\ \varepsilon_b,\ k)$.

Note that both the parameter σ describing characteristics of measurement errors and the parameters $k_{\min}$ and γ describing the characteristics of the power law are prescribed positive numbers. This means that although it appears difficult to obtain an analytic form for the optimal k, this optimum can be obtained through linear searches for which many methods are available. On the other hand, it is obvious that the first term of the above cost function is a nonnegative and decreasing function of k, whereas the second term is a nonnegative and increasing function of k. We can expect that $J(k)$ usually only has one local minimum, which has been confirmed by extensive computation experience in [18].

From the above analysis the following algorithms are suggested in [18] for identifying direct effects in a large-scale cellular network.

Algorithm 10.2.1. Algorithm I for Cellular Network Inference

(1) *Initialize the vector x and the cost function $J(0)$ respectively as $x = 0$ and a large positive number, for example, $J(0) = 10^{100}$.*

(2) *Compute the value of the matrix $\hat{A}$ and the value of the vector $\hat{A}^T b$. Denote the ith row element of the vector $\hat{A}^T b$ by y_i and assume that $|y_{j_1}| \geq |y_{j_2}| \geq \cdots \geq |y_{j_{n-1}}|$.*

(3) *Construct the matrix $\tilde{A}_k = [a_{j_1}\ a_{j_2}\ \cdots\ a_{j_k}]$. Compute the value of the cost function $J(k)$ defined as*

$$J(k) = \frac{\underline{\sigma}^2([\tilde{A}_k\ b])}{2\sigma^2} + \begin{cases} \gamma\,\mathbf{ln}(k_{\min}), & 1 \leq k \leq k_{\min}, \\ \gamma\,\mathbf{ln}(k), & k_{\min} < k \leq n. \end{cases}$$

(4) *If $J(k) < J(k-1)$, replace k by $k+1$ and repeat Step 3. If $J(k) \geq J(k-1)$, then go to the next step.*

(5) *Perform singular value decomposition for the matrix $[\tilde{A}_{k-1}\ b]$. Denote its right singular vector associated with its minimum singular value $\underline{\sigma}([\tilde{A}_{k-1}\ b])$ by v.*

(6) *Let v_i be the ith row element of the vector v. Replace the j_ith row element of the vector x by $-\frac{v_i}{v_{k+1}}$, $i = 1, 2, \cdots, k-1$.*

Algorithm 10.2.2. Algorithm II for Cellular Network Inference

This algorithm is completely the same as that of Algorithm 10.2.1, except that in the Step 2, the matrix $\hat{A}$ is replaced by the matrix A.

These algorithms have been compared extensively in [18] with the total least squares method using several typical examples to illustrate their effectiveness, which include an artificially constructed large-scale linear system with 100 nodes, a mitogen activated protein kinase (MAPK) pathway model with 103 nodes described by a set of ordinary nonlinear differential equations, some *in silico* data sets from the DREAM initiative given in [4], and some *in vivo* data sets taken from [2,15].

In these comparisons, in addition to some widely adopted specifications in network inferences, such as ROC (Receiver Operating Characteristics) curve, PR (Precision Recall) curve, AUROC (Area under a ROC curve), AUPR (Area under a PR curve), PPV (Positive Predictive Value), Se (Sensitivity), FP (False Positive) rate [4,15,27], and so on., some other specifications are also adopted, which respectively reflect the rate of false-sign errors and parameter estimation accuracies. These specifications are used in performance evaluations with the objectives to clarify that in predicting the behaviors of a network, not only structure of the network, but also regulation directions and strengths among its elements, are also important.

These comparisons make it clear that the aforementioned methods have distinguished advantages on both reduction of false-positive errors and improvement of parametric estimation accuracy. Moreover, one of the algorithms has a much faster convergence speed when either estimation error or estimation bias is considered.

10.3 Absolute and Relative Variations in GRN Structure Estimations

In structure estimation for a gene regulation networks, the so-called $\mathcal{Z}$-score method has been proven to be very simple and efficient [4]. Basically, this method estimates possibilities of the existence of a direct regulation from one gene to another gene through measuring the variations of the gene expression level when a gene is perturbed by some external efforts, which include gene knock-out, gene knock-down, and so on. In particular, when a direct regulation from gene A to gene B is required to be investigated, an external perturbation is added at gene A, and the expression level of gene B is measured before and after that perturbation. Let $x_B^{[\mathrm{wt}]}$ and $x_{AB}^{[\mathrm{pb}]}$ denote respectively these two measured values. Moreover, assume that the variance of measurement errors is σ_B. Then, the $\mathcal{Z}$-score on the direct regulation from gene A to gene B, denoted $\mathcal{Z}_{A\to B}$, is calculated as

$$\mathcal{Z}_{A\to B} = \frac{x_{AB}^{[\mathrm{pb}]} - x_B^{[\mathrm{wt}]}}{\sigma_B}. \tag{10.41}$$

The probability of the existence of a direct regulation from gene A to gene B is considered to be proportional to the absolute value of this $\mathcal{Z}$-score $\mathcal{Z}_{A\to B}$. Clearly, this calculation is based on absolute variations of gene expression levels before and after external perturbations.

Although this method has a very low computational complexity and is computationally efficient in structure estimations for a large-scale gene regulation network, it usually causes false positive errors. In addition, to get necessary data for revealing the structure of a gene regulation network using this score, each of its genes must be perturbed, which is usually economically expensive. Furthermore, it is sometimes not very easy to get actual knowledge about the variance of measurement errors. To overcome these disadvantages, a relative-variation-based approach is suggested in [28] for GRN structure estimations. In this approach, each gene is assumed to have three types of expression level, that is, low expression level, high expression level, and wild-type expression level. The low expression level means that the expression of a gene is repressed, whereas a high expression level means that the expression of a gene is activated. The wild-type expression level indicates that the expression of a gene has not been influenced by external perturbations. This assumption is extensively regarded as appropriate, noting that in a GRN, direct regulations among genes can usually be divided into two types, repression effect and activation effect. This means that in a steady-state cellular system, a gene is in one of these states, and its expression level usually does not change significantly in each of these states [1,15,29].

Let $x_{AB}^{[\mathrm{wt},0]}$ and $x_{AB}^{[\mathrm{pb},0]}$ denote respectively the actual value of the expression level of gene B before and after gene A is externally perturbed. The relative variation of the expression level of gene B due to the external perturbation of gene A, denoted by $\delta_{A\to B}$, is calculated as

$$\delta_{A\to B} = \frac{x_B^{[\mathrm{pb},0]} - x_B^{[\mathrm{wt},0]}}{x_B^{[\mathrm{wt},0]}}. \tag{10.42}$$

Compared with Eq. (10.41), it is clear that rather than the measured gene expression level, its actual value is used in the calculation of this relative variation. Naturally, this actual value is not available, and it must be estimated from experimental data. On the other hand, similar to the $\mathcal{Z}$-score in GRN structure estimations, the probability of the existence of a direct regulation from gene A to gene B is considered to be proportional to the absolute value of this relative variation. In addition, its computational complexity is comparable to that of the $\mathcal{Z}$-score, provided that the associated gene expression levels are available.

In [28,30–32], a likelihood maximization method is suggested for the estimation of actual gene expression levels from experiment data, incorporating the sparsity of a large-scale system. In this method, measurement errors are assumed to be independent of each other in every experiment and are distributed normally with zero mean. It is also assumed that for each gene in the network, measurement errors have the same variance in all the experiments. Under these assumptions, a computationally attractive method is developed for these estimations. On the basis of these estimates, relative variations of gene expressions in a GRN are inferred, which is further applied to the structure identification of the associated GRN.

10.3.1 Maximum Likelihood Estimation for Wild-Type Expression Level and Measurement Error Variance

For a GRN consisting of n genes, assume that the steady-state expression level of a gene can only take three discrete values, which are respectively called the value of low expression level, the value of high expression level, and the value of wild-type expression level. Denote these three values of gene i, $i = 1, 2, \cdots, n$, respectively by $x_i^{[\mathrm{L},0]}$, $x_i^{[\mathrm{H},0]}$, $x_i^{[\mathrm{wt},0]}$. Moreover, assume that measurement errors in the expression level of gene i have an identical normal distribution $N(0, \sigma_i^2)$ in different experiments, which are independent of those of other genes in the same experiment and those of each gene in a different experiment.

Let x_{ji}, $i = 1, 2, \cdots, n$, denote the measured expression level of gene i when gene j is externally perturbed, $j = 1, 2, \cdots, n$. Moreover, let $x_i^{[\mathrm{wt}]}$ denote the measured expression level of gene i when there is no external perturbation to this GRN, $i = 1, 2, \cdots, n$. In this subsection, we discuss how to use these experiment data to infer the three expression levels for each gene in the GRN and the variance of their measure errors.

To derive estimates for $x_i^{[\mathrm{L},0]}$, $x_i^{[\mathrm{H},0]}$, $x_i^{[\mathrm{wt},0]}$ and σ_i, $i = 1, 2, \cdots, n$, we first investigate their estimation for a particular gene in a GRN under the condition that the in-degree of this gene is known.

For a specific gene, say, gene i, assume that there are respectively k_{i1} and k_{i2} other genes that have direct activation and repression effects on it. Then, under the adopted assumptions, a maximum likelihood estimate (MLE) can be obtained for different three expression levels of this gene and the variance of their measurement errors.

Theorem 10.4. *For some prescribed k_{i1} and k_{i2}, the MLEs for $x_i^{[\mathrm{L},0]}$, $x_i^{[\mathrm{H},0]}$, $x_i^{[\mathrm{wt},0]}$ and σ_i are respectively as follows:*

$$\hat{x}_i^{[\mathrm{L},0]} = \frac{1}{k_{i1}} \sum_{l=1}^{k_{i1}} x_{j_l,i}, \quad \hat{x}_i^{[\mathrm{H},0]} = \frac{1}{k_{i2}} \sum_{l=n-k_{i2}}^{n-1} x_{j_l,i}$$

$$\hat{x}_i^{[\mathrm{wt},0]} = \frac{1}{n-k_{i1}-k_{i2}} \left(\sum_{l=k_{i1}+1}^{n-k_{i2}-1} x_{j_l,i} + x_i^{[\mathrm{wt}]} \right) \tag{10.43}$$

$$\hat{\sigma}_i = \left\{ \frac{1}{n} \left[\sum_{l=1}^{k_{i1}} \left(x_{j_l,i} - \hat{x}_i^{[\mathrm{L},0]} \right)^2 + \sum_{l=k_{i1}+1}^{n-k_{i2}-1} \left(x_{j_l,i} - \hat{x}_i^{[\mathrm{wt},0]} \right)^2 + \left(x_i^{[\mathrm{wt}]} - \hat{x}_i^{[\mathrm{wt},0]} \right)^2 \right.\right.$$
$$\left.\left. + \sum_{l=n-k_{i2}}^{n-1} \left(x_{j_l,i} - \hat{x}_i^{[\mathrm{H},0]} \right)^2 \right] \right\}^{\frac{1}{2}} \tag{10.44}$$

The proof of Theorem 10.4 is deferred to the appendix of this chapter.

In GRN structure identification, k_{i1} and k_{i2} are generally not available. But from the sparsity of a large-scale GRN some statistical information about the in-degree of a gene, which equals to the sum of k_{i1} and k_{i2}, can be obtained. More specifically, let $k_{\max}$ with $k_{\max} < n$ denote the maximum in-degree of a GRN, and let n_k denote the number of genes with its in-degree k. Then, from the power law given by Eq. (10.11) we can declare that

$$n_k = \begin{cases} \left\lfloor nck_{\min}^{-\gamma} \right\rfloor, & 0 \le k < k_{\min}, \\ \left\lfloor nck^{-\gamma} \right\rfloor, & k_{\min} \le k \le k_{\max}, \\ 0, & k_{\max} < k \le n, \end{cases} \tag{10.45}$$

where $\lfloor * \rfloor$ stands for the operation of taking the nearest integer that is not greater than $*$.

When only the sum of k_{i1} and k_{i2}, denoted k_i, is available, the optimal k_{i1} and k_{i2} are obtained simply through searching the minimizer in the $k_i + 1$ combinations of k_{i1} and k_{i2} satisfying $k_{i1} + k_{i2} = k_i$ and $0 \le k_{i1} \le k_i$. Denote

$$-\ln F_i \left(\hat{x}_i^{[\mathrm{L},0]}, \hat{x}_i^{[\mathrm{H},0]}, \hat{x}_i^{[wt,0]}, \hat{\sigma}_i \,\middle|\, k_{i1}, k_{i2} \right)$$

with the optimal k_{i1} and k_{i2} by $f_{ki,i}$. Obviously,

$$\exp(-f_{ki,i}) = \max_{\substack{x_i^{[\mathrm{L},0]}, x_i^{[\mathrm{H},0]}, x_i^{[\mathrm{wt},0]}, \sigma_i \\ k_{i1},\, k_{i2} \\ k_{i1}+k_{i2}=k_i \\ 0 \le k_{i1} \le k_i}} F_i \left(x_i^{[\mathrm{L},0]}, x_i^{[\mathrm{H},0]}, x_i^{[\mathrm{wt},0]}, \sigma_i \,\middle|\, k_{i1}, k_{i2} \right)$$

Let $\mathcal{N}_k$ denote the set of genes that have an in-degree k, that is,

$$\mathcal{N}_k = \{\, i \mid i \in \{1, 2, \cdots, n\},\ k_i = k \,\}.$$

Recall that measurement errors for gene expression levels are assumed to be independent of each other. Let $\#(\cdot)$ denote the number of elements of a set. Obviously, the MLEs for gene expression levels and measurement error variances are those $\left(x_i^{[\mathrm{L},0]},\ x_i^{[\mathrm{H},0]},\ x_i^{[\mathrm{wt},0]},\ \sigma_i\right)_{i=1}^{n}$ that make the following maximization problem achieve its optimum:

$$\max_{x_i^{[\mathrm{L},0]},\ x_i^{[\mathrm{H},0]} x_i^{[\mathrm{wt},0]},\ \sigma_i} \prod_{i=1}^{n} F_i\left(x_i^{[\mathrm{L},0]}, x_i^{[\mathrm{H},0]}, x_i^{[\mathrm{wt},0]}, \sigma_i \;\middle|\; k_{i1} + k_{i2} = k\right) \tag{10.46}$$
$$\text{subject to } \#(\mathcal{N}_k) = n_k,\ k = 0, 1, \cdots, k_{\max}$$

From the definition of $f_{ki,i}$ we can straightforwardly prove that this maximization problem is equivalent to the following minimization problem:

$$\min_{\zeta_{ki}} \sum_{k=0}^{k_{\max}} \sum_{i=1}^{n} f_{k,i}\zeta_{ki} \tag{10.47}$$

$$\text{subject to} \begin{cases} \sum_{k=0}^{k_{\max}} \zeta_{ki} = 1, & i = 1, \cdots, n, \\ \sum_{i=1}^{n} \zeta_{ki} = n_k, & k = 0, \cdots, k_{\max}, \\ \zeta_{ki} \in \{0,\ 1\}, & k = 0, \cdots, k_{\max}, i = 1, \cdots, n. \end{cases} \tag{10.48}$$

Note that $f_{k,i}$ can be obtained through separately maximizing

$$F_i\left(x_i^{[\mathrm{L},0]}, x_i^{[\mathrm{H},0]}, x_i^{[\mathrm{wt},0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right)$$

for every individual gene. This minimization problem is in fact a 0–1 integer programming problem, for which various efficient algorithms have been developed, such as the so-called linear programming-based branch-and-bound algorithm and so on [33,34]. More precisely, it has been proven that in this minimization problem, the constraint $\zeta_{ki} \in \{0,\ 1\}$ can be replaced by $\zeta_{ki} \in (0,\ 1)$, which can significantly reduce computational complexity of this optimization.

Denote the optimizer of the minimization problem (10.48) by $\hat{\zeta}_{ki}$, $k = 0, \cdots, k_{\max}$ and $i = 1, \cdots, n$. For gene i, if $\hat{\zeta}_{ki} = 1$ with $0 \le k \le k_{\max}$, then from the above problem description it is clear that the optimal estimate for the in-degree of this gene is k. When this information is available, the optimal estimate can be obtained for both k_{i1} and k_{i2}, which further leads to the optimal estimate for the wild-type expression level $x_i^{[\mathrm{wt},0]}$ and the measurement error variance σ_i. In addition, a rough estimate can also be obtained about the topology of

GRN. As a matter of fact, we can declare from these minimization results that genes numbered $j_1, j_2, \cdots, j_{k_{i1}}$ will most likely to have a direct repression effect on gene i, whereas genes numbered $j_{n-k_{i2}}, j_{n-k_{i2}+1}, \cdots, j_{n-1}$ will have a direct activation effect.

It is worth noting that the power law is generally used to describe the statistics of direct regulations in a large-scale network [1,18,35,36], but the expression level of a gene may also be changed by indirect regulations. However, if we can assume that direct regulations usually cause large gene expression variations, then the aforementioned utilization of the power law seems reasonable in estimations of wild-type gene expression levels and measurement error variances. Although there are still no solid biological evidences on the appropriateness of this assumption, actual computation results in [31,32], which include both simulated data and actual data, show that this adoption of the power law is really helpful in estimation accuracy improvements.

10.3.2 Estimation of Relative Expression Level Variations

Define the relative expression level variation (RELV) of gene i resulted from an external perturbation on gene j, denoted by δ_{ji}, as

$$\delta_{ji} = \frac{x_{ji}^{[0]} - x_i^{[\mathrm{wt},0]}}{x_i^{[\mathrm{wt},0]}}. \tag{10.49}$$

We hope that RELV is more efficient than the well-known $\mathcal{Z}$-score in distinguishing direct and indirect causal regulations. The rationale for this expectation is as follows. If in a pathway of a GRN, every direct regulation has the property that a relative change of the concentrations of the proteins or mRNAs, and so on, related to the regulated gene is at most as large as that of the regulation gene, then it is obvious that the magnitude of this relative change due to an indirect regulation, which is in fact a cascade connection of several direct regulations, is certainly not greater than that due to a direct regulation. Although it is still not very clear whether or not this assumption is reasonable for every pathway from a biology viewpoint, our computation experiences show that this assumption may have some nice biological interpretations and is satisfied by most regulations existent in a GRN. These arguments also imply that the larger the magnitude of δ_{ji} is, the more unlikely that the expression level variation of gene i after knocking out/down gene j is due to indirect regulations, and thus the larger the probability that gene i is directly regulated by gene j.

To reduce influences of measurement errors, define $\bar{\delta}_{ji}$ as the expectation of the absolute value of δ_{ji}. Then direct computations show that

$$\begin{aligned}\bar{\delta}_{ji} &= \frac{1}{x_i^{[\mathrm{wt},0]}}\left\{\sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\left(x_{ji}-x_i^{[\mathrm{wt},0]}\right)^2 (2\sigma_i^2)\right)\right. \\ &\qquad \left. -\left(x_{ji}-x_i^{[\mathrm{wt},0]}\right)\left[1-2\Phi\left((x_{ji}-x_i^{[\mathrm{wt},0]})/\sigma_i\right)\right]\right\}, \end{aligned} \tag{10.50}$$

where $\Phi(\cdot)$ is the error function. As $x_i^{[\mathrm{wt},0]}$ and σ_i are generally not available, an estimate for $\bar{\delta}_{ji}$ should be used in GRN topology inference. To obtain this estimate, $x_i^{[\mathrm{wt},0]}$ is replaced by $\hat{x}_i^{[\mathrm{wt},0]}$ of Eq. (10.43), and σ_i is replaced by $\hat{\sigma}_i$ of Eq. (10.44). Therefore,

$$\begin{aligned}\hat{\bar{\delta}}_{ji} &= \frac{1}{\hat{x}_i^{[\mathrm{wt},0]}}\left\{\sqrt{\frac{2}{\pi}}\hat{\sigma}_i \exp\left(-\left(x_{ji}-\hat{x}_i^{[\mathrm{wt},0]}\right)^2/(2\hat{\sigma}_i^2)\right)\right. \\ &\qquad \left. -\left(x_{ji}-\hat{x}_i^{[\mathrm{wt},0]}\right)\left[1-2\Phi\left((x_{ji}-\hat{x}_i^{[\mathrm{wt},0]})/\hat{\sigma}_i\right)\right]\right\} \end{aligned} \tag{10.51}$$

Note that in GRN topology inferences, the larger the value of $\hat{\bar{\delta}}_{ji}$, the higher the probability for the existence of a direct regulation from gene j to gene i. Let Δ denote the $n \times n$-dimensional matrix with its jth row ith column element being $\hat{\bar{\delta}}_{ji}$ when $i \neq j$ and its diagonal element being zero. Then, it is clear that this matrix contains information about regulation strengths between any two different genes in a GRN.

However, to infer the structure of a GRN from this matrix, an important fact must be taken into account that in a GRN, efforts required to regulate different genes are not completely the same [1,15]. This implies that although under the adopted assumption, it can be proved that a direct regulation usually leads to a bigger magnitude of the RELV of the regulated gene than an indirect one, direct regulations to different genes may lead to different magnitude orders of this variation. Therefore, to obtain a good estimate from the matrix Δ about the topology of a GRN, an appropriate normalization is still required for the estimate $\hat{\bar{\delta}}_i$ among different genes.

Although the problem of making a biologically significant normalization for RELVs is still theoretically challenging, in this paper, it is suggested to normalize them using the Euclidean norm of a vector, which is widely adopted in many fields like system analysis and synthesis, signal processing, and so on [18,37,38]. To use the aforementioned rough estimation about the topology of the GRN, RELVs of a gene estimated to have a direct regulation are normalized differently from those of a gene estimated not to have a direct regulation. Specifically, the j_lth row ith column element of the matrix Δ, that is, $\hat{\bar{\delta}}_{j_l,i}$, is normalized as

$$\delta_{j_l,i}^{[2]} = \begin{cases} \hat{\bar{\delta}}_{j_l i} \left(\sum_{s=k_1+1}^{n-k_2-1} \hat{\bar{\delta}}_{j_s,i}^2 \right)^{-\frac{1}{2}}, \ l = k_1 + 1, \cdots, n - k_2, \\ \hat{\bar{\delta}}_{j_l,i} \left(\sum_{s=1}^{k_1} \hat{\bar{\delta}}_{j_s,i}^2 + \sum_{s=n-k_2}^{n-1} \hat{\bar{\delta}}_{j_s,i}^2 \right)^{-\frac{1}{2}} \quad \text{otherwise.} \end{cases} \tag{10.52}$$

For presentation conciseness, denote the normalized matrix Δ using the Euclidean norm by $\Delta^{[2]}$.

Another important thing worth of consideration in GRN topology estimation is that genes estimated to have a direct regulation should correspond to a RELV with a magnitude greater than those estimated not to have a direct regulation. To achieve this purpose, the following adjustment is suggested in this paper. Define

$$\delta_0^{[2]} = \max_{1 \le i \le n} \max_{k_1+1 \le l \le n-k_2-1} \delta_{j_l,i}^{[2]} \tag{10.53}$$

With this value, the normalized RELVs for an arbitrary gene i are adjusted as follows:

$$\tilde{\delta}_{j_l,i}^{[2]} = \begin{cases} \delta_{j_l,i}^{[2]}, & k_1 + 1 \le l \le n - k_2 - 1, \\ \delta_{j_l,i}^{[2]} + \delta_0^{[2]} & \text{otherwise.} \end{cases} \tag{10.54}$$

Denote by $\tilde{\Delta}^{[2]}$ the $n \times n$-dimensional matrix with its jth row ith column element being $\tilde{\delta}_{ji}^{[2]}$. Elements of this matrix are directly used to infer the structure of a GRN. The greater the jth row ith element, the higher the probability that gene i is directly regulated by gene j.

10.3.3 Estimation Algorithm

In summary, on the basis of likelihood maximization and the concept of the RELV of a gene, the following algorithm is suggested in this paper for identifying direct regulations of a GRN, which consists of three main steps:

- Using available a priori information about the GRN under investigation, choose appropriate values of the parameters γ, $k_{\min}$, and $k_{\max}$ of the power law used to describe its sparsity. On the basis of Theorem 10.4, calculate $f_{ki,i}$ for all $i = 1, 2, \cdots, n$ and $k_i = 0, 1, \cdots, k_{\max}$.
- Solve the constrained minimization problem of Eq. (10.48). Using the estimated σ_i and $x_i^{[\mathrm{wt},0]}$ and Eqs. (10.50) and (10.52), calculate the matrix $\Delta^{[2]}$ consisting of the normalized magnitudes of the estimates of the RELVs of every gene in a GRN.

- Modify the matrix $\Delta^{[2]}$ according to Eqs. (10.53) and (10.54). Using elements of these modified matrices, queue possibilities of the existence of a direct regulation from the gene with the same number of the row to the gene with the same number of the column. The greater the element, the higher the confidence for the existence.

Actual computation results with the size 100 subchallenges of both DREAM3 and DREAM4, reported in [28,30–32], show that this method can outperform not only the widely used $\mathcal{Z}$-score-based method, but also the best team of these subchallenges who used an integration of some well-known methods. Precision analysis shows that compared with the widely adopted $\mathcal{Z}$-score-based method, highly confident predictions obtained by this method usually have a higher precision and therefore are more helpful in guiding designs of biology validation experiments. In addition to these, estimates obtained by this method are more accurate also for both wild-type gene expression levels and measurement error variances.

Another important property of the suggested algorithm is that its computational complexity increases only polynomially with the number of genes of a GRN, which is in sharp contrast to many other available methods whose computational complexity increases exponentially.

10.4 Estimation With Time Series Data

In the previous two sections, we investigated structure identification for a large-scale networked system using steady-state experimental data. In addition to this method, there are also various other approaches proposed to deal with this problem. For example, the Boolean network model-based method [7,39], the approach based on the correlation or partial correlation coefficients suggested in [40], the mutual information-based method [41], and their refinements suggested in [42–46]. Compared with these methods, an important characteristic of the algorithms in Sections 10.2 and 10.3 is that they have explicitly taken the sparsity of a large-scale system into account, which is helpful in improving estimation accuracies.

However, all these methods can model only static relations among subsystems. In various practical applications, dynamic experimental data are also able to be collected. Compared to steady-state experiment data, it is widely believed that dynamic experimental data contain more information about the system under investigation. An interesting theoretical issue is therefore to explore possibilities of estimating subsystem interactions in a networked system using time series experimental data. In this section, will discuss this problem for a gene regulation network in which dynamic interactions among genes are nonlinear. Although the discussed plant is restricted to cellular networks, the method itself is believed to be extendable to other types of plants.

To facilitate information extraction from time series expression profiles in structure identification of gene regulation networks, various dynamical models have been developed, such as dynamic Boolean networks [47], neural networks [48], Bayesian networks [29,49], and so on. Among the statistical techniques currently adopted in modeling GRNs, Bayesian inferences have received the most widespread attentions [11,50,51]. Under the dynamic Bayesian regime, the model of GRNs is extensively considered as a state-space model, which consists of gene expression measurement equations and gene regulation equations [11,13]. In this state-space model, gene expression values are assumed to depend not only on the current cellular states but also on external inputs or disturbances, which reflects the nature of a dynamic network. In the early works, it was generally assumed that gene regulations can be described by linear differential/difference equations, and the well-known Kalman filter is used to recover the structure of a GRN [11]. However, due to the inherent nonlinear nature of GRNs, there exist some restrictions when a linear model is applied to describe gene behaviors [52]. In short, linear approximation is valid only when a GRN has slow dynamics around its steady state. To capture complex gene interactions more efficiently, it is crucial to alleviate this linearity assumption. One way to make the GRN model more appropriate is to include nonlinear terms, such as the so-called S-system [53], sigmoid function [1,54–56], and so on.

When a nonlinear state-space model is adopted, the extended Kalman filter (EKF) is one efficient method for GRN structure recovering [55,56]. The EKF-based approach works well with both steady-state data and slow dynamical data. On the other hand, there may occur considerable performance deteriorations in this approach if either the initial state estimate is incorrect or there are appreciable deficiencies in the system model caused by first-order approximations [56]. More specifically, as the EKF-based approach does not take either unmodeled dynamics or parametric uncertainties into account, its estimation performances may not be satisfactory due to its slow convergence speed, which usually leads to low estimation accuracy. Mistakes are often caused by the low estimation accuracy of an estimation algorithm. For example, in inferring the structure of a GRN, an estimated parameter, say $\hat{g}^{[ij]}$, is often used to decide whether gene j directly regulates gene i. A false positive error is made when the actual value of $g^{[ij]}$ equals zero, but its estimate $\hat{g}^{[ij]}$ has a large magnitude. This means that unmodeled dynamics and parametric uncertainties should not be ignored in identifications.

To enhance estimation performances, GRN structure recovering is resorted in [24] to the robust state estimator suggested in [57] after the first-order approximation of GRNs. The suggested method is guaranteed to be robust against model errors due to GRN linearizations and state estimate inaccuracies, in which both parametric modeling errors and unmodeled dynamics are included. It has been proven that the estimated network topology by either the EKF-based method or the method suggested in [24] converges in the mean square sense to the actual network structure. An attractive property of this estimation algorithm, however, is that, on the basis of the specific structure of the network topology inference problem itself, it has been shown that under some weak requirements, the convergence speed of the suggested method can be guaranteed to be faster than that of the EKF-based method. This is quite important, noting that the data length in a biological experiment is usually not very long.

10.4.1 Robust Structure Identification Algorithm for GRNs

According to chemical principles, such as the Michaelis–Menten kinetics and so on, dynamic reactions occurred in a practical biochemical network are inherently nonlinear, which means that GRNs must be treated in general as a nonlinear dynamic system [1,7,39,53]. On the other hand, an extensively adopted way in dealing with dynamic systems is the so-called state-space approach. In particular, a nonlinear state evolution equation for GRNs consisting of n genes can be described by[2]

$$x(k+1) = f\left(x(k), \theta\right) + w(k), \tag{10.55}$$

where k stands for the temporal variable, $x(k) = col\{x_{k,i}|_{i=1}^{n}\}$ is the vector consisting of expression levels of all the genes, $f\left(x(k), \theta\right)$ is a vector of nonlinear functions, $\theta \in \mathbb{R}^p$ is a vector consisting of unknown parameters, and $w(k)$ denotes a noise vector, which is usually assumed to be uncorrelated and normally distributed with zero mean and covariance matrix $Q(k)$. Moreover, an m-dimensional vector of measurements $y(k)$, for example, microarray data, is related to the directly unobservable hidden state variables through the following observation equation:

$$y(k) = h\left(x(k), \theta\right) + v(k). \tag{10.56}$$

Here $h\left(x(k), \theta\right)$ is once again a vector of nonlinear functions, and $v(k)$ is a zero-mean uncorrelated Gaussian noise vector with covariance matrix $R(k)$. An extensively adopted assumption is that the random processes $w(k)|_{k=0}^{\infty}$ and $v(k)|_{k=0}^{\infty}$ are white and mutually independent.

The function $f\left(\cdot, \cdot\right)$ in Eq. (10.55) and the function $h\left(\cdot, \cdot\right)$ in Eq. (10.56) are quite general in describing regulation relationships among various genes in a GRN and measurements of their expression levels. These associated equations are able to approximately model most dynamic GRNs. However, one of their serious drawbacks is that they are mathematically very difficult, if not infeasible, to be used in handling problems like parameter estimation, structure identification, and so on. To make these problems mathematically tractable, some particular nonlinear functions are adopted. Among these functions, the most widely utilized one appears to be the sigmoid function, as it naturally reflect many characteristics of interactions occurred in a cell, such as the speedy change between expression and unexpression states of a gene, the

[2] It is worth emphasizing that in actual GRNs, rather than to directly act on another gene, a gene exerts its influence through its mRNAs, proteins, and so on. However, when relations among genes are discussed, models like those in Eqs. (10.55) and (10.57) are usually adopted [4,39,54,58].

ratio of chemical reactions depending on concentrations of the involved materials, and so on. Another attractive property of this function is its analyticity and simplicity. More specifically, the following model is widely adopted in describing state evolutions in a GRN [1,54–56]:

$$x(k+1) = G\,S\,(x(k)) + w(k), \tag{10.57}$$

where

$$S(x(k)) = \left[\frac{1}{1+\exp\left(-x_{k,1}\right)}, \ldots, \frac{1}{1+\exp\left(-x_{k,n}\right)}\right]^T.$$

The matrix $G \in \mathbb{R}^{n\times n}$ in this equation captures causal regulation relationships among genes, that is, if $g^{[ij]}$, the ith row jth column element of this matrix, has a magnitude significantly greater than zero, then, the jth gene directly regulates the ith gene. This means that this state-space model produces a directed graph among different genes in a GRN. Moreover, the microarray data obtained at the kth time instant can usually be simply described as

$$y(k) = x(k) + v(k). \tag{10.58}$$

The system model described by Eqs. (10.57) and (10.58) embraces a group of important features of GRNs, such as direct and causal gene regulations, nonlinear chemical reactions, dynamic gene expressions, external noise influences, microarray data, and so on. It also reflects the widely adopted approximation that each gene of a GRN is generally in one of two states, that is, expressed and unexpressed, and changes between these two states are usually very fast [1,39,53,54].

Motivated by these biological observations, the GRN identification problem discussed in this section is as follows. Given a set of noisy microarray data $y(k)\,|_{k=1}^{t}$, which are assumed to be generated by the state-space model described by Eqs. (10.57) and (10.58), recover the structure of GRNs through estimating the unknown but time-invariant matrix G.

It is worth mentioning that an accurate estimate of the matrix G enables us not only to reveal the actual GRN topology, but also to quantify interaction strengths among genes. In addition, direct comparison of Eqs. (10.55) and (10.57) shows that $\theta = vec(G)$.

Major challenges in inferring the nonlinear GRN model include estimation of both the states and parameters of the systems from noisy observations and that experimental data length is generally short and states/parameters to be estimated are usually of great amount.

In simultaneous estimation of system states and system parameters of a nonlinear dynamic system, a widely adopted method is using the EKF by taking the system parameters as additional states and augmenting state equations [55,59]. More specifically, let $z(k)$ denote the vector consisting of all gene expression levels of the GRN at the time instant k and all elements of the system parameter matrix G, that is,

$$z(k) \triangleq col\{x(k),\ Vec(G)\}.$$

Let $g\left(z(k)\right)$ and $C(k)$ represent respectively $\begin{bmatrix} GS\left(x(k)\right) \\ \theta(k) \end{bmatrix}$ and $\begin{bmatrix} I_{n\times n}, & 0_{n\times n^2} \end{bmatrix}$. Then, the augmented state-space model of the GRN described by Eqs. (10.57) and (10.58) can be reexpressed as

$$z(k+1) = \begin{bmatrix} x(k+1) \\ \theta(k+1) \end{bmatrix} = g\left(z(k)\right) + \eta(k), \tag{10.59a}$$

$$y(k) = C(k)z(k) + v(k), \tag{10.59b}$$

where

$$\eta(k) = \begin{bmatrix} w(k)^T, & \xi(k)^T \end{bmatrix}^T,$$

in which $\xi(k)$ is a zero-mean uncorrelated Gaussian noise vector with covariance matrix $\Phi(k)$.

In these expressions, the vector $\theta = vec(G)$ is replaced by $\theta(k)$, and an artificial disturbance $\xi(k)$ has been added for guaranteeing the well-posedness of the recursive calculation in the estimations. This is an approach extensively adopted in state estimation-based parameter identifications [38,60].

To employ a linear state estimation algorithm, for example, the Kalman filter or a linear robust state estimator, an unavoidable step is to have a linear state-space model. For this purpose, the nonlinear Eq. (10.59) needs to be first approximated by a linear equation. A widely adopted way is linearizing this nonlinear equation in a neighbor of the value of its current state estimate. Let $\hat{z}_{k|k}$ and $\hat{x}_{k,i}$ respectively represent an estimate of $z(k)$ and an estimate of $x_{k,i}$ at the time instant k. To realize the aforementioned idea, define the vector

$$\hat{g}\left(z(k)\right) = g\left(\hat{z}_{k|k}\right) + A(k, \hat{z}_{k|k})\left(z(k) - \hat{z}_{k|k}\right), \tag{10.60}$$

where

$$A(k, \hat{z}_{k|k}) = \left.\frac{\partial g\left(z\right)}{\partial z}\right|_{z=\hat{z}_{k|k}} = \begin{bmatrix} \hat{G}(k)W(k) & \left[\frac{1}{1+e^{-\hat{x}_{k,1}}}I_{n\times n}, \cdots, \frac{1}{1+e^{-\hat{x}_{k,n}}}I_{n\times n}\right] \\ 0_{n^2\times n} & I_{n^2\times n^2} \end{bmatrix} \tag{10.61}$$

and

$$W(k) = diag\left(\frac{e^{-\hat{x}_{k,1}}}{\left(1+e^{-\hat{x}_{k,1}}\right)^2}, \cdots, \frac{e^{-\hat{x}_{k,i}}}{\left(1+e^{-\hat{x}_{k,i}}\right)^2}, \cdots, \frac{e^{-\hat{x}_{k,n}}}{\left(1+e^{-\hat{x}_{k,n}}\right)^2}\right). \tag{10.62}$$

Then, a linear approximation of the nonlinear Eq. (10.59a) can be written as

$$z(k+1)=A(k,\hat{z}_{k|k})z(k)+\underbrace{\left(g\left(\hat{z}_{k|k}\right)-A(k)\hat{z}_{k|k}\right)}_{\text{known at time k}}+\eta(k), \tag{10.63}$$

which is a linear state-space model for the augmented state vector $z(k)$.

Based on the linear model (10.63), the Kalman filter can be directly applied to estimate the values of the parameter matrix G. These treatments are essentially the basic ideas developed in the EKF-based approach to nonlinear system identifications, for which rich literature exists, and examples include [55] and [59]. In general, this approach works well with steady-state data and/or systems with slow dynamics. On the other hand, there may exist considerable performance degradations if either the initial state estimate is incorrect, and/or there are appreciable deficiencies in modeling a dynamic system caused by the above first-order approximation [56]. Moreover, due to its slow convergence speed, which usually leads to a low estimation accuracy with finite experimental data, estimation performances of the EKF-based approach may be not very satisfactory in GRN topology identifications.

Studies in [24] show that one important reason for performance degradation of the EKF-based approach is the uncertainties of the linear model (10.63). Note that this linear model is obtained through the first-order approximation of the nonlinear model (10.59). As a result, the linear model (10.63) can hardly include all the dynamic features of actual GRNs. In other words, unmodeled dynamics usually exist in the errors of this linear model. On the other hand, note that the system matrix $A(k,\hat{z}_{k|k})$ in this linear model depends on the state estimate $\hat{z}_{k|k}$, which is generally different from the actual values of the system state at the time instant k. This means that parametric uncertainties also exist in the errors of this linear model (10.63).

In conclusion, due to the adopted linearization method, there inevitably exist both unmodeled dynamics and parametric uncertainties in the errors of this linear model (10.63). To reduce their influences on the accuracy of the GRN structure inference, these modeling errors should be explicitly taken into account in developing an estimation algorithm. More precisely, a desirable estimation algorithm is required to be robust against both the modeling errors due to linearizations and the modeling errors due to the state estimation errors.

For this purpose, the dynamics of the linear model (10.63) is modified to the following state-space model:

$$z(k+1)=A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)z(k)+a(k)+\eta(k), \tag{10.64}$$

where $\varepsilon(k)$ represents the difference between the actual value of the augmented plant state vector and its estimate at the time instant k, and $a(k)$ stands for $g\left(\hat{z}_{k|k}\right)-A(k,\hat{z}_{k|k})\hat{z}_{k|k}$. From their definitions we can straightforwardly prove that $A\left(k,\hat{z}_{k|k},0\right)=A(k)(\hat{z}_{k|k})$.

Remark 10.1. *From the definition of the matrix $A(k,\hat{z}_{k|k})$ it is clear that it varies with the estimate $\hat{z}_{k|k}$. To take into account the effects of the state estimation errors, the vector $\hat{z}_{k|k}$ in $A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)$ is replaced by $\hat{z}_{k|k}+\varepsilon(k)$, in which $\varepsilon(k)$ stands for the estimation error on the augmented plant state vector. To make the corresponding estimation problem mathematically tractable, elements of state estimation errors, that is, $\varepsilon_{k,i}$, $i=1,2,\cdots,n+n^2$, are assumed to be independent of each other. More precisely, the matrix $A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)$ is defined completely in the same way as $A(k,\hat{z}_{k|k})$ in Eq. (10.61):*

$$A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)=\left.\frac{\partial g(z)}{\partial z}\right|_{z=\hat{z}_{k|k}+\varepsilon(k)}. \tag{10.65}$$

This modification on the linear model (10.63) makes the assumption of [57] valid that the matrix $A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)$ is a known differentiable function of system parameter uncertainties.

Note that actual values of the augmented plant state vector are not available in general. To employ the robust state estimator introduced in [57], it is necessary to perform a first-order approximation on the matrix $A\left(k,\hat{z}_{k|k},\varepsilon(k)\right)$. For this purpose, the following matrices $\Lambda(k)$, $\Omega_{k,i}$, $\Xi_{k,i}$, and $\Delta_{k,i}$ are first defined:

$$\begin{aligned}
\Lambda(k) &= \left[\frac{e^{-\hat{x}_{k,1}}\left(-1+e^{-\hat{x}_{k,1}}\right)}{\left(1+e^{-\hat{x}_{k,1}}\right)^3},\cdots,\frac{e^{-\hat{x}_{k,n}}\left(-1+e^{-\hat{x}_{k,n}}\right)}{\left(1+e^{-\hat{x}_{k,n}}\right)^3}\right]^T,\\
\Delta_{k,i} &= \begin{bmatrix}\hat{G}(k)\Omega_{k,i} & \Xi_{k,i}\\ 0_{n^2\times n} & 0_{n^2\times n^2}\end{bmatrix},\\
\Omega_{k,i} &= diag\left(0,\cdots,0,\Lambda_{k,i},0,\ldots 0\right),\\
\Xi_{k,i} &= \left[0_{n\times n},\cdots,0_{n\times n},W_{k,ii}I_{n\times n},0_{n\times n},\cdots,0_{n\times n}\right]^T,
\end{aligned}$$

where $i=1,\ldots,n$. On the basis of these matrices, we construct the other matrix

$$\bar{S}_{k,i}=\begin{bmatrix}0_{n\times(n^2+n)}\\ C(k)\Delta_{k,i}\end{bmatrix}, i=1,\cdots,n. \tag{10.66}$$

Moreover, define the vector λ_j and matrices Γ_j and $\Theta_{k,j}$ as follows:

$$\begin{aligned}
\lambda_j &= [0,\cdots,0,1,0,\cdots,0]^T,\\
\Gamma_j &= reshape\left(\lambda_j,n,n\right),\\
\Theta_{k,j} &= \begin{bmatrix}\Gamma_j W(k) & 0_{n\times n^2}\\ 0_{n^2\times n} & 0_{n^2\times n^2}\end{bmatrix},
\end{aligned}$$

where $j = 1, \cdots, n^2$, and $reshape(j, n, n)$ represents constructing an $n \times n$ matrix from the vector λ_j whose elements are taken columnwise from that vector. Utilizing these vector and matrices, define the matrix

$$\overline{SS}_{k,j} = \begin{bmatrix} 0_{n\times(n^2+n)} \\ C(k)\Theta_{k,j} \end{bmatrix}, j = 1, \cdots, n^2. \tag{10.67}$$

Finally, on the basis of the matrices $\bar{S}_{k,j}$ and $\overline{SS}_{k,j}$, define the matrix

$$S(k) = col\left(col\left(\bar{S}_{k,i}\big|_{i=1}^{n}\right), \ col\left(\overline{SS}_{k,j}\big|_{j=1}^{n^2}\right)\right). \tag{10.68}$$

Then, from these definitions and constructions direct algebraic manipulations show that

$$\begin{aligned} A\left(k, \hat{z}_{k|k}, \varepsilon(k)\right) &= A\left(k, \hat{z}_{k|k}\right) + \sum_{j=1}^{n} \bar{S}_{k,j}(x_{k,j} - \hat{x}_{k,j}) + \sum_{j=1}^{n^2} \overline{SS}_{k,j}(\theta_j - \hat{\theta}_j) \\ &\quad + O\left(||z(k) - \hat{z}_{k|k}||_2^2\right) \\ &= A\left(k, \hat{z}_{k|k}\right) + (\varepsilon(k) \otimes I_{2n\times 2n})^T S(k) + O\left(||z(k) - \hat{z}_{k|k}||_2^2\right). \end{aligned} \tag{10.69}$$

Clearly, the matrices $\bar{S}_{k,j}$, $\overline{SS}_{k,j}$, and $S(k)$ are constituted from the first-order derivatives of the matrix $A\left(k, \hat{z}_{k|k}, \varepsilon(k)\right)$ with respect to estimation errors on gene expression values, GRN topologies, and the augmented plant states, respectively.

Based on all these discussions and definitions, a robust structure identification algorithm for GRNs is obtained as follows, which simply replaces the Kalman filter in the EKF-based method with the robust state estimator suggested in [57], where sensitivities are penalized of the innovation process in state estimations to modeling errors with the purpose of increasing the robustness of the estimate against parametric modeling errors.

- Initialization. Designate $P_{0|0}$ and $\hat{z}_{0|0}$ respectively as $P_{0|0} = \left(\pi_0^{-1} + C_0^T R_0^{-1} C_0\right)^{-1}$ and $\hat{z}_{0|0} = P_{0|0} C_0^T R_0^{-1} y_0$, where $\pi_0 = E\left[z_0 z_0^T\right] = diag\{\Pi_0, \ \pi_\theta\}$.
- Parameter modification. Define the matrix $S(k)$ as in Eq. (10.68). Moreover, define the matrices $\hat{A}(k, 0)$, $\hat{P}_{k|k}$, and $\Psi(k)$ as

$$\hat{A}(k, 0) = A\left(k, \hat{z}_{k|k}, 0\right)\left[I - \frac{(1-\gamma(k))}{\gamma(k)} \hat{P}_{k|k}\right], \tag{10.70a}$$

$$\hat{P}_{k|k} = \left(P_{k|k}^{-1} + \frac{(1-\gamma(k))}{\gamma(k)} S^T(k) S(k)\right)^{-1}, \tag{10.70b}$$

$$\Psi(k) = \begin{bmatrix} Q(k) & \\ & \Phi(k) \end{bmatrix}. \tag{10.70c}$$

- Plant state estimate updating. Calculate $\hat{z}_{k+1|k+1}$ and $P_{k+1|k+1}$ as

$$\begin{aligned} \hat{z}_{k+1|k+1} =& \hat{A}(k,0)\,\hat{z}_{k|k} + a(k) + P_{k+1|k+1} C^T(k+1) R(k+1)^{-1} \\ & \times \left[y(k+1) - C(k+1) \left(\hat{A}(k,0)\,\hat{z}_{k|k} + a(k) \right) \right], \end{aligned} \tag{10.71a}$$

$$P_{k+1|k} = A\left(k, \hat{z}_{k|k}, 0\right) \hat{P}_{k|k} A^T(k)\left(\hat{z}_{k|k}, 0\right) + \Psi(k), \tag{10.71b}$$

$$R_{e,k+1} = R(k+1) + C(k+1) P_{k+1|k} C^T(k+1), \tag{10.71c}$$

$$P_{k+1|k+1} = P_{k+1|k} - P_{k+1|k} C^T(k+1) R_{e,k+1}^{-1} C(k+1) P_{k+1|k}. \tag{10.71d}$$

Remark 10.2. *In this estimation procedure, $\gamma(k)$ is a design parameter belonging to* $[0,\ 1]$ *and taking a balance between the importance of nominal estimation performances and that of reducing estimation performance degradations due to modeling errors. The greater this parameter, the more important the nominal estimation performances. In the extreme case, that is, where $\gamma(k) = 1$ and/or $\frac{\partial A\left(k, \hat{z}_{k|k}, \varepsilon(k)\right)}{\partial \varepsilon_{k,i}} \equiv 0$, it is proved in [57] that the state estimator in the algorithm reduces to the well-known Kalman filter. This means that, under such situations, the above GRN structure estimation algorithm is equal to the EKF-based method. Although it is still theoretically not clear how to select this design parameter, a physically significant $\gamma(k)$ should generally satisfy $\gamma(k) \geq 0.5$.*

10.4.2 Convergence Analysis of the Robust Structure Identification Algorithm

In evaluating performances of an identification algorithm, one extensively utilized metric is related to its convergence. Although there are many works in the literature addressing the problem of estimating parameters of a nonlinear biochemical network using the EKF-based method, for example, [55] and [59], none of them provides convergence conditions. In this subsection, we derive some convergence conditions for the suggested robust structure identification algorithm and for the EKF-based method.

For brevity, we abbreviate the algorithm derived in the previous subsection as the RSE (robust state estimator) based method.

In probability theory, there exist several different notions about convergence of random variables. Examples include convergence in probability, mean-square convergence, convergence with probability 1, and so on [61,62]. In this subsection, mean-square convergence is adopted in the investigation of the properties of the suggested robust GRN structure identification method. Therefore, the definition of convergence in the mean-square sense is given in Definition 2.14, which is adopted from [62].

The recursive form of the RSE-based algorithm is very similar to that of the EKF-based method. Therefore, we first investigate convergence conditions of the EKF-based method and then turn to the suggested method.

An implicit assumption adopted in the EKF-based estimation method is that local linearizations are accurate enough in describing plant nonlinear dynamics in a neighborhood of the linearization point. Under such a situation, when $w(k)$ and $v(k)$ are random samples from some zero-mean Gaussian distributions and are statistically independent of each other, then the distribution of the estimate $\hat{z}_{k|k}$ can be effectively approximated by a normal distribution [63,64]. Two formulations are extensively utilized for the EKF-based estimation processes. One is the so-called two-step recursion formulation, which consists of a time-update step and a measurement-update step with a relinearization between these two steps. The other is a one-step formulation in terms of some a priori variables. Previous studies, for example, those in [63] and [65], have made it clear that these two formulations may have different steady-state performances and transient behaviors, but their convergence properties are the same. As a result, the convergence analysis of the EKF-based method is investigated in this subsection utilizing the one-step formulation, whose description is given explicitly as follows.

The EKF-Based Estimation Algorithm. The structure identification algorithm for nonlinear GRNs described by Eqs. (10.57) and (10.58) using the one-step EKF is constituted from the following four recursive steps:

- Linearization at an estimated state:

$$
\begin{aligned}
A(k,\hat{z}(k)) &= \left.\frac{\partial g\,(z)}{\partial z}\right|_{z=\hat{z}(k)} \\
&= \begin{bmatrix} \hat{G}(k)W(k) & \left[\frac{1}{1+e^{-\hat{x}_{k,1}}}I_{n\times n},\cdots,\frac{1}{1+e^{-\hat{x}_{k,n}}}I_{n\times n}\right] \\ 0_{n^2\times n} & I_{n^2\times n^2} \end{bmatrix}.
\end{aligned} \tag{10.72}
$$

- Compute the gain matrix of the Kalman filter:

$$
K(k) = A(k,\hat{z}(k))P(k)C^T(k)\Big(C(k)P(k)C^T(k)+R(k)\Big)^{-1}. \tag{10.73}
$$

- Update state estimates according to the following difference equation:

$$
\hat{z}(k+1) = g\left(\hat{z}(k)\right) + K(k)\left(y(k) - C(k)\hat{z}(k)\right). \tag{10.74}
$$

- Update the pseudo-covariance matrix of estimation errors through the following Riccati difference recursions:

$$
P(k+1) = A(k,\hat{z}(k))P(k)A^T(k,\hat{z}(k)) + \Psi(k) - K(k)\Big(C(k)P(k)C^T(k)+R(k)\Big)K^T(k). \tag{10.75}
$$

In the above descriptions, to distinguish the EKF-based method from that based on the robust state estimator, different symbols are adopted for state estimates and the pseudo-covariance matrix of estimation errors.

Recall that the augmented state vector $z(k)$ is defined as $z(k) = \left[x(k)^T, \theta(k)^T\right]^T$. Accordingly, estimation errors on the network topology, denoted $\tilde{\theta}(k)$, are given by

$$\tilde{\theta}(k) = \theta(k) - \hat{\theta}(k). \tag{10.76}$$

Besides, in line with the dimensions of the vectors $x(k)$ and $\theta(k)$, the pseudo-covariance matrix $P(k)$ can be partitioned into four submatrices:

$$P(k) = \begin{bmatrix} P_{x,k} & P_{x\theta,k} \\ P_{x\theta,k}^T & P_{\theta,k} \end{bmatrix}. \tag{10.77}$$

Note that in the state-space model (10.59a), the evolution of $\theta(k)$ is linear. This means that although the matrix $P(k)$ itself is a pseudo-covariance matrix, its 2nd block row 2nd block column submatrix $P_{\theta,k}$ is a covariance matrix. Substituting Eqs. (10.72) and (10.73) into Eq. (10.75), we obtain the following relation through some direct matrix operations:

$$P_{\theta,k+1} = P_{\theta,k} + \Phi(k) - P_{x\theta,k}^T\big(P_{x,k} + R(k)\big)^{-1} P_{x\theta,k}. \tag{10.78}$$

Let $\Sigma(k) = P_{x\theta,k}^T\big(P_{x,k} + R(k)\big)^{-1} P_{x\theta,k} - \Phi(k)$. Then, by this equality we are able to establish the convergence of the EKF-based estimation algorithm.

Theorem 10.5. *Assume that the following conditions are satisfied:*

$$P_{x\theta,k}^T\big(P_{x,k} + R(k)\big)^{-1} P_{x\theta,k} - P_{\theta,k} < \Phi(k) < P_{x\theta,k}^T\big(P_{x,k} + R(k)\big)^{-1} P_{x\theta,k}, \tag{10.79a}$$

$$\sum_{k=1}^{\infty}\left\{\lambda_{\min}^2\left(P_{\theta,k}^{-1}\right)\lambda_{\min}\left(\Sigma(k)\right)\right\} = \infty, \tag{10.79b}$$

$$A(k, \hat{z}(k)) \text{ is nonsingular for every } k \geq 0. \tag{10.79c}$$

Moreover, assume that there exist real positive constants $\bar{p}$ and $\underline{p}$, such that

$$\bar{p}I \leq P(k) \leq \underline{p}I. \tag{10.79d}$$

Then $\tilde{\theta}(k) \xrightarrow{m.s.} 0$.

The proof of this theorem is deferred to the appendix of this chapter.

Remark 10.3. $\tilde{\theta}(k) \xrightarrow{m.s.} 0$ *means that the estimate* $\hat{\theta}(k)$ *converges to* θ *in the mean square sense. Condition (10.79d) is closely related to the observability of a nonlinear dynamic system. More precisely, from Lemma 4.1 of [65] we know that if the matrix pair* $\left\{A(k, \hat{z}(k)),\ C(k)\right\}$ *satisfies a uniform observability condition, then condition (10.79d) is also satisfied. On the other hand, in addition to given conditions for the convergence of the EKF-based method, condition (10.79) also provides some insights on how to select the design parameter matrix* $\Phi(k)$*. In practice,* $\Phi(k)$ *is usually set as a constant matrix, such as* $\Phi(k) = rI$*, and the value of* r *is selected as a small positive number, for example,* $r = 10^{-5}$*. A more detailed discussion on how to choose appropriate parameters of a recursive estimation algorithm can be found, for example, in [60] and [38].*

Now we investigate the convergence properties of the suggested robust GRN structure identification algorithm. In fact, the suggested method is quite similar to the two-step recursions of the EKF-based method. More precisely, Eq. (10.71) can be divided into a time-update and measurement-update steps, which are respectively as follows.

- Time-update step: update of state predictions and pseudo-covariance matrix of prediction errors:

$$\hat{z}_{k+1|k} \triangleq g\left(\hat{z}_{k|k}\right) = \hat{A}\,(k,0)\,\hat{z}_{k|k} + a(k), \tag{10.80}$$

$$P_{k+1|k} = A\left(k, \hat{z}_{k|k}, 0\right) \hat{P}_{k|k}\, A^T\left(k, \hat{z}_{k|k}, 0\right) + \Psi(k). \tag{10.81}$$

- Measurement-update step: update of state estimation, pseudo-covariance matrix, and estimator gain:

$$\hat{z}_{k+1|k+1} = \hat{z}_{k+1|k} + K_{k+1|k+1}\left(y(k+1) - C(k+1)\hat{z}_{k+1|k}\right), \tag{10.82}$$

$$K_{k+1|k+1} = P_{k+1|k}\, C^T(k+1)\Big(C(k+1)\, P_{k+1|k}\, C^T(k+1) + R(k+1)\Big)^{-1}, \tag{10.83}$$

$$\begin{aligned} P_{k+1|k+1} = {} & P_{k+1|k} - P_{k+1|k}\, C^T(k+1)\Big(C(k+1)\, P_{k+1|k}\, C^T(k+1) + R(k+1)\Big)^{-1} \\ & \times C(k+1)\, P_{k+1|k}\,. \end{aligned} \tag{10.84}$$

These formulas are almost the same as their counterparts in the EKF-based estimation algorithm. The major difference is that in the EKF-based method, the difference between the values of the nonlinear function at the plant augmented state and its estimate is approximated as

$$g\,(z(k)) - g\left(\hat{z}_{k|k}\right) \approx A(k, \hat{z}_{k|k})\left(z(k) - \hat{z}_{k|k}\right),$$

whereas this difference in the RSE-based method is approximated as

$$g\,(z(k)) - g\left(\hat{z}_{k|k}\right) \approx \hat{A}(k)\,(0)\left(z(k) - \hat{z}_{k|k}\right).$$

Moreover, if a one-step formulation is reconstructed for the RSE-based method, we can establish a relation between the pseudo-covariance matrices of estimation errors on GRN topology at two successive time instants, which takes completely the same form as that of Eq. (10.78). As a result, the convergence properties of the RSE-based method can be expected to be close to those of the EKF-based method.

On the other hand, completely the same arguments as those in the derivation of Inequality (10.A.15) show that a similar inequality relationship can be established from Eq. (10.70) for the pseudo-covariance matrix $P_{\theta,k|k}$ and the pseudo-covariance matrix $P_{\theta,k+1|k+1}$ of the RSE-based GRN structure identification method, which is given in the following equation:

$$\begin{aligned}\lambda_{\min}\left(P^{-1}_{\theta,k+1|k+1}\right) \;\geq\;& \lambda_{\min}\left(P^{-1}_{\theta,k|k}\right)+\lambda^{2}_{\min}\left(P^{-1}_{\theta,k|k}\right)\lambda_{\min}\left(\Sigma_{k|k}\right)\\ &+\frac{1-\gamma(k)}{\gamma(k)}\lambda_{\min}\left(\left[S^{T}(k)S(k)\right]_{22}\right),\end{aligned} \tag{10.85}$$

where $\Sigma_{k|k}=P^{T}_{x\theta,k|k}\left(P^{T}_{x,k|k}+R(k)\right)^{-1}P_{x\theta,k|k}-\Phi(k)$. Moreover, similar to the partition of the matrix $P(k)$, $\left[S(k)^{T}S(k)\right]_{22}$ stands here for the 2nd block row 2nd block column submatrix of $S(k)^{T}S(k)$, which can be directly proved to be equal to $col\{\overline{SS}_{k,j}|_{j=1}^{n^2}\}col^{T}\{\overline{SS}_{k,j}|_{j=1}^{n^2}\}$. As the derivations of this inequality are completely the same as those of Inequality (10.85), we omit the details.

Note that $\gamma(k)\in[0,1]$ and the matrix $\left[S(k)^{T}S(k)\right]_{22}$ is at least positive semidefinite. Inequality (10.85) makes it clear that the convergence speed of the suggested robust GRN topology estimation method should not be slower than that of the EKF-based method.

Summarizing our discussion, we conclude by some convergence properties of the RSE-based method.

Corollary 10.1. *If the conditions in Theorem 10.5 hold, then the estimate $\hat{\theta}_{k|k}$ obtained by the suggested robust GRN structure identification algorithm converges to θ in mean square. Moreover, its convergence speed is not smaller than that of the EKF-based method.*

Remark 10.4. *Through expressing the RSE-based GRN structure identification algorithm into a recursive form similar to that of the EKF-based method, we obtain sufficient conditions for its convergence, which are completely the same as those of the EKF-based method. Nevertheless, the proof of this convergence property has also made it clear that the suggested method generally has a faster convergence speed. This property has a great significance in identifying a real GRN. In general, time series data obtained in an actual biological experiment is usually quite short. This means that an identification method with a rapid convergence speed is greatly appreciated in GRN structure estimations. By Corollary 1 it is safe to declare that the suggested method could be a competitive alternative for GRN topology identification in practice.*

To evaluate estimation performances of the suggested robust state estimation-based method for GRN structure identification algorithm, several comparisons are performed in [24], which include both simulated and actual gene expression data. False positive errors, false negative errors, and so on of the suggested RSE-based method are compared to those of the EKF-based method and those of the unscented Kalman filter (UKF) based method. Significant performance improvements have been achieved.

10.5 Bibliographic Notes

Studies on causality and/or structure estimation make it possible to describe not only data but also experiment design in a mathematical language. In [66], and [5], a diagraph is used to describe direct influences among different subsystems, whereas in [67], a framework is developed to estimate direct relations among different phenomena in which comparative experiments cannot be performed on the same individual. Techniques in time series analysis and Wiener filtering theory is utilized in [68] to establish causal relations among economy factors. A common feature in these studies is that stochastic analysis plays a central role. In [6], there are summarized several most important application areas of structure identification, characteristics of large-scale networked systems, and some major approaches developed for network estimations. In [4], there are given comprehensive comparisons of the characteristics of different methods in revealing the structure of a gene regulation network from experiment data. Many important topics about identification of the interactions in a gene regulation network have also been investigated in a special issue of Automatica [69].

Appendix 10.A

10.A.1 Proof of Theorem 10.4

When experimental data are available for its expression levels in which genes of a GRN are individually and systematically perturbed by external efforts, under the adopted assumptions, the likelihood function for the expression levels of gene i and their measurement error variance, denoted by

$$F_i\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right),$$

can be expressed as follows:

$$\begin{aligned}
& F_i\left(x_i^{[\mathrm{L},0]}, x_i^{[\mathrm{H},0]}, x_i^{[\mathrm{wt},0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right) \\
= \; & \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\left(x_i^{[\mathrm{wt}]} - x_i^{[\mathrm{wt},0]}\right)^2/(2\sigma_i^2)\right\} \prod_{j=1, j\neq i}^{n} \frac{1}{\sqrt{2\pi}\sigma_i}
\end{aligned}$$

$$\times \exp\left\{-\left(x_{ji}-x_{ji}^{[0]}\right)^2/(2\sigma_i^2)\right\}, \tag{10.A.1}$$

where $x_{ji}^{[0]} \in \left\{x_i^{[L,0]},\ x_i^{[H,0]},\ x_i^{[wt,0]}\right\}$.

Note that measurement errors for different genes and in different experiments are assumed to be independent of each other. Straightforward algebraic manipulations show that the likelihood function for all the gene expression levels and measurement error variances of the GRN, denoted by

$$F\left(\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right)\Big|_{i=1}^{n}\right),$$

can be expressed as

$$F\left(\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right)\Big|_{i=1}^{n}\right) = \prod_{i=1}^{n} F_i\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right) \tag{10.A.2}$$

This means that maximization of the likelihood function

$$F\left(\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right)\Big|_{i=1}^{n}\right)$$

is equivalent to independent maximization of the function

$$F_i\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right)$$

Sort the observed expression levels of gene i in a nondecreasing order and denote the sorted results by $x_{j_1,i} \le x_{j_2,i} \le \cdots \le x_i^{[wt]} \le \cdots \le x_{j_{n-1},i}$. Here, $j_l \neq j_k$ whenever $l \neq k$ and $j_l \in \{1, 2, \cdots, i-1, i+1, \cdots, n\}$. From the assumptions we know that there are respectively k_{i1} and k_{i2} genes that have direct activation and repression effects on gene i. On the other hand, direct algebraic manipulations show that

$$\begin{aligned} F_i\left(x_i^{[L,0]}, x_i^{[H,0]}, x_i^{[wt,0]}, \sigma_i \;\middle|\; k_{i1}, k_{i2}\right) &\le \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^n \prod_{l=1}^{k_{i1}} \exp\left\{-\left(x_{j_l,i}-x_i^{[L,0]}\right)^2/(2\sigma_i^2)\right\} \\ &\times \exp\left\{-\left(x_i^{[wt]}-x_i^{[wt,0]}\right)^2/(2\sigma_i^2)\right\} \prod_{l=k_{i1}+1}^{n-k_{i2}} \exp\left\{-\left(x_{j_l,i}-x_i^{[wt,0]}\right)^2/(2\sigma_i^2)\right\} \\ &\times \prod_{l=n-k_{i2}}^{n-1} \exp\left\{-\left(x_{j_l,i}-x_i^{[H,0]}\right)^2/(2\sigma_i^2)\right\}. \end{aligned} \tag{10.A.3}$$

For brevity, denote the minus of the natural logarithm of the right-hand side of the equation by f_{ki}. Direct algebraic operations show that f_{ki} can be equivalently expressed as

$$\begin{aligned} f_{ki} &= n\ln\sqrt{2\pi} + n\ln\sigma_i + \frac{1}{2\sigma_i^2}\left\{\sum_{l=1}^{k_1}\left(x_{j_l,i} - x_i^{[\mathrm{L},0]}\right)^2 + \left(x_i^{[\mathrm{wt}]} - x_i^{[\mathrm{wt},0]}\right)^2 + \right. \\ &\left. \sum_{l=1+k_1}^{n-k_2}\left(x_{j_l,i} - x_i^{[\mathrm{wt},0]}\right)^2 + \sum_{l=n-k_2+1}^{n}\left(x_{j_l,i} - x_i^{[\mathrm{H},0]}\right)^2\right\}. \end{aligned} \tag{10.A.4}$$

Differentiating the function f_{ki} with respect to σ_i, $x_i^{[\mathrm{wt},0]}$, $x_i^{[\mathrm{L},0]}$, and $x_i^{[\mathrm{H},0]}$, we have

$$\frac{\partial f_{ki}}{\partial x_i^{[\mathrm{L},0]}} = \frac{1}{\sigma_i^2}\sum_{l=1}^{k_1}\left(x_{j_l,i} - x_i^{[\mathrm{L},0]}\right), \tag{10.A.5}$$

$$\frac{\partial f_{ki}}{\partial x_i^{[\mathrm{wt},0]}} = \frac{1}{\sigma_i^2}\left\{\sum_{l=k_1+1}^{n-k_2}\left(x_{j_l,i} - x_i^{[\mathrm{wt},0]}\right) + \left(x_i^{[\mathrm{wt}]} - x_i^{[\mathrm{wt},0]}\right)\right\}, \tag{10.A.6}$$

$$\frac{\partial f_{ki}}{\partial x_i^{[\mathrm{H},0]}} = \frac{1}{\sigma_i^2}\sum_{l=n-k_2+1}^{n}\left(x_{j_l,i} - x_i^{[\mathrm{H},0]}\right), \tag{10.A.7}$$

$$\begin{aligned} \frac{\partial f_{ki}}{\partial \sigma_i} &= \frac{n}{\sigma_i} - \frac{1}{\sigma_i^3}\sum_{l=1}^{k_1}\left(x_{j_l,i} - x_i^{[\mathrm{L},0]}\right)^2 - \frac{1}{\sigma_i^3}\sum_{l=k_1+1}^{n-k_2}\left(x_{j_l,i} - x_i^{[\mathrm{wt},0]}\right)^2 \\ &\quad - \frac{1}{\sigma_i^3}\left(x_i^{[\mathrm{wt}]} - x_i^{[\mathrm{wt},0]}\right)^2 - \frac{1}{\sigma_i^3}\sum_{l=n-k_2+1}^{n}\left(x_{j_l,i} - x_i^{[\mathrm{H},0]}\right)^2. \end{aligned} \tag{10.A.8}$$

Then, from the first and second optimality conditions of the function f_{ki}, direct algebraic operations show that this function achieves its minimum at the following $\hat{x}_i^{[\mathrm{L},0]}$, $\hat{x}_i^{[\mathrm{H},0]}$, $\hat{x}_i^{[\mathrm{wt},0]}$, and $\hat{\sigma}_i$:

$$\hat{x}_i^{[\mathrm{L},0]} = \frac{\sum_{l=1}^{k_1} x_{j_l,i}}{k_1}, \qquad \hat{x}_i^{[\mathrm{wt},0]} = \frac{\sum_{l=1+k_1}^{n-k_2} x_{j_l,i} + x_i^{[\mathrm{wt}]}}{n+1-k_1-k_2}, \qquad \hat{x}_i^{[\mathrm{H},0]} = \frac{\sum_{l=n-k_2+1}^{n} x_{j_l,i}}{k_2} \tag{10.A.9}$$

$$\begin{aligned} \hat{\sigma}_i &= \frac{1}{\sqrt{n}}\left[\sum_{l=1}^{k_1}\left(x_{j_l,i} - \hat{x}_i^{[\mathrm{L},0]}\right)^2 + \sum_{l=1+k_1}^{n-k_2}\left(x_{j_l,i} - \hat{x}_i^{[\mathrm{wt},0]}\right)^2 + \left(x_i^{[\mathrm{wt}]} - \hat{x}_i^{[\mathrm{wt},0]}\right)^2\right. \\ &\quad \left. + \sum_{l=n-k_2+1}^{n}\left(x_{j_l,i} - \hat{x}_i^{[\mathrm{H},0]}\right)^2\right]^{\frac{1}{2}}. \end{aligned} \tag{10.A.10}$$

Note that the natural logarithm of a positive variable is an increasing function. Moreover, Inequality (10.A.3) becomes an equality with these $\hat{x}_i^{[\mathrm{L},0]}$, $\hat{x}_i^{[\mathrm{H},0]}$, $\hat{x}_i^{[\mathrm{wt},0]}$, and $\hat{\sigma}_i$. We can therefore declared that these values are MLEs for $x_i^{[\mathrm{L},0]}$, $x_i^{[\mathrm{H},0]}$, $x_i^{[\mathrm{wt},0]}$, and σ_i, respectively. This completes the proof. □

10.A.2 Proof of Theorem 10.5

The proof consists of the following two stages. First, we will show that the covariance matrix of the estimation error $\tilde{\theta}(k)$ converges to zero. Afterward, we will prove that the estimate $\hat{\theta}(k)$ converges to its actual value.

Denote the mean of the estimation error $\tilde{\theta}(k)$ by $\bar{\tilde{\theta}}(k)$. With this symbol, we can declare that to prove Theorem 10.5, the following property should be satisfied by the estimation error $\tilde{\theta}(k)$:

$$\begin{aligned}\lim_{k\to\infty}\mathbf{E}\left\{\left(\tilde{\theta}(k)-\bar{\tilde{\theta}}(k)\right)^T\left(\tilde{\theta}(k)-\bar{\tilde{\theta}}(k)\right)\right\} &= \lim_{k\to\infty}\mathbf{E}\left\{tr\left[\left(\tilde{\theta}(k)-\bar{\tilde{\theta}}(k)\right)\left(\tilde{\theta}(k)-\bar{\tilde{\theta}}(k)\right)^T\right]\right\}\\ &= \lim_{k\to\infty} tr\left(P_{\theta,k}\right)\\ &= 0, \end{aligned}\tag{10.A.11}$$

that is, it is necessary to show that $tr\left(P_{\theta,k}\right)\to 0$ as $k\to\infty$.

From Eq. (10.78) and the definition of $\Sigma(k)$ it is clear that $P_{\theta,k+1}=P_{\theta,k}-\Sigma(k)$. Based on this relation and Lemma 2.2, we can straightforwardly prove that

$$P_{\theta,k+1}^{-1}-P_{\theta,k}^{-1}=\left(P_{\theta,k}-\Sigma(k)\right)^{-1}\Sigma(k)P_{\theta,k}^{-1},\tag{10.A.12}$$

$$\left(P_{\theta,k}-\Sigma(k)\right)^{-1}-P_{\theta,k}^{-1}=\left(P_{\theta,k}\Sigma(k)^{-1}P_{\theta,k}-P_{\theta,k}\right)^{-1},\tag{10.A.13}$$

which further lead to

$$\begin{aligned}&\left(P_{\theta,k}-\Sigma(k)\right)^{-1}\Sigma(k)P_{\theta,k}^{-1}-P_{\theta,k}^{-1}\Sigma(k)P_{\theta,k}^{-1}\\ =\;&\left(P_{\theta,k}\Sigma(k)^{-1}P_{\theta,k}-P_{\theta,k}\right)^{-1}\left(P_{\theta,k}\Sigma(k)^{-1}\right)^{-1}\\ =\;&\left(P_{\theta,k}\Sigma(k)^{-1}P_{\theta,k}\Sigma(k)^{-1}P_{\theta,k}-P_{\theta,k}\Sigma(k)^{-1}P_{\theta,k}\right)^{-1}\\ =\;&P_{\theta,k}^{-1}\left[\Sigma(k)^{-1}\left(P_{\theta,k}-\Sigma(k)\right)\Sigma(k)^{-1}\right]^{-1}P_{\theta,k}^{-1}.\end{aligned}\tag{10.A.14}$$

Note that the constraint

$$P_{x\theta,k}^{T}\left(P_{x,k}+R(k)\right)^{-1}P_{x\theta,k}-P_{\theta,k}<\Phi(k)<P_{x\theta,k}^{T}\left(P_{x,k}+R(k)\right)^{-1}P_{x\theta,k}$$

is equivalent to $0<\Sigma(k)<P_{\theta,k}$. Therefore, Eq. (10.A.14) implies that

$$\left(P_{\theta,k}-\Sigma(k)\right)^{-1}\Sigma(k)P_{\theta,k}^{-1}>P_{\theta,k}^{-1}\Sigma(k)P_{\theta,k}^{-1}.$$

As a consequence, we can declare from Lemma 2.1 that Eq. (10.A.12) implies the following inequalities:

$$\begin{aligned}\lambda_{\min}\left(P_{\theta,k+1}^{-1}\right) &\geq \lambda_{\min}\left(P_{\theta,k}^{-1}\right)+\lambda_{\min}\left(\left(P_{\theta,k}-\Sigma(k)\right)^{-1}\Sigma(k)P_{\theta,k}^{-1}\right)\\ &\geq \lambda_{\min}\left(P_{\theta,k}^{-1}\right)+\lambda_{\min}\left(P_{\theta,k}^{-1}\Sigma(k)P_{\theta,k}^{-1}\right)\\ &\geq \lambda_{\min}\left(P_{\theta,k}^{-1}\right)+\lambda_{\min}^{2}\left(P_{\theta,k}^{-1}\right)\lambda_{\min}\left(\Sigma(k)\right).\end{aligned} \tag{10.A.15}$$

From condition (10.79a) it is clear that $P_{\theta,k+1}<P_{\theta,k}$, which implies that $P_{\theta,k+1}^{-1}>P_{\theta,k}^{-1}$. Therefore, when this assumption is satisfied, and if $\sum_{k=1}^{\infty}\left\{\lambda_{\min}^{2}\left(P_{\theta,k}^{-1}\right)\lambda_{\min}\left(\Sigma(k)\right)\right\}=\infty$, then $\lim_{k\to\infty}\lambda_{\min}\left(P_{\theta,k}^{-1}\right)=\infty$. Therefore, $tr\left(P_{\theta,k}\right)\to 0$ as $k\to\infty$.

Second, we will show that $\hat{\theta}(k)\to\theta$ as $k\to\infty$. Combining Eqs. (10.74)–(10.75), we have that

$$\hat{\theta}(k+1)=\hat{\theta}(k)+P_{x\theta,k}^{T}\left(P_{x,k}+R(k)\right)^{-1}e(k), \tag{10.A.16}$$

where $e(k)=y(k)-\hat{x}(k)$. Moreover, denoting $\theta-\hat{\theta}(k)$ by $\tilde{\beta}(k)$, we have that

$$\tilde{\beta}(k+1)=\tilde{\beta}(k)-P_{x\theta,k}^{T}\left(P_{x,k}+R(k)\right)^{-1}e(k). \tag{10.A.17}$$

Now, construct a Lyapunov function

$$V(k)=\tilde{\beta}^{T}(k)\left(P_{\theta,k}+I\right)\tilde{\beta}(k). \tag{10.A.18}$$

Then, the first difference of this Lyapunov function is

$$\begin{aligned}\Delta(k) &= V(k+1)-V(k)\\ &= \tilde{\beta}^{T}(k+1)\left(P_{\theta,k+1}+I\right)\tilde{\beta}(k+1)-\tilde{\beta}^{T}(k)\left(P_{\theta,k}+I\right)\tilde{\beta}(k)\\ &= \tilde{\beta}^{T}(k)\left(P_{\theta,k+1}-P_{\theta,k}\right)\tilde{\beta}(k)+\Omega(k)\\ &= -\tilde{\beta}^{T}(k)\Sigma(k)\tilde{\beta}(k)+\Omega(k),\end{aligned} \tag{10.A.19}$$

where

$$\begin{aligned}\Omega(k) = {} & \left(\tilde{\beta}^T(k+1)\tilde{\beta}(k+1) - \tilde{\beta}^T(k)\tilde{\beta}(k)\right) \\ & + \left[\left(\tilde{\beta}^T(k+1)P_{\theta,k+1}\tilde{\beta}(k+1) - \tilde{\beta}^T(k)P_{\theta,k}\tilde{\beta}(k)\right) - \tilde{\beta}^T(k)\left(P_{\theta,k+1} - P_{\theta,k}\right)\tilde{\beta}(k)\right].\end{aligned} \tag{10.A.20}$$

Note that from the definition of $P(k)$ it is clear that $P(k) \geq 0$. On the other hand, note that

$$P(k) = \begin{bmatrix} P_{x,k} & P_{x\theta,k} \\ P_{x\theta,k}^T & P_{\theta,k} \end{bmatrix}$$

and

$$\lim_{k\to\infty} P_{\theta,k} = 0.$$

We can directly prove that

$$\lim_{k\to\infty} P_{x\theta,k} = 0.$$

Therefore, for an arbitrary positive number ε, there exists a positive integer k_0 such that for $k > k_0$, the norm of the pseudo-covariance matrix $P_{x\theta,k}$ satisfies $\left\| P_{x\theta,k} \right\| < \varepsilon$. Moreover, from the results of Theorem 3.1 in [65] it is clear that when the constraints (10.79c) and (10.79d) are satisfied and the initial estimation error of the augmented state vector $z(k)$ is bounded, that is, $\left\| z_0 - \hat{z}_0 \right\| \leq \delta$, $e(k)$ is also exponentially bounded in the mean square sense. We can therefore declare that, for an arbitrary positive number $\bar{\varepsilon}$, there exists a positive integer $\bar{k}_0$ such that

$$\left\| \tilde{\beta}(k+1) - \tilde{\beta}(k) \right\| < \bar{\varepsilon}$$

for $k > \bar{k}_0$. This means that, for an arbitrary positive number $\bar{\tilde{\varepsilon}}$, there exists a positive integer $\bar{\tilde{k}}_0$ such that $|\Omega(k)| < \bar{\tilde{\varepsilon}}$ for $k > \bar{\tilde{k}}_0$. Hence, when k is large enough, it is certain that $|\Omega(k)| < \frac{1}{2}\tilde{\beta}(k)^T \Sigma(k)\tilde{\beta}(k)$, that is, there exists a positive integer $\tilde{k}_0$ such that, for $k > \tilde{k}_0$, $\Delta(k)$ defined in Eq. (10.A.19) has the following property:

$$\begin{aligned}\Delta(k) &= V(k+1) - V(k) \\ &\leq -\tilde{\beta}^T(k)\Sigma(k)\tilde{\beta}(k) + \frac{1}{2}\tilde{\beta}^T(k)\Sigma(k)\tilde{\beta}(k) \\ &= -\frac{1}{2}\tilde{\beta}^T(k)\Sigma(k)\tilde{\beta}(k) \\ &\leq 0.\end{aligned} \tag{10.A.21}$$

On the other hand, assume that $\tilde{\beta}^T(k)\Sigma(k)\tilde{\beta}(k) = 0$. Then, based on condition (10.79a), which implies that $\Sigma(k)$ is positive definite, we have that $\tilde{\beta}(k) = 0$. Therefore, by Lyapunov stability theory it is clear from the last inequality that $\hat{\theta}(k) \to \theta$ as $k \to \infty$.

We can therefore conclude that if condition (10.79) holds, then $\tilde{\theta}(k) \xrightarrow{m.s.} 0$.

This completes the proof. □

References

[1] H. Jong, Modeling and simulation of genetic regulatory systems: a literature review, Journal of Computational Biology 9 (2002) 67–103.

[2] T.S. Gardner, D. di Bernardo, D. Lorenz, J.J. Collins, Inferring genetic networks and identifying compound mode of action via expression profiling, Science 301 (2003) 102–105.

[3] B.N. Kholodenko, A. Kiyatkin, F.J. Bruggeman, E. Sontag, H.V. Westerhof, J.B. Hoek, Untangling the wires: a strategy to trace functional interactions in signaling and gene networks, Proceeding of the National Academy of Science, USA 99 (2002) 12841–12846.

[4] R.J. Prill, D. Marbach, J. Saez-Rodriguez, P.K. Sorger, L.G. Alexopoulos, X.W. Xue, N.D. Clarke, G. Altan-Bonnet, G. Stolovitzky, Towards a rigorous assessment of systems biology models: the DREAM3 challenges, PLoS ONE 5 (2010) e9202.

[5] J. Pearl, Causality: Models, Reasoning, and Inference, second edition, Cambridge University Press, UK, 2009.

[6] E.D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, 2009.

[7] T. Akutsu, S. Miyano, S. Kuhara, Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, in: Proceedings of the Pacific Symposium on Biocomputing, vol. 4, World Scientific Maui, Hawaii, USA, pp. 17–28.

[8] Y. Zheng, C.K. Kwoh, Reconstruction Boolean networks from noisy gene expression data, in: International Conference on Control, Automation, Robotics and Vision, vol. 4, pp. 58–72.

[9] I. Shmulevich, E.R. Dougherty, S. Kim, W. Zhang, Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks, Bioinformatics 18 (2002) 261–274.

[10] F. Ferrazzi, P. Sebastiani, M.F. Ramoni, R. Bellazzi, Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear gaussian networks, BMC Bioinformatics 8 (Suppl. 5) (2007) S2.

[11] B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, F. d'Alché-Buc, Gene networks inference using dynamic Bayesian networks, Bioinformatics 19 (2003) II138–II148.

[12] A. la Fuente, N. Bing, I. Hoeschele, P. Mendes, Discovery of meaningful associations in genomic data using partial correlation coefficients, Bioinformatics 20 (2004) 3565–3574.

[13] M. Bansal, G.D. Gatta, D. Bernardo, Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, Bioinformatics 22 (2006) 815–822.

[14] E.D. Sontag, Network reconstruction based on steady-state data, Essays in Biochemistry 45 (2008) 161–176.

[15] I. Cantone, L. Marucci, F. Iorio, M.A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. Bernardo, D. Bernardo, M.P. Cosma, A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches, Cell 137 (2009) 172–181.

[16] M. Andrec, B.N. Kholodenko, R.M. Levy, E. Sontag, Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy, Journal of Theoretical Biology 232 (2005) 427–441.

[17] P. Berman, B. Das Gupta, E. Sontag, Randomized approximation algorithms for set multicover problems with applications to reverse engineering of protein and gene networks, Discrete Applied Mathematics 155 (2007) 733–749.

[18] T. Zhou, Y.L. Wang, Causal relationship inference for a large-scale cellular network, Bioinformatics 26 (2010) 2020–2028.

[19] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Review 51 (2009) 661–703.

[20] D.D. Siljak, Large-scale Dynamic Systems: Stability and Structure, North-Holland Books, New York, USA, 1978.

[21] T. Zhou, Controllability and observability of networked dynamic systems, Automatica 52 (2015) 63–75.

[22] A.L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nature Reviews Genetics 5 (2004) 101–113.

[23] R. Chang, M. Stetter, W. Brauer, Quantitative inference by qualitative semantic knowledge mining with Bayesian model averaging, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 1587–1599.

[24] J. Xiong, T. Zhou, Structure identification for gene regulatory networks via linearization and robust state estimation, Automatica 50 (2014) 2765–2776.

[25] S. Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia, USA, 1991.

[26] G.H. Gloub, C. Loan, Matrix Computation, 2nd edition, The John Hopkins University Press, Baltimore, USA, 1989.

[27] T.S. Gardner, J.J. Faith, Reverse-engineering transcription control networks, Physics of Life Reviews 2 (2005) 65–88.

[28] Y.L. Wang, T. Zhou, A relative variation-based method to unraveling gene regulatory networks, PLoS ONE 7 (2012) e31194.

[29] N. Friedman, M. Linial, I. Nachman, D. Pe'er, Using Bayesian networks to analyze expression data, Journal of Computational Biology 7 (2000) 601–620.

[30] Y.L. Wang, Causal Relationship Identification and Multi-Target Analysis of a Gene Regulation Network, Ph.D. thesis, Department of Automation, Tsinghua University, China, 2011 (in Chinese).

[31] J. Xiong, Direct Causal Relationship Identification for a Gene Regulation Network, Ph.D. thesis, Department of Automation, Tsinghua University, China, 2014 (in Chinese).

[32] T. Zhou, J. Xiong, Y.L. Wang, GRN topology identification using likelihood maximization and relative expression level variations, in: Proceedings of the 31th Chinese Control Conference, Heifei, Anhui Province, China, pp. 7408–7417.

[33] G.L. Nemhauser, L.A. Wolsey, Integer and Combinatorial Optimization, John Wiley & Sons, 1988.

[34] L.A. Wolsey, Integer Programming, John Wiley & Sons, USA, 1988.

[35] R. Albert, Scale-free networks in cell biology, Journal of Cell Science 118 (2005).

[36] M. Andrecut, S.A. Kauffman, A.M. Madni, Evidence of scale-free topology in gene regulatory network of human tissues, International Journal of Modern Physics C 19 (2008).

[37] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[38] L. Ljung, System Identification: Theory for the User, Prentice Hall PTR, Upper Saddle River, New Jersey, USA, 1999.

[39] S.A. Kauffman, The Origins of Order: Self-organization and Selection in Evolution, Oxford University Press, USA, 1993.

[40] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences 95 (1998) 14863–14868.

[41] A.J. Butte, I.S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, in: Pacific Symposium on Biocomputing, vol. 5, pp. 418–429.

[42] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, T.S. Gardner, Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles, PLoS Biology 5 (2007) e8.

[43] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, A. Califano, Reverse engineering cellular networks, Nature Protocols 1 (2006) 662–671.

[44] P.E. Meyer, K. Kontos, F. Lafitte, G. Bontempi, Information-theoretic inference of large transcriptional regulatory networks, EURASIP Journal on Bioinformatics and Systems Biology 2007 (2007) 79879.

[45] A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods, PLoS ONE 5 (2010) e12776.

[46] J. Xiong, T. Zhou, Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses, PLoS ONE 7 (2012) e43819.

[47] S. Martin, Z. Zhang, A. Martino, J.L. Faulon, Boolean dynamics of genetic regulatory networks inferred from microarray time series data, Bioinformatics 23 (2007) 866–874.

[48] T.H. Tian, K. Burrage, Stochastic neural network models for gene regulatory networks, in: Evolutionary Computation, 2003. CEC'03. The 2003 Congress on, vol. 1, IEEE, 2003, pp. 162–169.

[49] T.F. Liu, W.K. Sung, A. Mittal, Model gene network by semi-fixed Bayesian network, Expert Systems with Applications 30 (2006) 42–49.

[50] K. Murphy, S. Mian, Modelling Gene Expression Data Using Dynamic Bayesian Networks, Technical report, Computer Science Division, University of California, Berkeley, California, USA, 1999.

[51] S.Y. Kim, S.Y. Imoto, S. Miyano, Inferring gene networks from time series microarray data using dynamic Bayesian networks, Briefings in Bioinformatics 4 (2003) 228–235.

[52] L.J. Qian, H.X. Wang, E.R. Dougherty, Inference of noisy nonlinear differential equation models for gene regulatory networks using genetic programming and Kalman filtering, IEEE Transactions on Signal Processing 56 (2008) 3327–3339.

[53] H. Wang, L. Qian, E. Dougherty, Inference of gene regulatory networks using S-system: a unified approach, IET Systems Biology 4 (2010) 145–156.

[54] Y.F. Huang, I. Tienda-Luna, Y.F. Wang, Reverse engineering gene regulatory networks, Signal Processing Magazine, IEEE 26 (2009) 76–97.

[55] Z.D. Wang, X.H. Liu, R.R. Liu, J.L. Liang, V. Vinciotti, An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 6 (2009) 410–419.

[56] A. Noor, E. Serpedin, M. Nounou, H. Nounou, Inferring gene regulatory networks via nonlinear state-space models and exploiting sparsity, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9 (2012) 1203–1211.

[57] T. Zhou, Sensitivity penalization based robust state estimation for uncertain linear systems, IEEE Transactions on Automatic Control 55 (2010) 1018–1024.

[58] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, G. Stolovitzky, Wisdom of crowds for robust gene network inference, Nature Methods 9 (2012) 796–804.

[59] X.D. Sun, L. Jin, M.M. Xiong, Extended Kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks, PLoS ONE 3 (2008) e3758.

[60] L. Ljung, T. Soderstrom, Theory and Practice of Recursive Identification, The MIT Press, Cambridge, Massachusetts, USA, 1987.

[61] A. Doucet, N. Freitas, N. Gordon, Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York, USA, 2001.

[62] G.R. Grimmett, D.R. Stirzaker, Probability and Random Processes, Oxford University Press, USA, 2001.

[63] L. Ljung, Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems, IEEE Transactions on Automatic Control 24 (1979) 36–50.

[64] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking, IEEE Transactions on Signal Processing 50 (2002) 174–188.

[65] K. Reif, S. Gunther, E. Yaz, R. Unbehauen, Stochastic stability of the discrete-time extended Kalman filter, IEEE Transactions on Automatic Control 44 (1999) 714–728.

[66] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, second edition, The MIT Press, London, UK, 2000.

[67] D.B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Education Psychology 66 (1974) 688–701.

[68] C.W. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (1969) 424–438.

[69] F. Allgower, F. Doyle (Eds.), Special Issue on Systems Biology, Automatica 47 (2011).

CHAPTER 11

Attack Identification and Prevention in Networked Systems

11.1 Introduction

A large-scale networked system usually consists of a great amount of subsystems, and these subsystems are often spatially distributed and far from each other geometrically. This asks for timely data exchanges among subsystems, which can be supported by advanced communication technologies, including both wired and wireless channels. Examples include electric power systems, unmanned vehicle systems, water supply networks, and so on.

Although integrations of several subsystems are greatly expected to increase performances of the whole system, they also provide more opportunities of being affected by malicious operators. There are many real world examples illustrating these possibilities, and in some cases, significant damages have been caused to the physical processes targeted by the attackers [1]. One example is the advanced computer worm Stuxnet, which happened in 2010 and infected some industrial control systems, reportedly leading to damages of approximately 1,000 centrifuges at these plants, although the attack itself was rather naive from a control engineer's point of view [2]. Another example is the Maroochy water breach in 2000 [3]. In this incident, an attacker managed to hack into some controllers that activate and deactivate some valves, which caused flooding of the grounds of a hotel, a park, and a river, with about a million liters of sewage. In 2003, the SQL Slammer worm attacked the Davis–Besse nuclear plant [4]. The recent multiple power blackouts in Brazil are also believed to be caused by malicious attacks [5].

Differently from faults and disturbances in a control system, which happen naturally and occasionally, attacks may happen in a well-organized way, simultaneously or sequentially over several spatial places of a large-scale networked system, with the objective to significantly violate performances of the system, or even to destroy the system itself. In other words, although faults also affect system behaviors, simultaneous events are usually not considered to be colluding. On the contrary, simultaneous attacks are often cooperative. In addition, faults are usually constrained by some physical dynamics and do not have an intent or objective, but attacks are usually not straightforwardly restricted by the dynamics of the plant physical process and/or chemical process, usually have a malicious objective, and often perform in a secrete way.

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00011-7

System safety is in fact not a new topic [6,7]. Particularly, in a power system, a supervisory control and data acquisition system, which is usually abbreviated as SCADA, is well developed and has been widely adopted. This system is originally designed to prevent electricity theft and to monitor abrupt voltage changes. In this system, measurement data of plant static outputs are used to estimate plant states, and the estimates are normally used in an internal control area of a transmission system operator. A basic assumption adopted in SCADA is that the system is in a steady-state behavior within its computation interval, which is usually of tens of seconds. When the plant states have a fast change during some alert or emergency situations, its estimation accuracy deteriorates drastically. Moreover, new power generation sources, such as large-scale wind power penetration, smart transmission devices, and so on have not been taken into account appropriately. In addition, inaccurate estimations about the state of a directly connected subsystem may cause some misleading estimates about system security and lead to a wrong decision of security control. These factors ask investigations on attack preventions using dynamic input–output data.

On the other hand, with the development of communication technologies and computer technologies, it is extensively anticipated that a plant and even subsystems of a plant are connected by some public communication networks, such as the world wide web and so on, to reduce hardware costs and increase maintenance flexibilities. In such a situation, a communication network becomes also an essential part of a networked system, such as that played by a physical plant and a digital feedback controller. When a system is connected by communication channels, many other possibilities arise for the system to be attacked. For example, it becomes easier for a hacker to insert some destructive disturbances in a more secrete way into a signal being transmitted in a communication channel [7–9]. In particular, a cyber-physical system usually suffers from some specific vulnerabilities, which do not exist in classical control systems and for which appropriate detection and identification techniques are required. For instance, the reliance on communication networks and standard communication protocols to transmit measurements and control packets increases the possibility of intentional and worst-case attacks against physical plants and/or the controller. On the other hand, information security methods, such as authentication, access control, message integrity, and so on, may not be adequate for a satisfactory protection of these attacks. Indeed, these security methods are only some passive ways to prevent a signal being disturbed during its transmissions. They do not exploit compatibilities of the received signals with the underlying physical/chemical process of the plant and/or the control mechanism. These characteristics may make them ineffective against an insider attack targeting the physical dynamics and/or the controller.

Generally, the dynamics of a networked system with possible attacks can be described as

$$x(k+1) = f(k, x(k), u(k), d(k)), \tag{11.1}$$

$$y(k) = g(k, x(k), u(k), d(k)), \tag{11.2}$$

where $x(k) \in \mathcal{R}^n$, $y(k) \in \mathcal{R}^m$, $u(k) \in \mathcal{R}^p$, and $d(k) \in \mathcal{R}^q$ respectively represent the plant state vector, output vector, input vector, and the vector consisting of possible attacks. Moreover, $f(k, \cdot)$ and $g(k, \cdot)$ are in general nonlinear functions that may be time varying. For a physical plant to work properly, there are usually some strict restrictions on its states. For example, in a water supply network, the levels of its reservoirs are in general restricted in some suitable ranges. The set consisting of these permissible plant states is called the plant safe region, whereas the complimentary set of the plant states is called the unsafe region. A plant is said to be safe if its state vector belongs to its safe region; otherwise, the plant is said to be unsafe. The objectives of the malicious attackers are often manipulating the plant state vector from its safe region to its unsafe region through adding an attack disturbance vector process $d(k)$ into the system with the smallest efforts, as well as to avoid being detected. In other words, to manipulate the plant from being safe to being unsafe in a theft and economic way. The input signal $u(k)$ is delivered from a controller through a communication network designed to make the plant work satisfactory. Due to the imperfectness of communication channels, this delivery may have some time delays and even failures and be corrupted by environment noises. Without detectors, the plant is usually impossible to distinguish the input vector $u(k)$ from that carrying the attack vector $d(k)$. In other words, the attack vector $d(k)$ is usually injected into the input vector $u(k)$ possibly in a transmission process, and this injection combines these two vectors into a single vector that is input to the plant. Differently from external disturbances widely studied in control system analysis and synthesis, the attack vector $d(k)$ is usually not random. More precisely, it might be designed with a clear objective using some measurements of the plant output vector $y(k)$.

When $d(k) \equiv 0$, the networked system described by Eqs. (11.1) and (11.2) is said to be in its normal situation, and the associated trajectories of its state and output vectors are said to be in their nominal behavior. Otherwise, the system is said to be in an abnormal situation, and the corresponding trajectories of its state and output vectors are said to be in their abnormal behavior. The objective of attack estimation is to clarify whether or not the system is in its normal situation using measurements of its output vector, whereas the objective of attack identification is to estimate the amount and positions of attackers when the system is in an abnormal situation. Generally, these two problems are dealt with through investigating characteristics of the residues of some detectors.

Taking into account that for nonlinear dynamic systems, there still does not exist a general analysis and/or synthesis method, our attention is restricted to linear systems in this chapter, in which both the function $f(k, \cdot)$ of Eq. (11.1) and the function $g(k, \cdot)$ of Eq. (11.2) are linear. This significantly reduces mathematical difficulties in handling the attack estimation and prevention problems, keeping the associated results significant and relevant from an engineering viewpoint. With this simplification, the input vector $u(k)$ can be removed from the model, as its existence does not affect conclusions on either attack identification or attack prevention due to the linearities of the system.

In this chapter, we first discuss attack identification and prevention using static data, which has been extensively adopted in large-scale electric power networks. Afterward, relations between system observability and attack prevention are investigated. With these relations, estimation and identification of attacks are respectively dealt with, as well as the relations between system security and optimal sensor placements. Further research topics will also be briefly discussed.

To clarify basic concepts and ideas, rather than the general model of a networked system given by Eqs. (3.25) and (3.26), which has been discussed in several chapters of this book, in this chapter the state-space model of Eq. (3.1) is utilized.

11.2 The SCADA System

Very possibly, the most successful attack estimator until now is the SCADA system, which is originally designed to protect an electricity network from working in an unsafe region. In this system, the plant is assumed to be in steady state, and its state vector is estimated from its output measurements using a least squares method. Clearly, to let this system work properly, it is necessary that the number of sensors in the plant is not smaller than the dimension of its state vector. Otherwise, the output of the associated attack detector cannot be uniquely determined, which will invalidate its usefulness. Generally, the more the sensors the plant has, the better the detection performance the system has.

When the plant is linear and works in its steady state, and only output measurements are utilized in estimating its state vector, the relations between the plant outputs and states, described by Eq. (11.2), can be expressed in a much simpler way given as follows:

$$y(k) = C_k x(k) + D_k d(k) + w(k). \tag{11.3}$$

In this equation, a vector $w(k)$ is introduced to represent measurement noises. For simplicity, we assume that its mathematical expectation is constantly equal to zero, whereas its covariance matrix is constantly equal to the identity matrix. As pointed out before, to get a physically meaningful estimate about the plant states, it is necessary that the dimension of the output vector $y(k)$ is not smaller than that of the state vector $x(k)$. This means that the output matrix C_k has more rows than columns.

When no attacks are available, a reasonable estimate about the plant state vector, denoted $\hat{x}(k)$, which is extensively adopted and widely called a least squares estimate, is

$$\hat{x}(k) = (C_k^T C_k)^{-1} C_k^T y(k). \tag{11.4}$$

With this estimate, the predicted plant output $\hat{y}(k)$ and its prediction errors $\tilde{y}(k)$ are respectively

$$\hat{y}(k) = C_k(C_k^T C_k)^{-1} C_k^T y(k), \tag{11.5}$$

$$\tilde{y}(k) = y(k) - \hat{y}(k) = \left[I - C_k(C_k^T C_k)^{-1} C_k^T\right] y(k). \tag{11.6}$$

Obviously, to make this estimate physically meaningful, it is necessary that the matrix C_k is of full column rank, which guarantees the existence of the inverse of $C_k^T C_k$. This condition can be satisfied through an appropriate sensor placement in the plant.

When the measurement error vector $w(k)$ has a normal distribution, we can easily prove that $\tilde{y}^T(k)\tilde{y}(k)$ has a $\mathcal{X}^2$ distribution. Hence, an anomaly detector, which is based on a quantity of the residue of plant output measurement, can be defined as

$$r(k) = Sy(k), \qquad S = I - C_k(C_k^T C_k)^{-1} C_k^T, \tag{11.7}$$

which is in fact the SCADA system [6]. Obviously, this detector shares the same form of the output prediction error in the normal situation that is given by $\tilde{y}(k)$. In this system, $\sqrt{r^T(k)r(k)}$ is utilized to decide whether or not there exists an attack in the networked system. When its value is greater than a prescribed threshold, it is believed that an attack exists. Otherwise, there are no attacks in the networked system [6,7].

This anomaly detector is usually efficient in detecting a single attack which caused a significant change in one of the plant output measurements. However, when a coordinated malicious attack exists in the plant and may cause simultaneous change of several plant output measurements, very possibly, this detector does not work quite well. For example, assume that there is a coordinated attack $\bar{d}(k)$ satisfying

$$D_k\bar{d}(k) = C_k\bar{x}(k). \tag{11.8}$$

This is possible if the attackers appropriately select their attack places and strategies and coordinate their actions. Under such a situation, we have that the output of the anomaly detector satisfies the following equalities:

$$\begin{aligned}
r(k) &= Sy(k) \\
&= \left[I - C_k(C_k^T C_k)^{-1} C_k^T\right] \times \left[C_k x(k) + D_k\bar{d}(k) + w(k)\right] \\
&= \left[I - C_k(C_k^T C_k)^{-1} C_k^T\right] \times [C_k x(k) + w(k)] + \left[I - C_k(C_k^T C_k)^{-1} C_k^T\right] C_k\bar{x}(k) \\
&= \left[I - C_k(C_k^T C_k)^{-1} C_k^T\right] \times [C_k x(k) + w(k)],
\end{aligned} \tag{11.9}$$

that is, no matter how large the magnitude of an element in the vector $\bar{x}(k)$ is, the output of the anomaly detector, that is, $r(k)$, does not change. This invalidates its capabilities of detecting the attacks satisfying Eq. (11.8).

Note that when the plant is linear and the input signal $u(t)$ is constantly set to zero, by Eq. (11.1) its state evolutions can be expressed as

$$x(k+1) = A_k x(k) + B_k d(k). \tag{11.10}$$

In the plant steady state, we have that $x(k+1) = x(k)$. Hence,

$$x(k) = [I - A_k]^{-1} B_k d(k), \tag{11.11}$$

which may be very far from the steady state of the plant without an attack, which is equal to zero by the adopted assumption that $u(k) \equiv 0$. In fact, when the attack of Eq. (11.8) is applied,

$$x(k) = [I - A_k]^{-1} B_k D_k^{\dagger} C_k \bar{x}(k), \tag{11.12}$$

where $D_k^{\dagger}$ stands for the pseudo-inverse of the matrix $D(k)$.[1] As each element of the vector $\bar{x}(k)$ can be arbitrarily large in magnitude without being detected by the anomaly detector, this equation implies that an appropriate selection of either the matrix B_k or the matrix D_k, or both of them, may cause a significant change of the plant states, which are capable of deriving the plant to its unsafe region. On the other hand, through choices of attack positions in a networked system, a selection of either the matrix B_k or the matrix D_k can be realized in a relatively easy way, provided that the attackers have accurate information on the system model, which is mostly about the system output matrix C_k in this case.

These conclusions are valid even if the plant is not in its steady state. In addition, numerical studies show that the attack in the form (11.8), which is extensively called stealthy attack, is usually sparse [6,7].

To analyze the security of a networked system, some security indices have also been introduced. When the plant is also time invariant, the following index α_j is suggested in [7] to measure the security degree of the plant jth output measurement:

$$\alpha_j \overset{\text{def}}{=} \min_{x \in \mathcal{R}^n} ||Cx||_0 \quad \text{subject to } C(j,:)x \neq 0, \tag{11.13}$$

[1] When $C_k \bar{x}(k)$ belongs to the space spanned by the column vectors of the matrix D_k, there may exist infinitely many vectors $\bar{d}(k)$ that satisfy Eq. (11.8). A complete parameterization for all the solutions to this equation is available, which is given by Theorem 2.5. As this is not a very essential issue here, we omit the details and only use $D_k^{\dagger} C_k \bar{x}(k)$ in the discussions.

where to make the expressions more concise, the temporal variable k has been omitted from the system matrices. Moreover, $C(j,:)$ stands for the jth row vector of the output matrix C.

From its definition it is clear that when an attacker is intended to change the plant jth output measurement without being detected by the SCADA system, α_j is in fact the minimum number of plant output measurements that he/she needs compromise. The smaller this index, the easier the jth plant output measurement to be attacked by a stealthy attack, that is, an attack that cannot be detected in principle. This means that knowledge about this security index for all the plant output measurements may provide some useful information about the security vulnerabilities of the networked system, which may be helpful in protecting the networked system with restricted resources.

Although the security index is significant from an engineering viewpoint, the associated optimization problem itself cannot be easily solved. In fact, this minimization problem has been proven to be NP-hard, and various modifications have been suggested as an alternative. One example is to replace the vector 0-norm with the vector 1-norm.

11.3 Attack Prevention and System Transmission Zeros

The previous section illustrates briefly the SCADA system, which uses only plant output measurements to detect whether or not there are attacks in the plant. The attack detection method is quite simple, but it reveals almost all essential issues in attack preventions: an attacker usually intends to significantly violate or even destroy a system in a secrete way, whereas a detector should be efficient in revealing all the attacks as soon as possible. This implies that there may exist close relations between attack preventions and system observability. In this section, we clarify these relations under the situation in which both the detector and the attacker have accurate knowledge about the plant dynamics.

As general conclusions are still not available for nonlinear time-varying systems, our attention is also restricted to linear and time-invariant systems in this section. We assume that an attacker is intended to alter the states of a plant through injecting some exogenous inputs on the basis of the system matrices of the plant at each time. In addition, the attacker is assumed to have unrestricted computation capabilities. An attacker with these characteristics is called an omniscient attacker. Under these assumptions, stealthy colluding attacks are closely related to plant transmission zeros.

Assume that the dynamics of a networked system with possible attacks is described as

$$x(k+1) = Ax(k) + Bd(k), \tag{11.14}$$

$$y(k) = Cx(k) + Dd(k). \tag{11.15}$$

As in the previous section, here $d(k)$ represents an attack disturbance. Once again, we assume that $x(k) \in \mathcal{R}^n$, $y(k) \in \mathcal{R}^m$, and $d(k) \in \mathcal{R}^q$. Noting that the plant is assumed to be linear, it is not necessary to include control signals in the adopted model.

To simplify discussions, we assume throughout the rest of this chapter that the above state space model is a minimal realization, that is, it is both controllable and observable. Note that if the system is not controllable, then a decomposition can divide its state space into the direct sum of controllable and uncontrollable subspaces, and the system state vector can be manipulated by an attacker only when it is in the controllable subspace. On the other hand, if the system is unobservable, then the system state space can be decomposed into the direct sum of observable and unobservable subspaces. As the part of a state vector belonging to the unobservable subspace does not have any influence on the plant output vector, variations of this part cannot be detected by the plant output vectors. That is, under such a situation, there always exist attacks that cannot be detected and therefore cannot be identified by monitoring only these plant outputs. This observation means that the aforementioned minimal realization assumption is reasonable.

Another assumption adopted in the rest of this chapter is that the matrix $\begin{bmatrix} B \\ D \end{bmatrix}$ is of full column rank, which is constituted from the system input matrix B and its direct feedthrough matrix D. This is necessary for the identifiability of attacks, noting that if this matrix is not of full column rank, then different attacks may cause the same value of the vector $\begin{bmatrix} B \\ D \end{bmatrix} d(k)$.

More precisely, for an arbitrary $\bar{d}(k) \in \mathcal{N}ull\left(\begin{bmatrix} B \\ D \end{bmatrix}\right)$, it is obvious from the definition of the null space of a matrix that

$$\begin{bmatrix} B \\ D \end{bmatrix} [d(k) + \bar{d}(k)] = \begin{bmatrix} B \\ D \end{bmatrix} d(k).$$

In addition, when the matrix $\begin{bmatrix} B \\ D \end{bmatrix}$ is not of full column rank, $\mathcal{N}ull\left(\begin{bmatrix} B \\ D \end{bmatrix}\right)$ has at least one nonzero vector as its element. Under such a situation, it is clear from Eqs. (11.14) and (11.15) that these different attacks lead to the same trajectory of the system state vector and to the same trajectory of the system output vector. Obviously, this is not an appreciative property in attack detection and/or attack identification.

Similarly to those in the SCADA system, an attack detector, which is sometimes also called an attack monitor, exploits only plant dynamics and output measurements to reveal the existence of attacks and identify their positions. Hence, an attack cannot be detected in principle if the associated plant output measurements are consistent with those without any attacks. On the contrary, if an attack causes a plant output measurement that has some characteristics different from those in regular/normal situations, some methods may be developed to detect this attack. Note that the output of a plant depends on both its initial conditions and external inputs. To clarify this dependence, rather than $y(k)$, $y(k, x(0), d(j)|_{j=0}^{k})$ is used more often in the following discussions. Here, $x(0)$ belongs to the set $\mathcal{R}^n$ and represents the value of the plant state vector at the time instant $k = 0$.

Similarly, $x(k, x(0), d(j)|_{j=0}^{k})$ is sometimes adopted to indicate the dependence of the state vector on its initial values and on the input vector sequence $d(k)|_{k=0}^{\infty}$.

Definition 11.1. *(Undetectable attack) Concerning the system described by Eqs. (11.14) and (11.15), an attack $d(k)|_{k=0}^{\infty}$ is undetectable if for each initial state vector $x(0) \in \mathcal{R}^n$, there exists at least one other initial state vector $\bar{x}(0) \in \mathcal{R}^n$ such that $y(k, \bar{x}(0), d(j)|_{j=0}^{k}) = y(k, x(0), 0)$ at each $k = 0, 1, 2, \cdots$.*

From an engineering point of view, this definition means that a detector cannot be constructed if the plant outputs associated with some attacks are completely the same as those due to modifications of plant initial conditions. One of the basic motivations behind this definition is that plant initial conditions are usually not exactly known in many applications, which makes the mapping from a plant input series to a plant output series not bijective and therefore leaves opportunities for an attacker to inject destructive disturbances.

In addition, to detect whether or not there exists an attack in a networked system, it is usually necessary for a detector to determine where the attack is from if it exists. This leads to an attack identification problem and requires to construct a detector that has capabilities of identifying attack locations. As attacks in a networked system are usually sparse, only a few elements of the attack vector $d(k)$ are usually not constantly equal to zero. Hence, some upper bounds can be put on the number of colluding attackers. Here, we assume that at most K attackers collude in one attack.

Definition 11.2. *(Unidentifiable attack) Concerning the system described by Eqs. (11.14) and (11.15), an attack vector sequence $d(k)|_{k=0}^{\infty}$ with K elements not constantly zero is unidentifiable if for each initial state vector $x(0) \in \mathcal{R}^n$, there exist at least one other initial state vector $\bar{x}(0) \in \mathcal{R}^n$ and one other attack vector sequence $\bar{d}(k)|_{k=0}^{\infty}$ with $\bar{K}$ elements not constantly zero, where $0 \leq \bar{K} \leq K$, such that at every time instant $k = 0, 1, 2, \cdots$, the plant output satisfies $y(k, \bar{x}(0), \bar{d}(j)|_{j=0}^{k}) = y(k, x(0), d(j)|_{j=0}^{k})$.*

Similarly to an undetectable attack, an unidentifiable attack cannot be identified because another attack can generate completely the same plant output, which makes plant outputs alone not sufficiently informative in distinguishing these attacks.

Let $\mathcal{K}$ denote a set consisting of K nonzero positive integers that belong to the set $\{1,\ 2,\ \cdots,\ q\}$, where q stands for the dimension of the attack vector $d(k)$. Associating with this set, define the attack set $\mathcal{A}_K$ in which an entry of the attack vector $d(k)$ is not constantly set to zero only if it is in the row whose number is one of the elements of the set $\mathcal{K}$. More precisely, assume that $\mathcal{K} = \{\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_K\}$ with $\alpha_i \in \{1,\ 2,\ \cdots,\ q\}$ for each $i = 1, 2, \cdots, q$. Then,

$$\mathcal{A}_{\mathcal{K}} = \{\, d(k)\ |\ d(i,k) \equiv 0,\ \ \forall i \in \{1,\ 2,\ \cdots,\ q\} \backslash \mathcal{K};\ \ d(i,k) \not\equiv 0,\ \ \forall i \in \mathcal{K} \,\},$$

where $d(i,k)$ stands for the ith row element of the attack vector $d(k)$.

An attack set is undetectable if there exists at least one undetectable attack in that set. Similarly, an attack set is unidentifiable if there exists at least one unidentifiable attack in that set. As the attack set $\mathcal{A}_K$ only provides information about the position, in which an attack is injected into the networked system, which is completely the same as that of the set $\mathcal{K}$, sometimes the set $\mathcal{K}$ is also called as an attack set. In other words, when an attack from the attack set $\mathcal{A}_K$ is injected into the networked system described by Eqs. (11.14) and (11.15), $Bd(k)$ and $Dd(k)$ in these equations can be respectively rewritten as $B_{\mathcal{K}} d_{\mathcal{K}}(k)$ and $D_{\mathcal{K}} d_{\mathcal{K}}(k)$, where $B_{\mathcal{K}}$ is constituted from the columns of the matrix B whose numbers are consistent with the elements of the set $\mathcal{K}$, whereas $D_{\mathcal{K}}$ is constituted from the columns of the matrix D whose numbers are consistent with the elements of the set $\mathcal{K}$. In addition, $d_{\mathcal{K}}(k)$ is a K-dimensional real-valued vector.

In [9], a different attack detection problem is formulated and investigated. It is assumed there that only sensors or actuators of a networked system are attacked, and a characterization is given for the maximum number of attacks that can be detected and corrected, which is expressed as a function of the plant state transition matrix and the plant output matrix. Although this problem formulation appears to be different from that given in Definition 11.1, they are actually closely related to each other [7].

From linearity assumption of the networked system and Definition 11.1 it is obvious that an attack $d(k)|_{k=0}^{\infty}$ is undetectable if and only if there exists a plant initial state vector $\tilde{x}(0) \in \mathcal{R}^n$ such that

$$y(k, \tilde{x}(0), d(j)|_{j=0}^{k}) \equiv 0. \tag{11.16}$$

The latter is closely related to some fundamental properties of a dynamic system. In fact, Eq. (11.16) has completely the same form as that of the conditions adopted in the definition of the zero dynamics of a system. More precisely, under some weak conditions, the equality in this equation can be satisfied if and only if the attack $d(k)|_{k=0}^{\infty}$ only stimulates the zero dynamics of the networked system [10,11]. Due to these relations, we can establish an algebraic criterion to verify whether or not an attack set is undetectable.

Similarly, we can straightforwardly prove that an attack $d(k)|_{k=0}^{\infty}$ with K elements that are not constantly zero is unidentifiable if and only if there exist a plant initial state vector $\tilde{x}(0) \in \mathcal{R}^n$ and an attack $\bar{d}(k)|_{k=0}^{\infty}$ with $\bar{K}$ elements in $d(k)|_{k=0}^{\infty}$ that is not constantly zero and the integer $\bar{K}$ satisfying $0 \leq \bar{K} \leq K$ such that

$$y(k, \tilde{x}(0), d(j) - \bar{d}(j)|_{j=0}^{k}) \equiv 0. \tag{11.17}$$

Note that Eqs. (11.16) and (11.17) are quite similar in their forms. It is not hard to understand that attack identification is once again closely related to transmission zeros of a networked system, which leads to an easily verifiable algebraic condition for an unidentifiable attack set.

11.3.1 Zero Dynamics and Transmission Zeros

These discussions reveal that both attack detection and attack identification are in fact a verification problem on the existence of an initial state vector and an input sequence that make the plant output vector be constantly equal to zero, which is closely related to zero dynamics and therefore to transmission zeros of a system. To illustrate these relations, the following conclusions are first established for a discrete-time system, which are similar to the results about continuous-time systems given in [10,11].

Lemma 11.1. *Let $G(z)$ be the transfer function matrix with minimal realization given by Eqs. (11.14) and (11.15). If the system input vector process $d(k)$ is of the form $d(k) = \lambda^k d(0)$ for each $k \geq 0$ and the system initial conditions satisfy $x(0) = (\lambda I - A)^{-1} Bd(0)$, in which λ is a constant scalar not equal to any eigenvalue of the matrix A, and $d(0)$ is a constant vector with compatible dimension. Then for an arbitrary integer $k \geq 0$, the system output vector process $y(k)$ can be expressed as*

$$y(k) = \lambda^k G(\lambda) d(0). \tag{11.18}$$

Proof. When λ is not an eigenvalue of the matrix A, the matrix $\lambda I - A$ is invertible, which means that the condition $x(0) = (\lambda I - A)^{-1} Bd(0)$ is well defined. Note that when a system has a minimal realization of Eqs. (11.14) and (11.15), the $\mathcal{Z}$-transformation of its output vector process $y(k)$ can be expressed as

$$y(z) = C(zI - A)^{-1} [x(0) + Bd(z)] + Dd(z). \tag{11.19}$$

Hence, when $d(k) = \lambda^k d(0)$ for each $k \geq 0$, we have that

$$\begin{aligned}
y(z) &= C(zI-A)^{-1}\left\{x(0)+B\left[(z-\lambda)^{-1}d(0)\right]\right\}+D\left[(z-\lambda)^{-1}d(0)\right] \\
&= \left[C(\lambda I-A)^{-1}B+D\right]\left[(z-\lambda)^{-1}d(0)\right] \\
&\quad +C(zI-A)^{-1}\left\{x(0)+B\left[(z-\lambda)^{-1}d(0)\right]\right\} \\
&\quad -C(\lambda I-A)^{-1}B\left[(z-\lambda)^{-1}d(0)\right] \\
&= G(\lambda)\left[(z-\lambda)^{-1}d(0)\right] \\
&\quad +C(zI-A)^{-1}\left\{x(0)+\left[I-(zI-A)(\lambda I-A)^{-1}\right]B\left[(z-\lambda)^{-1}d(0)\right]\right\} \\
&= G(\lambda)\left[(z-\lambda)^{-1}d(0)\right]+C(zI-A)^{-1}\left\{x(0)-(\lambda I-A)^{-1}Bd(0)\right\}. \qquad (11.20)
\end{aligned}$$

Therefore, the condition $x(0)=(\lambda I-A)^{-1}Bd(0)$ leads to the following equality:

$$y(z)=G(\lambda)\left[(z-\lambda)^{-1}d(0)\right] \qquad (11.21)$$

Recall that λ is a constant scalar. The proof can now be completed through taking the inverse $\mathcal{Z}$-transform of the vector-valued function $y(z)$. □

From Lemma 11.1 it is clear that if $d(0)$ belongs to the null space of $G(\lambda)$, then when the system initial state vector is set as $x(0)=(\lambda I-A)^{-1}Bd(0)$, the plant output vector $y(k)$ is constantly equal to zero for each $k=0,1,2,\cdots$. However, it is worth mentioning that when λ is a real scalar and $d(0)$ is a real-valued vector, both $\lambda^k d(0)$ with any $k\geq 0$ and $(\lambda I-A)^{-1}Bd(0)$ are real-valued vectors, which can be realized in principle in actual engineering problems, noting that both the state transition matrix A and the system input matrix B are real valued. However, when either λ is a complex scalar or $d(0)$ is a complex-valued vector, the associated $\lambda^k d(0)$ with $k\geq 0$ and/or $(\lambda I-A)^{-1}Bd(0)$ may also be complex valued. This may make the associated values not be realizable by any actual system input vector process and/or any system initial states.

When λ is equal to an eigenvalue of the state transition matrix A, the discussions become more complicated. However, with the results of Theorem 2.5, similar results can still be obtained. On the other hand, using a concept called system matrix, similar results can also be established without distinguishing whether or not λ is an eigenvalue of the state transition matrix A, which are given in the following Lemma 11.2.

To guarantee the existence of a nontrivial input vector process $d(k)$ and a nontrivial initial state vector $x(0)$ such that all the conditions of Lemma 11.1 are satisfied and the plant output vector process is constantly equal to zero, it is necessary that the matrix $G(\lambda)$ is not of full column rank. When the system is of single input and single output, it is obvious that λ must be one of the zeros of its transfer function, as the minimal realization assumption does not permit the existence of a zero in the transfer function that is equal to one of its poles. The situation becomes much more complicated when the system is of multiple inputs and multiple outputs. In this case, even if its state space model is a minimal realization, there still exists possibility that some of its zeros and some of its poles share the same value. To clarify the mathematical descriptions of a system zero and its relations to the zero dynamics of the system, the following concepts are required.

Definition 11.3. *A vector $x(0) \in \mathcal{R}^n$ is called weakly unobservable if there exists an input vector sequence $d(k)|_{k=0}^{\infty}$, such that the corresponding output vector sequence $y(k)$ is constantly equal to zero, that is, $y\left(k, x(0), d(j)|_{j=0}^{k-1}\right) = 0$ for each $k = 0, 1, 2, \cdots$.*

Denote by $\mathcal{V}$ the set consisting of all weakly unobservable vectors of a system. From the linearity of the system we can straightforwardly proven that this set is actually a subspace of $\mathcal{R}^n$. Due to this reason, the set $\mathcal{V}$ is sometimes called the weakly unobservable subspace of the system [11,12]. In fact, if we define a subspace $\mathcal{V}_k$ with $k = 0, 1, 2, \cdots$, as

$$\mathcal{V}_k = \left\{x(0) \;\middle|\; x(0) \in \mathcal{R}^n, \text{ there is an input sequence } d(k)|_{k=0}^{\infty}, \right.\\ \left. \text{such that } y(i) = 0 \text{ for each } i = 0, 1, \cdots, k-1\right\},$$

then it is obvious from the definition of this subspace sequence that

$$\mathcal{V}_0 \supseteq \mathcal{V}_1 \supseteq \mathcal{V}_2 \supseteq \mathcal{V}_3 \supseteq \cdots.$$

In addition, we can directly prove that there exists a nonnegative integer $0 \le k \le n$ such that

$$\mathcal{V} = \mathcal{V}_k = \mathcal{V}_{k+i} \qquad \text{for all } i \ge 0.$$

Let $F \in \mathcal{R}^{q\times n}$ and $L \in \mathcal{R}^{q\times s}$ satisfy

$$(A+BF)\mathcal{V} \subseteq \mathcal{V}, \qquad \mathcal{N}ull(C+DF) \supseteq \mathcal{V}, \qquad \mathcal{S}pan(L) = \mathcal{N}ull(D)\bigcap(B^{\dagger}\mathcal{V}).$$

The existence of these matrices is shown in [11]. Then we can prove that, for each $x(0) \in \mathcal{V}$, an input vector sequence $d(k)|_{k=0}^{\infty}$ that satisfies

$$y\left(k, x(0), d(j)|_{j=0}^{k-1}\right) = 0 \quad \text{for each} \quad k = 0, 1, 2, \ldots,$$

can be parameterized so that, for an arbitrary $k \in \{0, 1, 2, \cdots\}$,

$$d(k) = Fx(k) + Lw(k),$$

where $w(k)$ is an arbitrary real vector sequence with compatible dimension. Moreover, we can also prove that if $x(0) \in \mathcal{V}$ and $d(i)|_{i=0}^{\infty}$ with $d(i) \in \mathcal{R}^q$ is an associated input vector sequence such that $y(k, x(0), d(i)|_{i=0}^{k-1}) = 0$ for each $k = 1, 2, \cdots$, then the associated state vector sequence $x(k)|_{k=0}^{\infty}$, denoted $x(k, x(0), d(i)|_{i=0}^{k-1})|_{k=0}^{\infty}$, satisfies $x(k, x(0), d(i)|_{i=0}^{k-1}) \in \mathcal{V}$ for each $k = 1, 2, \cdots$.

Some other properties of this weakly unobservable subspace can be found, for example, in [11,12].

The concept of zero dynamics of a system is established on the basis of its weakly unobservable subspace.

Definition 11.4. *When the initial state vector of the system described by Eqs. (11.14) and (11.15) is restricted to belong to its weakly unobservable subspace $\mathcal{V}$, its associated input–output relation is called its zero dynamics.*

The zero dynamics of a system is also connected to its weakly unobservable subspace through its system matrix, a concept suggested originally by Rosenbrock [11,13].

Lemma 11.2. *Concerning the system described by Eqs. (11.14) and (11.15), assume that the matrix $\begin{bmatrix} B \\ D \end{bmatrix}$ is of full column rank. Define its system matrix*

$$P(z) = \begin{bmatrix} zI - A & -B \\ C & D \end{bmatrix}. \tag{11.22}$$

Suppose that there exist a vector $x(0) \in \mathcal{R}^n$ and a vector $d(0) \in \mathcal{R}^q$ such that at least one of these two vectors is not equal to zero, and at a real value of the complex variable z, denoted λ, the following equality is satisfied:

$$P(\lambda) \begin{bmatrix} x(0) \\ d(0) \end{bmatrix} = 0. \tag{11.23}$$

Then for each $k = 1, 2, \cdots$, the input vector sequence $d(k) = \lambda^k d(0)$ satisfies both $y(k, x(0), d(i)|_{i=0}^{k}) = 0$ and $d(k) \in \mathcal{R}^q$.

Proof. Note that the system defined by Eqs. (11.14) and (11.15) is well defined. This means that for a fixed initial state vector $x(0)$ and each input vector sequence $d(k)|_{k=0}^{\infty}$, there are only one state vector sequence $x(k)|_{k=0}^{\infty}$ and only one output vector sequence $y(k)|_{k=0}^{\infty}$ that satisfy these two equations simultaneously.

When condition (11.23) is satisfied, define $x(k)$ as $x(k)=\lambda^k x(0)$ for every $k=0,1,2,\cdots$. Then, we can straightforwardly prove that

$$x(1)=\lambda x(0)=Ax(0)+Bd(0).$$

Assume now that, for an integer i with $i\geq 0$, Eq. (11.14) is satisfied by $x(i)=\lambda^i x(0)$ and $d(i)=\lambda^i d(0)$. Then, we can straightforwardly prove that

$$\begin{aligned}x(i+2)&=\lambda^{i+2}x(0)=\lambda x(i+1)=\lambda\left[Ax(i)+Bd(i)\right]=A\left[\lambda x(i)\right]+B\left[\lambda d(i)\right]\\&=Ax(i+1)+Bd(i+1),\end{aligned}$$

that is, when $k=i+1$, the aforementioned state vector process and input vector process also satisfy Eq. (11.14). Therefore, $x(k)=\lambda^k x(0)$ and $d(k)=\lambda^k d(0)$ satisfy this equation for every $k\geq 0$. Hence, $x(k)=\lambda^k x(0)$ is the solution to this difference equation with its initial state vector $x(0)$ and input vector process $d(k)=\lambda^k d(0)$.

On the other hand, note that, for each $k=0,1,2,\ldots,$

$$Cx(k)+Dd(k)=C\left(\lambda^k x(0)\right)+D\left(\lambda^k d(0)\right)=\lambda^k\left[Cx(0)+Dd(0)\right]=0,$$

provided that the vectors $x(0)$ and $d(0)$ satisfy conditions (11.23). Therefore, $y(0)=0$, and for each $k=1,2,\cdots,$

$$y(k,x(0),d(i)|_{i=0}^{k-1})=0.$$

Now, assume that λ takes a complex value. Denote its real and imaginary parts respectively by λ_r and λ_j. Recall that the matrices A, B, C, and D and the vectors $x(0)$ and $d(0)$ are real valued. Satisfaction of Eq. (11.23) implies that

$$\begin{bmatrix}\lambda_r I-A\\ C\end{bmatrix}x(0)+\begin{bmatrix}-B\\ D\end{bmatrix}d(0)=0,\qquad \lambda_j x(0)=0.$$

In addition, note that

$$\begin{bmatrix}-B\\ D\end{bmatrix}=\begin{bmatrix}-I&0\\0&I\end{bmatrix}\begin{bmatrix}B\\ D\end{bmatrix},$$

which means that

$$\operatorname{rank}\left(\begin{bmatrix}-B\\ D\end{bmatrix}\right)=\operatorname{rank}\left(\begin{bmatrix}B\\ D\end{bmatrix}\right).$$

If $\lambda_j\neq 0$, then it is necessary that $x(0)=0$. As the matrix $\begin{bmatrix}B\\ D\end{bmatrix}$ is assumed to be of full column rank, $x(0)=0$ and the first subequation in the above equation imply that the vector

$d(0)$ is also equal to zero. This is a contradiction with the requirement that at least one of the vectors $x(0)$ and $d(0)$ is not equal to zero. Hence, $\lambda_j = 0$, that is, λ takes a real value.

These arguments show that if there is λ satisfying Eq. (11.23), then it is necessary that this λ is real valued. Hence, $\lambda^k d(0)$ is real valued for each $k = 0, 1, 2, \cdots$. This completes the proof. □

Clearly, a real vector satisfying Eq. (11.23) belongs to the weakly unobservable subspace $\mathcal{V}$ of the system.

From the proof of Lemma 11.2 it is obvious that if both the system initial state vector $x(0)$ and its input vector sequence $d(k)|_{k=0}^{\infty}$ are allowed to take complex values, then the associated results are valid even when the complex variable z takes a complex value that satisfies Eq. (11.23). In this case, both $x(0)$ and $d(0)$ may be complex vectors, which will lead to a complex-valued sequence $d(k)|_{k=0}^{\infty}$ that makes the associated system output vector sequence constantly equal to zero.

On the other hand, Lemma 11.2 does not require that λ is different from every eigenvalue of the state transition matrix A, which makes its results more general than those on system zero dynamics derived from Lemma 11.1. In fact, if λ is not an eigenvalue of the state transition matrix A, then Eq. (11.23) straightforwardly leads to $[C(\lambda I - A)^{-1}B + D]d(0) = 0$, which can be rewritten as $G(\lambda)d(0) = 0$. This is completely in the same form as the condition obtained from Lemma 11.1 for the system output vector sequence constantly equal to zero. These observations imply that when λ is not an eigenvalue of the matrix A, the conditions of Lemma 11.2 are equal to those derived from Lemma 11.1. Hence, we can regard that Lemma 11.2 includes some conclusions of Lemma 11.1 as its particular case.

To guarantee the existence of a nonzero vector $x(0)$ and/or a nonzero vector $d(0)$ such that Eq. (11.23) is satisfied, it is necessary that the matrix $P(\lambda)$ is not of full column rank. A complex value λ at which the system matrix $P(z)$ defined by Eq. (11.22) is not of full column rank is called a transmission zero of this system, which is an extension of a zero in a transfer function of a single-input single-output system to a zero of the transfer function matrix of a multiple-input multiple-output system [10,13].

Various methods have been developed for the computation of the transmission zeros of a multiple-input multiple-output system, among which one of most widely adopted methods appears to be the approach based on the so-called McMillan–Smith form [10,11]. In addition, the proof of the lemma reveals that to construct an actually realizable system initial state vector and an actually realizable system input vector process such that the system output vector process is constantly equal to zero, real-valued transmission zeros must be considered.

Our discussions show that if a system has a real transmission zero, then there certainly exist a real-valued initial state vector and a real-valued input vector sequence that lead to a constantly zero output vector sequence. However, it is still not clear that if there is a real-valued initial state vector $x(0)$ and a real-valued input vector sequence $d(k)|_{k=0}^{\infty}$ such that the associated output vector sequence $y(k, x(0), d(j)|_{j=0}^{k})$ is constantly zero, whether or not the system must have at least one real transmission zero.

To attack this problem, we need the following concept.

Definition 11.5. *If for all permissible input vector sequences $u_1(k)|_{k=0}^{\infty}$ and $u_2(k)|_{k=0}^{\infty}$, the satisfaction by the system output vector sequence $y(k)|_{k=0}^{\infty}$ of the condition*

$$y\left(k, 0, u_1(j)|_{j=0}^{k}\right) = y\left(k, 0, u_2(j)|_{j=0}^{k}\right)$$

for all $k = 0, 1, 2, \ldots$ implies that

$$u_1(k) = u_2(k)$$

for all $k = 0, 1, 2, \ldots,$ then this system is called left invertible.

When a system is linear, it is obvious from the definition that a necessary and sufficient condition for its left invertibility is that $y\left(k, 0, u(j)|_{j=0}^{k}\right) \equiv 0$ if and only if $u(k) \equiv 0$. For the linear time-invariant system described by Eqs. (11.14) and (11.15) that has a minimal realization, let $G(z)$ denote its transfer function matrix $C(zI - A)^{-1}B + D$. It has been proven that the following statements are equivalent to each other [11]:

- This system is left invertible.
- There exists a rational matrix-valued function $G_l(z)$ such that $G_l(z)G(z) = I$, that is, the transfer function matrix $G(z)$ has a left inverse.
- The system matrix $P(z)$ defined in Eq. (11.22) has a rank of $n+q$ for all but finitely many $z \in \mathcal{C}$.
- Let $\mathcal{V}$ denote the weakly unobservable subspace of the system. Then

$$\mathcal{V} \bigcap \{B \times \mathcal{N}ull(D)\} = \{\,0\,\} \quad \text{and} \quad \begin{bmatrix} B \\ D \end{bmatrix} \quad \text{is of full column rank.}$$

Related to the concept of weak observability, there is also a concept called strong observability.

Definition 11.6. *If for every permissible initial state vector $x(0)$ and every feasible input vector sequence $u(k)|_{k=0}^{\infty}$, the system output vector sequence $y(k)|_{k=0}^{\infty}$ satisfies the equality*

$$y\left(k, 0, u(j)|_{j=0}^{k}\right) = 0$$

for all $k = 0, 1, 2, \cdots$ *only when*

$$x(0) = 0,$$

then this system is called strongly observable.

For a linear time invariant system, the following results have been established in [11].

Lemma 11.3. *A system with its state space model described by Eqs. (11.14) and (11.15) is strongly observable, if and only if at every complex value of the complex variable z,*

$$\operatorname{rank}\left(\begin{bmatrix} zI - A & -B \\ C & D \end{bmatrix}\right) = n + \operatorname{rank}\left(\begin{bmatrix} B \\ D \end{bmatrix}\right).$$

From Definitions 11.3 and 11.6 it is obvious that a system is strongly observable if and only if its weakly unobservable subspace is constituted only from the zero vector. It has also been proven that, for the linear time-invariant system of Eqs. (11.14) and (11.15), its strong observability is equivalent to that, for each real matrix F with compatible dimension, the matrix pair $(A + BF,\ C + DF)$ is always observable [11].

Based on these results, the following conclusions are derived, which reveal situations under which the weakly unobservable subspace of a system is not restricted only to the zero vector.

Theorem 11.1. *Assume that the system described by Eqs. (11.14) and (11.15) is left invertible. Then there exists at least one nonzero initial state vector* $x(0)$ *such that there is an input vector sequence* $d(k)|_{k=0}^{\infty}$ *that makes the system output vector* $y\left(k, x(0), d(j)|_{j=0}^{k}\right)$ *constantly equal zero if and only if it has a transmission zero.*

Proof. Assume that the system has a real transmission zero. Then by Lemma 11.2 there certainly exists a nonzero initial state vector satisfying the requirements.

Now, assume that all the transmission zeros of the system are complex. Let λ be any one of them. Then there exist vectors $x_*(0)$ and $d_*(0)$ such that at least one of them is not equal to zero and the following equality is satisfied:

$$\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}\begin{bmatrix} x_*(0) \\ d_*(0) \end{bmatrix} = 0. \tag{11.24}$$

Recall that all the system parameter matrices A, B, C, and D are real. Taking the conjugates of both sides of this equation, we obtain the following equality:

$$\overline{\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}\begin{bmatrix} x_*(0) \\ d_*(0) \end{bmatrix}} = \begin{bmatrix} \bar{\lambda} I - A & -B \\ C & D \end{bmatrix}\begin{bmatrix} \bar{x}_*(0) \\ \bar{d}_*(0) \end{bmatrix} = 0. \tag{11.25}$$

As $\mathbf{col}\{x_*(0),\ d_*(0)\}$ does not equal zero, it is obvious that the vector $\mathbf{col}\{\bar{x}_*(0),\ \bar{d}_*(0)\}$ is not a zero vector either. These imply that $\bar{\lambda}_0$ is also a transmission zero of the system.

Define $d_*(k)$ as $d_*(k) = \lambda^k d(0)$, $k = 0, 1, 2, \cdots$. Let $d_*(z)$ and $d_\diamond(z)$ denote respectively the $\mathcal{Z}$-transformations of the sequences $d_*(k)|_{k=0}^{\infty})$ and $\bar{d}_*(k)|_{k=0}^{\infty})$. Moreover, let $y_*(z)$ and $y_\diamond(z)$ denote respectively the $\mathcal{Z}$-transformations of the output vector sequences $y(k, x_*(0), d_*(j)|_{j=0}^{k})$ and $y(k, \bar{x}_*(0), \bar{d}_*(j)|_{j=0}^{k})$. According to the proof of Lemma 11.2, we have that

$$y(k, x_*(0), d_*(j)|_{j=0}^{k-1}) = 0, \qquad y(k, \bar{x}_*(0), \bar{d}_*(j)|_{j=0}^{k-1}) = 0 \qquad \text{for all } k = 0, 1, 2, \cdots. \tag{11.26}$$

Define $x(0)$ and $d(k)$ further as

$$x(0) = x_*(0) + \bar{x}_*(0), \qquad d(k) = d_*(k) + \bar{d}_*(k) \qquad \text{for all } k = 0, 1, 2, \cdots.$$

Then, both $x(0)$ and $d(k)$ with $k \in \{0, 1, 2, \cdots\}$ are real valued and therefore can be realized in principle.

Let $y(z)$ denote the $\mathcal{Z}$-transformation of the output vector sequence associated with the system initial state vector $x(0)$ and the input vector sequence $d(k)|_{k=0}^{\infty}$. Then by Eq. (11.19) we have that

$$\begin{aligned} y(z) &= C(zI - A)^{-1}[x(0) + Bd(z)] + Dd(z) \\ &= C(zI - A)^{-1}\{[x_*(0) + \bar{x}_*(0)] + B[d_*(z) + d_\diamond(z)]\} + D[d_*(z) + d_\diamond(z)] \\ &= \Big\{C(zI - A)^{-1}[x_*(0) + Bd_*(z)] + Dd_*(z)\Big\} \\ &\quad + \Big\{C(zI - A)^{-1}[x_*(0) + Bd_\diamond(z)] + Dd_\diamond(z)\Big\} \\ &= y_*(z) + y_\diamond(z). \end{aligned} \tag{11.27}$$

We can therefore declare from Eq. (11.26) that, for each $k = 0, 1, 2, \cdots$,

$$y(k, x(0), d(j)|_{j=0}^{k}) = 0. \tag{11.28}$$

On the contrary, assume that the system described by Eqs. (11.14) and (11.15) does not have a transmission zero but there is a nonzero initial state vector $x(0)$ such that there exists an input sequence $d(k)|_{k=0}^{\infty}$ that makes $y(k, x(0), d(j)|_{j=0}^{k}) = 0$ for every $k = 0, 1, 2, \cdots$. Then, for arbitrary $\lambda \in \mathcal{C}$, the matrix

$$\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}$$

is of full column rank. Hence

$$\operatorname{rank}\left(\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}\right) = n + q = n + \operatorname{rank}\left(\begin{bmatrix} B \\ D \end{bmatrix}\right)$$

By Lemma 11.3 this system is strongly observable. Hence, if there is an input sequence $d(k)|_{k=0}^{\infty}$ that makes $y(k, x(0), d(j)|_{j=0}^{k}) = 0$ for every $k = 0, 1, 2, \cdots$, then $x(0) = 0$. This contradicts the assumption $x(0) \neq 0$. Therefore, the system must have some transmission zeros.

This completes the proof. □

Different from Lemma 11.2, the transmission zero in the theorem is not required to be real. On the other hand, both the initial state vector and the input vector sequence are required to be real.

11.3.2 Attack Prevention

From the results in the previous subsection it is clear that a linear time-invariant system is left invertible if and only if its transfer function matrix is of full normal column rank. This observation leads further to the following conclusions.

Corollary 11.1. *If an LTI system is not left invertible, then it is certainly not attack detectable/identifiable.*

Proof. Assume that a transfer function matrix $G(z)$ is not of FNCR. Then there exists a nonzero vector α such that, for each $\lambda \in \mathcal{C}$, the following equality is satisfied:

$$G(z)\alpha = 0. \tag{11.29}$$

Let $(A,\ B,\ C,\ D)$ be a real matrix quadruplex that satisfies $C(zI - A)^{-1}B + D = G(z)$ and associates with a state space model having a minimal realization. Take an arbitrary real scalar λ that is not an eigenvalue of the matrix A and denote the associated real and imaginary parts of the vector α by α_r and α_j, respectively. Then, the Eq. (11.29) means that the following two equalities are satisfied:

$$[C(\lambda I - A)^{-1}B + D]\alpha_r = 0, \qquad [C(\lambda I - A)^{-1}B + D]\alpha_j = 0. \tag{11.30}$$

Note that $\alpha \neq 0$ is equivalent to $\begin{bmatrix} \alpha_r \\ \alpha_j \end{bmatrix} \neq 0$, that is, the assumption $\alpha \neq 0$ implies that at least one of α_r and α_j is not equal to zero.

Assume that $\alpha_r \neq 0$. Then, the first equality of Eq. (11.30) can be equivalently rewritten as

$$\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} (\lambda I - A)^{-1} B\alpha_r \\ \alpha_r \end{bmatrix} = 0. \tag{11.31}$$

Note that both vectors $(\lambda I - A)^{-1} B\alpha_r$ and α_r are real. Moreover, the vector $\begin{bmatrix} (\lambda I - A)^{-1} B\alpha_r \\ \alpha_r \end{bmatrix}$ is not a zero vector. By Lemma 11.2 there always exist a nonzero $x(0) \in \mathcal{R}^n$ and an input vector sequence $d(k) \in \mathcal{R}^q$, not constantly equal to zero, such that the output vector process of the LTI system, which can be expressed as $y(k, x(0), d(i)|_{i=0}^{k})$, is constantly equal to zero for each $k = 0, 1, \cdots$. Hence the system is not attack detectable/identifiable.

When $\alpha_j \neq 0$, the same arguments show that the system is not attack detectable/identifiable.

This completes proof. □

These observations reveal that the problem of attack preventions is not trivial only when the associated system is left invertible.

With these preparations on system zero dynamics, we discuss now conditions respectively for the detectability and identifiability of attacks in a networked system.

Theorem 11.2. *Concerning the system described by Eqs. (11.14) and (11.15), an attack set $\mathcal{K}$ is undetectable if and only if there exists a number λ such that the system matrix*

$$P_d(\lambda) = \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} \\ C & D_{\mathcal{K}} \end{bmatrix} \tag{11.32}$$

is not of full column rank.

Proof. Assume that the matrix $P_d(\lambda)$ is not of full column rank at a particular real value of the complex variable λ, denoted λ_0. Then, there exist vectors $x(0)$ and $d(0)$ such that at least one of these two vectors is not a zero vector and

$$(\lambda_0 I - A)x(0) - B_{\mathcal{K}} d(0) = 0, \tag{11.33}$$

$$Cx(0) + D_{\mathcal{K}} d(0) = 0. \tag{11.34}$$

As all the involved matrices, that is, the state transition matrix A, the input matrix $B_{\mathcal{K}}$, the output matrix C, and the direct feedthrough matrix $D_{\mathcal{K}}$, and the scalar λ_0 are real valued, we can assume, without any loss of generality, that both vectors $x(0)$ and $d(0)$ are real valued. In fact, if they are not real valued, then denote their real and imaginary parts respectively by $x_r(0)$, $d_r(0)$, $x_j(0)$, and $d_j(0)$. Then Eqs. (11.33) and (11.34) imply that

$$(\lambda_0 I - A)x_r(0) - B_{\mathcal{K}}d_r(0) = 0, \qquad (\lambda_0 I - A)x_j(0) - B_{\mathcal{K}}d_j(0) = 0, \tag{11.35}$$

$$Cx_r(0) + D_{\mathcal{K}}d_r(0) = 0, \qquad Cx_j(0) + D_{\mathcal{K}}d_j(0) = 0. \tag{11.36}$$

Note that by their definitions all the vectors $x_r(0)$, $d_r(0)$, $x_j(0)$, and $d_j(0)$ are real valued. Moreover, at least one of them is not equal to zero. Reasonability of the adopted assumption immediately follows.

We now can declare the undetectability of the attack set $\mathcal{K}$ through a direct application of Lemma 11.2.

Assume now that the system matrix $P_d(\lambda)$ is not of full column rank at a particular complex value λ_0. If the system is not left invertible, then according to the aforementioned arguments, we can declare that the system is not detectable of the attack set $\mathcal{K}$.

Now, assume that the system is left invertible. Then from Theorem 11.1 we can declare that the attack set $\mathcal{K}$ is not detectable.

Assume that the attack set $\mathcal{K}$ is detectable. Then, by Corollary 11.1 its transfer function matrix must be of full normal column rank. From this observation and the definition of attack detectability, this system cannot have any transmission zero.

This completes proof. □

From Eqs. (11.33) and (11.34) we can also see that for the existence of an undetectable attack, it is necessary that the cardinality of the attack set is sufficiently large [7]. On the other hand, it is possible to improve attack detectability of a networked system through introducing state feedback into the system. The motivations are that a state feedback can modify transmission zeros of a system and the feedback gain can be designed with objectives that are unknown to an attacker.

Similarly, we have the following criterion for attack identifiability of the networked system described by Eqs. (11.14) and (11.15).

Theorem 11.3. *Concerning the system described by Eqs. (11.14) and (11.15), an attack set $\mathcal{K}$ is unidentifiable, if and only if there exist an attack set $\mathcal{R}$ with its cardinality not greater than that of the attack set $\mathcal{K}$ and a number λ such that the matrix*

$$P_i(\lambda) = \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} & -B_{\mathcal{R}} \\ C & D_{\mathcal{K}} & D_{\mathcal{R}} \end{bmatrix} \tag{11.37}$$

is not of full column rank.

Proof. Assume that at a particular real value of the complex variable λ, denoted λ_0, the matrix-valued function $P_i(\lambda)$ is not of full column rank. Then there exist vectors $x(0)$, $d_{\mathcal{K}}(0)$, and $d_{\mathcal{R}}(0)$ such that at least one of these three vectors is not a zero vector, and

$$(\lambda_0 I - A)x(0) - B_{\mathcal{K}} d_{\mathcal{K}}(0) - B_{\mathcal{R}} d_{\mathcal{R}}(0) = 0, \tag{11.38}$$

$$Cx(0) + D_{\mathcal{K}} d_{\mathcal{K}}(0) + D_{\mathcal{R}} d_{\mathcal{R}}(0) = 0. \tag{11.39}$$

Using the same arguments as in the proof of Theorem 11.2, we can assume, without loss of any generality, that the vectors $x(0)$, $d_{\mathcal{K}}(0)$, and $d_{\mathcal{R}}(0)$ are real valued. Then by Lemma 11.2 there exist an initial state vector $x(0) \in \mathcal{R}^n$ and an input vector sequence $u(k) \in \mathcal{R}^{K+R}$, $k = 0, 1, \cdots$, such that the output vector sequence $y(k)$ of the following system is constantly equal to zero.

$$x(k+1) = Ax(k) + [B_{\mathcal{K}} \;\; B_{\mathcal{R}}]u(k), \tag{11.40}$$

$$y(k) = Cx(k) + [D_{\mathcal{K}} \;\; D_{\mathcal{R}}]u(k). \tag{11.41}$$

Moreover, at least either the initial state vector $x(0)$ is not a zero vector, or the input vector sequence $u(k)|_{k=0}^{\infty}$ is not constantly equal to zero. Note that both matrices $\begin{bmatrix} B_{\mathcal{K}} \\ D_{\mathcal{K}} \end{bmatrix}$ and $\begin{bmatrix} B_{\mathcal{R}} \\ D_{\mathcal{R}} \end{bmatrix}$ are of full column rank. Direct algebraic manipulations show that, for this plant initial state vector $\tilde{x}(0) \in \mathcal{R}^n$, there exist an attack $d(k)|_{k=0}^{\infty}$ with K elements in $d(k)|_{k=0}^{\infty}$ not constantly zero, an attack $\bar{d}(k)|_{k=0}^{\infty}$ with $\bar{K}$ elements in $d(k)|_{k=0}^{\infty}$ not constantly zero, and the integer $\bar{K}$ satisfying $0 \leq \bar{K} \leq K$ such that

$$y(k, \tilde{x}(0), d(j) - \bar{d}(j)|_{j=0}^{k}) \equiv 0, \qquad \left|\left| d(k)|_{k=0}^{\infty} - \bar{d}(k)|_{k=0}^{\infty} \right|\right| \neq 0,$$

that is, the attack set $\mathcal{K}$ is not identifiable.

By the same token as that adopted in the proof of Theorem 11.2, we can prove that even if the aforementioned λ_0 takes a particular complex value, the system is also unidentifiable with respect to the attack set $\mathcal{K}$.

This completes the proof. □

On the basis of geometric linear system theories, some graph theoretic conditions have also been established for the existence of an undetectable attack set in a networked system and for the existence of an unidentifiable attack set in a networked system. These conditions depend only on system structures and are valid for almost all system parameters when the networked system has a compatible structure. A system property with these characteristics is extensively called generic and has the advantage of strong robustness against parametric modeling errors [14]. When the initial state vector of a networked system is known, it is proven that if the attack-state-output graph is sufficiently connected, then undetectable attacks do not exist in almost all structurally compatible networked systems. When the initial state vector is not known for a networked system, the criterion becomes more complicated, and left invertibility of the system may be required. Detailed discussions can be found, for example, in [7,8].

11.4 Detection of Attacks

In the above section, we developed some criteria to verify whether or not an attack can be detected or identified. These criteria are closely related to the transmission zeros of a networked system and can be verified in principle. When the scale of the system is large, however, computations of its transmission zeros may be computationally prohibitive and/or numerically unstable. This means that further efforts are required to develop a computationally attractive method for verifying attack detectability and identifiability. To achieve this objective, the methods given in Chapter 3 may be helpful, in which structure information is explicitly and efficiently utilized in the verification of controllability and observability of a large-scale networked system, noting that both the system matrix $P_d(\lambda)$ of Eq. (11.32) and the system matrix $P_i(\lambda)$ of Eq. (11.37) have a similar form of the matrix-valued polynomial $M(\lambda)$ of Eq. (3.29). In this section, we investigate how to detect an attack using a centralized observer.

The design of an attack detector consists of a discrete-time modified Luenberger observer, which is sometimes also called a residual filter, with its input being the plant output measurements $y(k)|_{k=0}^{\infty}$ and its output being a residual signal $r(k)|_{k=0}^{\infty}$:

$$z(k+1) = Az(k) - L[y(k) - Cz(k)], \tag{11.42}$$

$$r(k) = Cz(k) - y(k), \tag{11.43}$$

where the matrix L is usually called the gain of the observer, which is selected to make the matrix $A + LC$ stable, that is, all its eigenvalues have a magnitude smaller than 1. When the matrix pair $(A,\ C)$ is observable, the existence of a desirable gain matrix L is always guaranteed. The residue signal $r(k)$ is used to detect the existence of an attack. Ideally, this signal is constantly equal to zero if and only if there do not exist any attacks in the networked system. As the initial state vector of the observer also affects the residual signal, these expectations cannot be satisfied in general, and some modifications are required to develop a practically meaningful criterion to check the existence of an attack. However, we have the following theoretical results on the aforementioned residual filter.

Theorem 11.4. *Concerning the system described by Eqs. (11.14) and (11.15), assume that an attack set $\mathcal{K}$ is detectable and the initial state vector of the system, that is, $x(0)$, is known. Assume also that the initial state vector of the residual filter described by Eqs. (11.42) and (11.43), that is, $z(0)$, is set to $z(0) = x(0)$, and the gain matrix L is selected such that the matrix $A + LC$ is stable. Then, the residual signal of the filter is constantly equal to zero if and only if the attack disturbance is constant equal to zero.*

Proof. Define the vector-valued function $e(k) = z(k) - x(k)$. Then from Eqs. (11.14), (11.15), (11.42), and (11.43) we can straightforwardly prove that

$$\begin{aligned} e(k+1) &= z(k+1) - x(k+1) \\ &= \{Az(k) - L[Cx(k) + D_{\mathcal{K}}d_{\mathcal{K}}(k) - Cz(k)]\} - \{Ax(k) + B_{\mathcal{K}}d_{\mathcal{K}}(k)\} \\ &= (A+LC)e(k) - (LD_{\mathcal{K}} + B_{\mathcal{K}})d_{\mathcal{K}}(k), \qquad (11.44) \\ r(k) &= Cz(k) - [Cx(k) + D_{\mathcal{K}}d_{\mathcal{K}}(k)] \\ &= Ce(k) - D_{\mathcal{K}}d_{\mathcal{K}}(k). \qquad (11.45) \end{aligned}$$

When the initial state vector of the detector is set to be the same value of the initial state vector of the system, that is, $z(0) = x(0)$, we have that $e(0) = 0$. Hence, when the matrix $A + LC$ has all its eigenvalues being smaller than 1 in magnitude and the attack disturbances do not exist, that is, $d_{\mathcal{K}}(k) \equiv 0$, it is obvious from Eq. (11.44) that the error vector sequence $e(k)$ with $k = 0, 1, 2, \cdots$ is also constantly equal to zero. This further implies that the output vector of the detector, that is, the residual vector $r(k)$ is also constantly equal to zero when the temporal variable k takes any value from the set $\{0,\ 1,\ 2,\ \cdots\}$.

Now assume that there exists a nonzero attack vector sequence $d_{\mathcal{K}}(k)|_{k=0}^{\infty}$ such that the output of the attack detector described by Eqs. (11.42) and (11.43) is constantly equal to zero. Then by Theorem 11.4 there must exist at least one $\lambda \in \mathcal{C}$, one vector $\bar{e}(0)$, and one vector $\bar{d}_{\mathcal{K}}(0)$ such that at least one of these two vectors is not equal to zero and

$$\begin{bmatrix} \lambda I - (A+LC) & LD_{\mathcal{K}} + B_{\mathcal{K}} \\ C & -D_{\mathcal{K}} \end{bmatrix} \begin{bmatrix} \bar{e}(0) \\ \bar{d}_{\mathcal{K}}(0) \end{bmatrix} = 0. \qquad (11.46)$$

Note that

$$\begin{bmatrix} \lambda I - (A+LC) & LD_{\mathcal{K}} + B_{\mathcal{K}} \\ C & -D_{\mathcal{K}} \end{bmatrix} = \begin{bmatrix} I & -L \\ 0 & I \end{bmatrix} \begin{bmatrix} \lambda I - A & B_{\mathcal{K}} \\ C & -D_{\mathcal{K}} \end{bmatrix}.$$

Note also that the matrix $\begin{bmatrix} I & -L \\ 0 & I \end{bmatrix}$ is invertible for every real-valued matrix L with compatible dimension. It is obvious that the satisfaction of Eq. (11.46) is equivalent to the satisfaction of the equation

$$\begin{bmatrix} \lambda I - A & B_{\mathcal{K}} \\ C & -D_{\mathcal{K}} \end{bmatrix} \begin{bmatrix} \bar{e}(0) \\ \bar{d}_{\mathcal{K}}(0) \end{bmatrix} = 0, \tag{11.47}$$

which contradicts the assumption that the attack set $\mathcal{K}$ is detectable. This completes the proof. □

Although the results of Theorem 11.4 are promising, the system initial state vector $x(0)$ is usually not known exactly. Under such a situation, the condition $z(0) = x(0)$ can generally not be satisfied properly for the initial state vector of the residual filter, and $z(0)$ may be selected with some arbitrariness. This means that even if there do not exist attacks in the networked system, the residual signal $r(k)$ may not be constantly equal to zero. However, it can be proven that when the networked system has not been attacked, the residual signal $r(t)$ asymptotically converges to zero with the increment of the temporal variable k. On the other hand, measurement errors and external random disturbances are usually unavoidable in actual networked systems, which means that even if the conditions of Theorem 11.4 are perfectly satisfied, the residual signal $r(k)$ may not be constantly equal to zero and some statistical hypothesis verification methodologies appear necessary for the determination of whether or not there are attacks in the networked system. In addition to these, parametric uncertainties and unmodeled dynamics usually exist in the model of Eqs. (11.14) and (11.15). Hence, in actual implementations of attack detections, in addition to construct a stable residual filter, the gain matrix L should be selected to reduce sensitivities of the residual signal $r(k)$ to modeling errors while keeping it satisfactorily sensitive to attacks. Furthermore, for a large-scale networked system, it is usually more attractive to implement an attack detector in a distributed way using only local subsystem output measurements. Detailed discussions on these issues are given in various literature, for example, [9], [8], and [7].

11.5 Identification of Attacks

In actual applications, it is usually not only necessary to know whether or not there exists an attack in a networked system, but also essential to know where the attack is from when it exists. Differently from attack detection, attack identification is computationally much more difficult due to its combinatorial characteristics, although their essential treatments are the same. In fact, it has been proven in [8] that attack identifications are generally NP hard with respect to the number of colluding attackers. More precisely, the identification of an attack from the attack set $\mathcal{K}$ usually requires a combinatorial procedure, noting that even when the number q of attackers is known, the actual attack is only one of the $\begin{pmatrix} q \\ K \end{pmatrix}$ possible attacks. To identify this attack, each possibility must be considered, and for each possible attack, a residual filter needs be designed. When the attack set is identifiable, there exists one and only one residual filter whose outputs are constantly equal to zero. It is this residual filter that indicates the actual attack.

Unlike the residual filter in attack detection, the model of Eqs. (11.14) and (11.15) is not straightforwardly utilized in constructing the residual filters in attack identification. Usually, the design of these residual filters consists of three steps. First, a transformation is applied to the plant output vector to construct a vector without any attacks. Second, a state transformation is performed to divide the plant state vector into two parts, one affected by attack disturbances and the other is free from attack disturbances. The third step is to construct a residual filter with the transformed plant output vector and the transformed plant state vector that are not affected by attack disturbances, which is completely the same as that in the construction procedure for attack detections.

Now, we discuss the first step, in which the attack identification problem for the networked system described by Eqs. (11.14) and (11.15) is converted to that of a modified system in which the associated system output vector is not attacked.

For this purpose, let $D_{\mathcal{K}}^{\dagger}$ denote the pseudo-inverse of the matrix $D_{\mathcal{K}}$. Define an auxiliary networked system as

$$x(k+1) = [A - B_{\mathcal{K}} D_{\mathcal{K}}^{\dagger} C]x(k) + B_{\mathcal{K}}[I - D_{\mathcal{K}}^{\dagger} D_{\mathcal{K}}]d(k), \tag{11.48}$$

$$y(k) = [I - D_{\mathcal{K}} D_{\mathcal{K}}^{\dagger}]Cx(k). \tag{11.49}$$

Then, we have the following conclusions.

Lemma 11.4. *Let the networked system be described by Eqs. (11.14) and (11.15). The attack set $\mathcal{K}$ is identifiable if and only if this attack set is identifiable for the networked system described by Eqs. (11.48) and (11.49).*

Proof. According to Theorem 11.3, if the attack set $\mathcal{K}$ is identifiable for the networked system described by Eqs. (11.14) and (11.15), then for each attack set $\mathcal{R}$ with its cardinality not greater than that of the attack set $\mathcal{K}$, the following system matrix $P_i(\lambda)$ is of full column rank at each complex number λ:

$$P_i(\lambda) = \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} & -B_{\mathcal{R}} \\ C & D_{\mathcal{K}} & D_{\mathcal{R}} \end{bmatrix}. \tag{11.50}$$

It is equivalent to that the matrix $\bar{P}_i(\lambda)$ defined as

$$\bar{P}_i(\lambda) = \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} & -B_{\mathcal{R}} \\ C & D_{\mathcal{K}} & D_{\mathcal{R}} \\ C & D_{\mathcal{K}} & D_{\mathcal{R}} \end{bmatrix} \tag{11.51}$$

is of full column rank at each complex number λ.

Multiplying both sides of the above equation from the left by $\mathbf{diag}\left\{I,\ D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger},\ I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}\right\}$, we have that

$$\begin{bmatrix} I & & \\ & D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger} & \\ & & I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger} \end{bmatrix} \bar{P}_i(\lambda) = \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} & -B_{\mathcal{R}} \\ D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C & D_{\mathcal{K}} & D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}D_{\mathcal{R}} \\ [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C & 0 & [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]D_{\mathcal{R}} \end{bmatrix}. \tag{11.52}$$

Note that

$$\begin{aligned} &\begin{bmatrix} I & B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger} & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} \lambda I - A & -B_{\mathcal{K}} & -B_{\mathcal{R}} \\ D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C & D_{\mathcal{K}} & D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}D_{\mathcal{R}} \\ [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C & 0 & [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]D_{\mathcal{R}} \end{bmatrix} \\ &= \begin{bmatrix} \lambda I - A + B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C & -B_{\mathcal{K}}(I - D_{\mathcal{K}}^{\dagger}D_{\mathcal{K}}) & -B_{\mathcal{R}} + B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}D_{\mathcal{R}} \\ [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C & 0 & [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]D_{\mathcal{R}} \\ D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C & D_{\mathcal{K}} & D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}D_{\mathcal{R}} \end{bmatrix}. \end{aligned} \tag{11.53}$$

Moreover, the matrix

$$\begin{bmatrix} I & B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger} & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}$$

is always invertible, and it is obvious that the above equations mean that the matrix-valued function

$$\begin{bmatrix} \lambda I - A + B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C & -B_{\mathcal{K}}(I - D_{\mathcal{K}}^{\dagger}D_{\mathcal{K}}) & -B_{\mathcal{R}} + B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}D_{\mathcal{R}} \\ [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C & 0 & [I - D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]D_{\mathcal{R}} \end{bmatrix}$$

is also of full column rank at each complex λ. It can therefore be concluded from Theorem 11.3 that the attack set $\mathcal{K}$ is also identifiable for the networked system described by Eqs. (11.48) and (11.49).

On the contrary, if the attack set $\mathcal{K}$ is identifiable for the networked system described by Eqs. (11.48) and (11.49), then similar arguments show that this attack set is also identifiable for the networked system described by Eqs. (11.14) and (11.15).

This completes the proof. □

To construct a residual signal that is able to identify an attack, the state space model of Eqs. (11.48) and (11.49) is further transformed. To develop this transformation, however, some concepts and results are needed from geometric control theory, which can be found, for example, in [11,12].

Definition 11.7. *Assume that the dynamics of a discrete-time and linear time-invariant system can be described by the following state space model:*

$$x(k+1) = Ax(k) + Bu(k),$$
$$y(k) = Cx(k),$$

where $x(k) \in \mathcal{R}^n$, $y(k) \in \mathcal{R}^p$, and $u(k) \in \mathcal{R}^q$ denote respectively the state vector, input vector, and output vector of the system. Moreover, assume that $\mathcal{S}$ is a subspace of $\mathcal{R}^n$. If

$$A\left(\mathcal{S}\bigcap\mathcal{N}ull(C)\right) \subseteq \mathcal{S},$$

then this subspace is said to be $(A,\ \mathcal{N}ull(C))$-conditioned invariant.

A conditioned invariant subspace is closely related to estimator designs. A well-known result is that the smallest $(A,\ \mathcal{N}ull(C))$-conditioned invariant subspace that contains the subspace $\mathcal{S}pan(B)$ as its subset is the largest subspace in the state space $\mathcal{R}^n$ of the system, which can be estimated in the presence of an unknown input sequence $u(k)|_{k=0}^{\infty}$. This conclusion is obviously of great significance in engineering, as it clarifies situations for the existence of a state estimator with unknown inputs. Another well-known result is that the null space of the observability matrix, that is,

$$\mathcal{N}ull\left(\begin{bmatrix} C^T & A^TC^T & (A^T)^2C^T & \cdots & (A^T)^{n-1}C^T \end{bmatrix}^T\right),$$

which gives all the initial state vectors that result in a zero system output when there do not exist any external inputs are an $(A,\ \mathcal{N}ull(C))$-conditioned invariant subspace.

It has been proven that for each $(A,\ \mathcal{N}ull(C))$-conditioned invariant subspace $\mathcal{S}$, there exists at least one real matrix L such that

$$(A+LC)\mathcal{S}\subseteq\mathcal{S},$$

that is, an output injection matrix that renders this subspace invariant.

To apply these concepts and results to the system described by Eqs. (11.48) and (11.49), let $\bar{\mathcal{S}}_{\mathcal{K}}$ denote the smallest $\left(A-B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C,\ \mathcal{N}ull([I-D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C)\right)$-conditioned invariant subspace that contains $\mathcal{S}pan(B_{\mathcal{K}}[I-D_{\mathcal{K}}^{\dagger}D_{\mathcal{K}}])$ as its subspace. Moreover, let L be a matrix that satisfies

$$\bar{\mathcal{S}}_{\mathcal{K}}\supseteq\left(A-B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C+L[I-D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]C\right)\bar{\mathcal{S}}_{\mathcal{K}}$$

Existence of a desirable matrix L is guaranteed by the properties of the subspace $\bar{\mathcal{S}}_{\mathcal{K}}$. Furthermore, let T be a nonsingular square matrix satisfying

$$T[A-B_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}C+LC]T^{-1}=\begin{bmatrix}\bar{A}_{11} & \bar{A}_{12}\\ 0 & \bar{A}_{22}\end{bmatrix},\qquad TB_{\mathcal{K}}[I-D_{\mathcal{K}}^{\dagger}D_{\mathcal{K}}]=\begin{bmatrix}\bar{B}_{\mathcal{K}}\\ 0\end{bmatrix},\tag{11.54}$$

$$[I-D_{\mathcal{K}}D_{\mathcal{K}}^{\dagger}]CT^{-1}=[\bar{C}_1\ \ \bar{C}_2]\tag{11.55}$$

where $\bar{B}_{\mathcal{K}}$ is a matrix of full row rank. We can prove using geometric control theories that a desirable matrix T always exists [8,11,12]. As a matter of fact, the matrix T can be constituted from a basis of the subspace $\bar{\mathcal{S}}_{\mathcal{K}}$ and a basis of the quotient subspace $\mathcal{R}^n\backslash\bar{\mathcal{S}}_{\mathcal{K}}$ [11,12].

With the aforementioned matrix transformation, we can establish the following conclusions.

Lemma 11.5. *Let the networked system be described by Eqs. (11.14) and (11.15). The attack set $\mathcal{K}$ is identifiable if and only if this attack set is identifiable for the following networked system:*

$$\begin{bmatrix}x_1(k+1)\\ x_2(k+1)\end{bmatrix}=\begin{bmatrix}\bar{A}_{11} & \bar{A}_{12}\\ 0 & \bar{A}_{22}\end{bmatrix}\begin{bmatrix}x_1(k)\\ x_2(k)\end{bmatrix}+\begin{bmatrix}\bar{B}_{\mathcal{K}}\\ 0\end{bmatrix}d(k),\tag{11.56}$$

$$y(k)=[\bar{C}_1\ \ \bar{C}_2]\begin{bmatrix}x_1(k)\\ x_2(k)\end{bmatrix}.\tag{11.57}$$

Proof. Assume that the networked system described by Eqs. (11.56) and (11.57) is attack identifiable for the attack set $\mathcal{K}$. Then, for each complex number λ and each attack set $\mathcal{R}$ with its attacker number not greater than k, if

$$\begin{bmatrix}\lambda I-\bar{A}_{11} & -\bar{A}_{12} & \bar{B}_{\mathcal{K}} & -B_{\mathcal{R}1}\\ 0 & \lambda I-\bar{A}_{22} & 0 & -B_{\mathcal{R}2}\\ \bar{C}_1 & \bar{C}_2 & 0 & D_{\mathcal{R}}\end{bmatrix}\begin{bmatrix}x_1(0)\\ x_2(0)\\ g_{\mathcal{K}}\\ g_{\mathcal{R}}\end{bmatrix}=0,\tag{11.58}$$

then it is necessary that all the vectors $x_1(0)$, $x_2(0)$, $g_{\mathcal{K}}$, and $g_{\mathcal{R}}$ are zero vectors, that is,

$$\left[x_1^T(0)\ x_2^T(0)\ g_{\mathcal{K}}^T\ g_{\mathcal{R}}^T\right]^T = 0. \tag{11.59}$$

Note that

$$\begin{aligned}
\begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} & \bar{B}_{\mathcal{K}} \\ 0 & \bar{A}_{22} & 0 \\ \bar{C}_1 & \bar{C}_2 & 0 \end{bmatrix} &= \begin{bmatrix} T[A - B_{\mathcal{K}} D_{\mathcal{K}}^{\dagger} C + LC]T^{-1} & T B_{\mathcal{K}}[I - D_{\mathcal{K}}^{\dagger} D_{\mathcal{K}}] \\ \left[I - D_{\mathcal{K}} D_{\mathcal{K}}^{\dagger}\right] C T^{-1} & 0 \end{bmatrix} \\
&= \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A - B_{\mathcal{K}} D_{\mathcal{K}}^{\dagger} C + LC & B_{\mathcal{K}}[I - D_{\mathcal{K}}^{\dagger} D_{\mathcal{K}}] \\ \left[I - D_{\mathcal{K}} D_{\mathcal{K}}^{\dagger}\right] C & 0 \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}.
\end{aligned} \tag{11.60}$$

We therefore have that

$$\begin{aligned}
&\begin{bmatrix} \lambda I - \bar{A}_{11} & -\bar{A}_{12} & \bar{B}_{\mathcal{K}} & -B_{\mathcal{R}1} \\ 0 & \lambda I - \bar{A}_{22} & 0 & -B_{\mathcal{R}2} \\ \bar{C}_1 & \bar{C}_2 & 0 & D_{\mathcal{R}} \end{bmatrix} = \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \\
&\times \begin{bmatrix} \lambda I - [A - B_{\mathcal{K}} D_{\mathcal{K}}^{\dagger} C + LC] & -B_{\mathcal{K}}[I - D_{\mathcal{K}}^{\dagger} D_{\mathcal{K}}] & -T^{-1} \begin{bmatrix} B_{\mathcal{R}1} \\ B_{\mathcal{R}2} \end{bmatrix} \\ -\left[I - D_{\mathcal{K}} D_{\mathcal{K}}^{\dagger}\right] C & 0 & D_{\mathcal{R}} \end{bmatrix} \begin{bmatrix} T^{-1} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}.
\end{aligned} \tag{11.61}$$

As the matrix T is regular by its definition, it is obvious that both two matrices

$$\begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix}, \quad \begin{bmatrix} T^{-1} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

are well defined and are of full rank. We can therefore claim from Eqs. (11.58) and (11.61) that the matrix

$$\begin{bmatrix} \lambda I - [A - B_{\mathcal{K}} D_{\mathcal{K}}^{\dagger} C + LC] & -B_{\mathcal{K}}[I - D_{\mathcal{K}}^{\dagger} D_{\mathcal{K}}] & B_{\mathcal{R}} \\ -\left[I - D_{\mathcal{K}} D_{\mathcal{K}}^{\dagger}\right] C & 0 & D_{\mathcal{R}} \end{bmatrix}$$

is also of full column rank for each complex number λ and each attack set $\mathcal{R}$ with its attacker number not greater than that of the attack set $\mathcal{K}$. Hence, from Lemma 11.4 and Theorem 11.3

we can declare attack identifiability for the attack set $\mathcal{K}$ in the networked system described by Eqs. (11.14) and (11.15).

Similar arguments show that if the attack set $\mathcal{K}$ is identifiable for the networked system described by Eqs. (11.14) and (11.15), then it is also identifiable for the networked system described by Eqs. (11.56) and (11.57).

This completes the proof. □

Note that in the state space model of Eqs. (11.56) and (11.57), the state vector $x_2(k)$ is completely isolated from any attack disturbances. Using this property, a residual filter can be constructed as

$$w(k+1) = [\bar{A}_{22} - \bar{L}(I - \bar{C}_1\bar{C}_1^\dagger)]w(k) - \bar{L}\bar{y}(k), \tag{11.62}$$

$$\bar{y}(k) = [I - \bar{C}_1\bar{C}_1^\dagger]y(k), \tag{11.63}$$

$$r_{\mathcal{K}}(k) = [I - \bar{C}_1\bar{C}_1^\dagger]\bar{C}_2 w(k) - \bar{y}(k). \tag{11.64}$$

This residual filter is very similar to that for attack detections given in the previous section, except that the output vector $y(k)$ is replaced by a modified output vector $\bar{y}(k)$. Obviously, this modification is removing $\bar{C}_1 x_1(k)$ from the output vector $y(k)$ with the purpose of constructing an output vector that is not affected by any attack disturbances, noting that the substate vector $x_2(k)$ is independent of the attack disturbance vector $d(k)$ and

$$\begin{aligned}
\bar{y}(k) = [I - \bar{C}_1\bar{C}_1^\dagger]y(k) &= [I - \bar{C}_1\bar{C}_1^\dagger]\begin{bmatrix}\bar{C}_1 & \bar{C}_2\end{bmatrix}\begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} \\
&= [I - \bar{C}_1\bar{C}_1^\dagger]\bar{C}_2 x_2(k).
\end{aligned}$$

With this residual filter, we have the following results.

Theorem 11.5. *Let the networked system be described by Eqs. (11.14) and (11.15). Assume that the attack set $\mathcal{K}$ is identifiable. Moreover, assume that the initial state vector $x(0)$ of the networked system is known. Furthermore, assume that the initial state vector of the residual filter described by Eqs. (11.62)–(11.64) is set as $w(0) = x_2(0)$ and that the gain matrix $\bar{L}$ is selected such that the matrix $\bar{A}_{22} - \bar{L}(I - \bar{C}_1\bar{C}_1^\dagger)\bar{C}_2$ is stable. Then the residual signal $r(k)$ of the above filter is constantly equal to zero if and only if both matrices $B_{\mathcal{K}}$ and $D_{\mathcal{K}}$ coincide with those of the actual colluding attackers.*

Proof. Using the state vector $w(k)$ of the attack identifier described by Eqs. (11.62)–(11.64), define the new state vector $\bar{w}(k)$ evolving as

$$\bar{w}(k+1) = \bar{A}_{11}w(k) + \bar{A}_{12}\bar{w}(k). \tag{11.65}$$

Moreover, define the error vector $e(k) = \mathbf{col}\{e_1(k),\ e_2(k)\}$, in which two suberror vectors $e_1(k)$ and $e_2(k)$ are respectively defined as $e_1(k) = \bar{w}(k) - x_1(k)$ and $e_2(k) = w(k) - x_2(k)$. Then on the basis of Eqs. (11.14), (11.15), (11.62)–(11.64), and (11.65), straightforward algebraic manipulations show that

$$e(k+1) = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & \bar{A}_{22} - \bar{L}(I - \bar{C}_1\bar{C}_1^{\dagger})\bar{C}_2 \end{bmatrix} e(k) - \begin{bmatrix} \bar{B}_{\mathcal{K}} \\ 0 \end{bmatrix} d(k), \qquad (11.66)$$

$$r_{\mathcal{K}}(k) = \begin{bmatrix} 0 & (I - \bar{C}_1\bar{C}_1^{\dagger})\bar{C}_2 \end{bmatrix} e(k). \qquad (11.67)$$

When $w(0)$ is set to be equal to $x_2(0)$, the definition of the suberror vector $e_2(k)$ implies that $e(0) = 0$. As the residual signal $r_{\mathcal{K}}(k)$ is not directly affected by the attack disturbances $d(k)$, and the matrix $\bar{L}$ is selected to assure that the magnitude of each eigenvalue of the matrix $\bar{A}_{22} - \bar{L}(I - \bar{C}_1\bar{C}_1^{\dagger})\bar{C}_2$ is smaller than 1, this residual signal $r_{\mathcal{K}}(k)$ is constantly equal to zero, provided that the attack set is $\mathcal{K}$. This completes the proof. □

In case that there exist modeling errors in the adopted state space model, and/or there are random disturbances in the networked system, and/or the initial state vector $x(0)$ is not exactly known, observations in attack detections are also valid for the above attack identification methods. On the other hand, from the above results it is clear that to identify an attack, it is necessary to construct at least one residual filter for each possible attack. This might be not feasible when the number K is large.

Similarly to the attack detector discussed in the previous section, it is also possible to realize the above attack identifier in a distributed way. The details are not included in this book. Some preliminary results can be found in [8,9].

11.6 System Security and Sensor/Actuator Placement

The previous sections make it clear that to detect and/or identify an attack in networked systems, system observability is an essential property. In Chapter 3, we have discussed how to verify observability of a large-scale networked system and how many sensors are required to construct an observable networked system. However, it is usually not sufficient in actual engineering to only detect and/or identify an attack. A general requirement is to achieve this objective in an allowed time period after the occurrence of an attack, if not in the shortest time period. In fact, the longer the time is needed to detect/identify an attack, the larger the damages the attack may bring.

Note that both attack detection and attack identification depend on a residual generator. As system initial states cannot be known exactly in general and both modeling errors and external disturbances are usually unavoidable in actual applications, the time period and/or data length needed to detect/identify an attack depend heavily on the performances of that residual generator in its robustness against modeling errors and its capabilities of reducing influences from external disturbances. From the results of Section 11.5 it is clear that this residual generator can be easily rewritten into a form of the Luenberger observer discussed in Chapter 4, which includes the Kalman filter as its particular form. On the other hand, recall that the Kalman filter is optimal when the plant is linear and the external disturbances are normally distributed and when the sensitivity penalization-based robust state estimator given in Chapter 4 has completely the same form as that of the Kalman filter. It is interesting from both theoretical aspects and application aspects to investigate relations among estimation performances of the Kalman filter, the number of sensors, and the length of system output measurements [15]. A dual problem is about relations about control efforts, the number of actuators and optimal control performances, which is also an interesting topic in the analysis and synthesis of networked systems [16,17]. In fact, an attacker often intends to destroy a system with minimal efforts in a stealthy way, whereas a controller usually wants to improve system performances with the smallest energy consumptions.

In this section, we investigate a sensor placement problem and an actuator placement problem, both of them related closely to system security.

From a mathematical viewpoint, both sensor placement problems and actuator placement problems can be regarded as set function optimizations. When a sensor placement problem is under investigation, a set $\mathcal{V}$ is usually adopted to denote the set consisting of all potential positions of a networked system in which a sensor can be placed, whereas the function $f(\cdot)$ represents metric measuring factors like the best performances that a state estimator can achieve with the measurements from several sensors belonging to a set $\mathcal{S}$, the costs of these sensors, and so on. Then, when there are restrictions on the number of sensors in a networked system, say, it does not exceed k, a sensor placement problem may be mathematically described as

$$\max_{\mathcal{S}\subseteq\mathcal{V},\ |\mathcal{S}|\leq k} f(\mathcal{S}). \tag{11.68}$$

Here $|\mathcal{S}|$ stands for the number of elements in the set $\mathcal{S}$. Basically, this is a finite combinatorial optimization problem, which can be solved in principle through exhausting all possible sensor sets that have its element number not greater than k and comparing their associated cost function values. When the number of subsystems in a networked system is large, as possibilities of sensor locations in general increase exponentially with the subsystem number, such a brute-force-based method is usually computationally prohibitive.

Similar statements are valid for actuator placements.

To deal with these sensor/actuator placement problems, assume that the dynamics of a discrete time and linear time-variant system can be described by the following state space model:

$$x(k+1) = A(k)x(k) + B(k)w(k), \tag{11.69}$$

$$y(k) = C(k)x(k) + v(k). \tag{11.70}$$

Once again, here, $x(k) \in \mathcal{R}^n$, $y(k) \in \mathcal{R}^p$, and $w(k) \in \mathcal{R}^q$ denote respectively the state vector, input vector, and output vector of the system. Moreover, assume that $\mathcal{S}$ is a subspace of $\mathcal{R}^n$.

11.6.1 Some Properties of the Kalman Filter

It is now widely known that when a plant is linear and its external disturbances are normally distributed, the Kalman filter is the optimal state estimator under the criterion of mean squares errors [18,19]. Recent studies, however, show that when mean squares errors are adopted in the determination of optimal sensor locations, the associated optimization problem generally does not have the submodularity property and is in general NP-hard. Moreover, the widely adopted greedy heuristic algorithm may perform arbitrarily poorly, and there does not exist a constant-factor (polynomial-time) approximation algorithm [20,21].

To deal with the problem of appropriately locating sensors for a networked system, in this section, we investigate relations among estimation error, data length, and plant output number in the Kalman filtering. As only stochastic properties of $B(k)w(k)$ affect estimation accuracies of the Kalman filter [18,19,22], we can assume without any loss of generality that $B(k) \equiv I_n$. To simplify mathematical expressions, this assumption is adopted throughout this and next subsections. Moreover, we assume that the system initial conditions, process disturbances, and measurement errors are white and uncorrelated. More precisely:

- $x(0)$, $w(k)$, and $v(k)$ are uncorrelated with each other for $k \geq 0$;
- $w(k)$ and $v(k)$ are uncorrelated with $w(k')$ and $v(k')$ for every nonnegative integer $k' \neq k$;
- $\mathbf{E}(x(0)) = 0$ and $\mathbf{E}(w(k)) = \mathbf{E}(v(k)) = 0$ for each $k = 0, 1, \cdots$;
- $\mathbf{Var}(x(0)) > 0$, $\mathbf{Var}(w(k)) > 0$, and $\mathbf{Var}(v(k)) = \sigma^2 I$ with $\sigma > 0$ for every nonnegative integer k.

Using the same arguments as in the derivation of Eq. (3.6), we can show straightforwardly from Eqs. (11.69) and (11.70) that, for arbitrary nonnegative integers k and k' with $0 \leq k' \leq k$, the following equality is valid:

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(k) \\ x(k') \end{bmatrix} = \begin{bmatrix} \Gamma(k) \\ L(k') \end{bmatrix} \begin{bmatrix} x(0) \\ w(0) \\ w(1) \\ w(2) \\ \vdots \\ w(k-1) \end{bmatrix} + \begin{bmatrix} v(0) \\ v(1) \\ v(2) \\ \vdots \\ v(k) \\ 0 \end{bmatrix}, \tag{11.71}$$

where

$$\Gamma(k) = \begin{bmatrix} C_0 & 0 & 0 & 0 & \cdots & 0 \\ C(1)A(0) & C(1) & 0 & 0 & \cdots & 0 \\ C(2)A(1)A(0) & C(2)A(1) & C(2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C(k)\prod_{j=k-1}^{0} A(j) & C(k)\prod_{j=k-2}^{0} A(j) & C(k)\prod_{j=k-3}^{0} A(j) & C(k)\prod_{j=k-4}^{0} A(j) & \cdots & C(k) \end{bmatrix},$$

$$L(k') = \begin{bmatrix} \prod_{j=k'-1}^{0} A(j) & \prod_{j=k'-2}^{0} A(j) & \cdots & I & 0 & \cdots & 0 \end{bmatrix}.$$

If $x(0)$, $w(j)|_{j=0}^{k-1}$, and $v(k)|_{j=0}^{k}$ are normally distributed, then we can declare from Eq. (11.71) that

$$\left[y^T(0) \;\; y^T(1) \;\; y^T(2) \;\; \cdots \;\; y^T(k) \;\; x^T(k') \right]^T$$

is also normally distributed, which is equivalent to that

$$x(k') \quad \text{and} \quad \left[y^T(0) \;\; y^T(1) \;\; y^T(2) \;\; \cdots \;\; y^T(k) \right]^T$$

are jointly normally distributed.

For simplicity, denote the vectors

$$\left[y^T(0) \;\; y^T(1) \;\; \cdots \;\; y^T(k) \right]^T, \quad \left[x^T(0) \;\; w^T(0) \;\; w^T(1) \;\; \cdots \;\; w^T(k-1) \right]^T, \quad \text{and} \quad \left[v^T(0) \;\; v^T(1) \;\; \cdots \;\; v^T(k) \right]^T$$

respectively by $Y(k)$, $Z(k-1)$, and $V(k)$. Then Eq. (11.71) can be rewritten as follows:

$$\begin{bmatrix} Y(k) \\ x(k') \end{bmatrix} = \begin{bmatrix} \Gamma(k) & I \\ L(k') & 0 \end{bmatrix} \begin{bmatrix} Z(k-1) \\ V(k) \end{bmatrix}. \tag{11.72}$$

In addition, from the system output vector measurements $y(j)$, $j=0,1,\cdots,k$, the best estimate of the system state vector $x(k')$ under the criterion of mean squares errors, denoted $\hat{x}(k')$, has been proven to have the following closed-form expression [18,19,22]:

$$\hat{x}(k') = \mathbf{E}\left(x(k') \mid Y(k) \right), \tag{11.73}$$

that is,

$$\hat{x}(k') = \arg\min_{\alpha} \mathbf{E}\left(\left[\alpha - x(k')\right]^T \left[\alpha - x(k')\right]\right) \tag{11.74}$$

under the constraints of Eq. (11.71).

Based on this relation and the adopted assumptions on the external disturbance process $w(k)|_{k=0}^{\infty}$ and the measurement error process $v(k)|_{k=0}^{\infty}$, direct algebraic manipulations show that

$$\begin{aligned} &\begin{bmatrix} Y(k) \\ x(k') \end{bmatrix} \\ &\quad \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Gamma(k)\mathbf{Var}(Z(k-1))\Gamma^T(k) + \sigma^2 I & \Gamma(k)\mathbf{Var}(Z(k-1))L^T(k') \\ L(k')\mathbf{Var}(Z(k-1))\Gamma^T(k) & L(k)\mathbf{Var}(Z(k-1))L^T(k') \end{bmatrix} \right). \end{aligned} \tag{11.75}$$

Concerning joint normal distributions, the following result is well known [23,24].

Lemma 11.6. *Assume that random vectors x and y are jointly normally distributed. Moreover, assume that*

$$\mathbf{E}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \qquad \mathbf{Var}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} V_x & V_{yx}^T \\ V_{yx} & V_y \end{bmatrix}.$$

Then the conditional random vector $y|x$ is also normally distributed. In addition,

$$\mathbf{E}\,(y|x) = \mu_y + V_{yx}V_x^{-1}[x - \mu_x], \qquad \mathbf{Var}\,(y|x) = V_y - V_{yx}V_x^{-1}V_{yx}^T.$$

On the basis of Eqs. (11.73) and (11.72) and Lemma 11.6, the following equality can be straightforwardly established:

$$\begin{aligned} \hat{x}(k') &= \mathbf{E}\left(x(k') \mid Y(k) \right) \\ &= L(k')\mathbf{Var}(Z(k-1))\Gamma^T(k)\left[\Gamma(k)\mathbf{Var}(Z(k-1))\Gamma^T(k) + \sigma^2 I\right]^{-1} Y(k). \end{aligned} \tag{11.76}$$

Recall that when a linear space is constituted from n-dimensional random vectors with zero mean and finite covariance matrix, an inner product can be defined as

$$< x,\ y > = \mathbf{E}(x^T y).$$

Obviously, the norm induced from this inner product is consistent with the cost function adopted in the optimization problem of Eq. (11.74), that is, the mean squares errors. Hence, from the projection-based optimization condition [25,26] we have that

$$< x(k') - \hat{x}(k'),\ \hat{x}(k') > = \mathbf{E}\left(\left[x(k') - \hat{x}(k')\right]^T \hat{x}(k')\right) = 0. \tag{11.77}$$

Denote the estimation error vector $x(k') - \hat{x}(k')$ by $\tilde{x}(k')$. From this equality and Eq. (11.75) direct algebraic operations show that

$$\begin{aligned}
\mathbf{Var}\left(\tilde{x}(k')\right) &= \mathbf{E}\left(\left[x(k') - \hat{x}(k')\right]\left[x(k') - \hat{x}(k')\right]^T\right) \\
&= L(k')\mathbf{Var}(Z(k-1))L^T(k') \\
&\quad - L(k')\mathbf{Var}(Z(k-1))\Gamma^T(k)\left[\Gamma(k)\mathbf{Var}(Z(k-1))\Gamma^T(k) + \sigma^2 I\right]^{-1} \\
&\quad \times \Gamma(k)\mathbf{Var}(Z(k-1))L^T(k') \\
&= L(k')\left[\mathbf{Var}^{-1}(Z(k-1)) + \frac{1}{\sigma^2}\Gamma^T(k)\Gamma(k)\right]^{-1} L^T(k').
\end{aligned} \tag{11.78}$$

A recursive realization of Eqs. (11.76) and (11.78) is the well-known Kalman filter [18,19]. Rather than the recursive formula of the Kalman filter, which is given by Eqs. (4.8) and (4.9), expression (11.78) for the covariance matrix of estimation errors appears to be much more convenient in the analysis of the relations among sensor positions, sensor number, measurement data length, and estimation accuracy.

Define the random variable

$$\delta(k') = \tilde{x}^T(k')\mathbf{Var}^{-1}\left(\tilde{x}(k')\right)\tilde{x}(k') \tag{11.79}$$

Recall that the random vector $x(k')$ is normally distributed. On the other hand, it is clear from Eqs. (11.72) and (11.76) that the random vector $\hat{x}(k')$ is also normally distributed. We can therefore declare that the estimation error vector $\tilde{x}(k')$ is also normally distributed. In addition, note that $\mathbf{E}\left[\mathbf{E}\left(y|x\right)\right] = \mathbf{E}(y)$ for arbitrary random vectors x and y [23,24]. We therefore have

$$\mathbf{E}(\tilde{x}(k')) = \mathbf{E}(x(k') - \hat{x}(k')) = \mathbf{E}(x(k')) - \mathbf{E}\left(\mathbf{E}\left(\ x(k')\ \middle|\ Y(k)\ \right)\right) = 0. \tag{11.80}$$

Hence the random variable $\delta(k')$ has the χ_n^2 distribution, that is, the χ^2 distribution with n degrees of freedom [23,24]. This property is quite helpful in developing algorithms for attack detections and attack identifications [6,7].

It is also worth mentioning that

$$\begin{aligned} \mathbf{E}\left\{\left[x(k')-\hat{x}(k')\right]^T\left[x(k')-\hat{x}(k')\right]\right\} &= \mathbf{E}\left\{\mathrm{tr}\left(\left[x(k')-\hat{x}(k')\right]\left[x(k')-\hat{x}(k')\right]^T\right)\right\} \\ &= \mathrm{tr}\left\{\mathbf{E}\left(\left[x(k')-\hat{x}(k')\right]\left[x(k')-\hat{x}(k')\right]^T\right)\right\} \\ &= \mathrm{tr}\left\{\mathbf{Var}\left(\tilde{x}(k')\right)\right\}. \end{aligned} \tag{11.81}$$

These observations mean that the covariance matrix $\mathbf{Var}\left(\tilde{x}(k')\right)$ of the estimation error is closely related to both some statistics adopted in attack detections/identifications and the estimation accuracy measured by the mathematical expectation of squared estimation errors.

11.6.2 Sensor Placements

In both attack detections and attack identifications, it is essential to distinguish an abnormal situation of a networked system from its normal situation as fast as possible. This means that it is preferable to have sensors in positions that can make a detector achieve the minimal value of a metric on its residual signal when there do not exist any attacks in the associated networked system and reach a value of that metric as large as possible if the associated networked system has been attacked. Keeping in mind relations among state observers and detectors, in this subsection, we investigate sensor placements with the objective of minimizing state estimation errors.

For this purpose, we adopt the following assumption for the system described by Eqs. (11.69) and (11.70) to reflect characteristics in optimizations of sensor locations.

Assumption 11.1. *The matrix $C(k)$ is time invariant. Moreover, each of its rows has a nonzero element equal to* 1. *Furthermore, each of its columns has at most one nonzero element.*

This assumption reflects that all sensor positions have been fixed, each sensor only directly measures one of the states of the networked system, and if a state in the networked system is directly measured, then it can only be measured by one sensor. Under such an assumption, we can directly prove that the number of the sensors in the networked system is equal to the number of nonzero elements in the matrix C, which is an abbreviation of the matrix $C(k)$ when Assumption 11.1 is satisfied.

Define the set

$$\mathcal{S} = \left\{ i \;\middle|\; \begin{array}{c} C = \left[c_{jl}|_{j=1,l=1}^{j=p,l=n} \right] \text{ satisfies Assumption 11.1, and} \\ \text{there exists } \; j \in \{1, 2, \cdots, p\} \;\; \text{such that } \; c_{ji} = 1 \end{array} \right\}.$$

It is clear that this set indicates all the states in the networked system described by Eqs. (11.69) and (11.70) for which there is a sensor measuring its value. With a little abuse of terminology, in this subsection, we call this set the sensor set for that networked system.

For an arbitrary $t \times s$-dimensional matrix Ξ, let $(\Xi)_{ij}$ represent its ith row jth column element, $i = 1, 2, \cdots, t$ and $j = 1, 2, \cdots, s$. Define scalars $\mu(k)$, $v_w(k)$, v_0, and v_{0in} respectively as

$$\mu(k) = \max_{0 \le i \le k} \sigma_{\max}(A(i)), \qquad v_w(k) = \max_{0 \le i \le k} \max_{1 \le j \le n} (\mathbf{Var}(w(i)))_{jj},$$
$$v_0 = \max_{1 \le j \le n} (\mathbf{Var}(x(0)))_{jj}, \qquad v_{0in} = \max_{1 \le j \le n} \left(\mathbf{Var}^{-1}(x(0)) \right)_{jj}.$$

On the basis of the results given in the previous subsection on the Kalman filtering, the following results have been obtained in [15], which establish some relations among the number of sensors, data length, and state estimation accuracy in a networked system.

Theorem 11.6. *Let the networked system be described by Eqs. (11.69) and (11.70), and let $\hat{x}(k)$ stand for the estimate of its state vector $x(k)$ using the Kalman filter. Assume that $\sigma > 0$ and $\mu(k) \neq 1$. Then for every $k = 1, 2, \cdots$,*

$$\frac{n\sigma^2 \lambda_{\min}\left(L^T(k)L(k)\right)}{|\mathcal{S}| \frac{1-\mu^{2(k+1)}}{1-\mu^2} + \sigma^2 v_{0in}} \le \mathbf{E}\left(||x(k) - \hat{x}(k)||_2^2 \right) \le n(k+1)\lambda_{\max}\left(L^T(k)L(k)\right) \max\{v_0, v_w(k)\}, \tag{11.82}$$

where $|\mathcal{S}|$ is the number of elements in the sensor set $\mathcal{S}$.

This theorem reveals that the lower bound of state estimation accuracy decreases only inversely proportionally to the number of the sensors in the system and increases linearly with the increment of the number of the states in the system. This means that estimation errors cannot be significantly reduced through only adding sensors to the system, and for a large-scale system, measurements with a short data length cannot lead to an acceptable estimation accuracy in general.

From this theorem conditions can be derived on the number of sensors and data length required to meet some prescribed estimation accuracy. More precisely, assume that it is required that $\mathbf{E}\left(||x(k) - \hat{x}(k)||_2^2 \right) \le \alpha$ with some prescribed positive number α. Then the following conclusions can be derived from Theorem 11.6.

- Assume that the date length k is fixed. Then, to satisfy $\mathbf{E}\left(\left|\left|x(k) - \hat{x}(k)\right|\right|_2^2\right) \le \alpha$, it is necessary that

$$|\mathcal{S}| \ge \left[\frac{n\sigma^2\lambda_{\min}\left(L^T(k)L(k)\right)}{\alpha} - \sigma^2 v_{0in}\right]\frac{1-\mu^2}{1-\mu^{2(k+1)}}$$

- Assume that the number of sensors $|\mathcal{S}|$, is fixed. Then, to satisfy $\mathbf{E}\left(\left|\left|x(k) - \hat{x}(k)\right|\right|_2^2\right) \le \alpha$, it is necessary that

$$k \ge \frac{\log\left\{1 - \left[\frac{n\sigma^2\lambda_{\min}\left(L^T(k)L(k)\right)}{\alpha} - \sigma^2 v_{0in}\right]\frac{1-\mu^2}{|\mathcal{S}|}\right\}}{2\log(\mu)} - 1.$$

These relations imply that the required number of sensors increases linearly with the increment of the number of the system states, but the required data length increases only logarithmically.

Now, we investigate relations between sensor sets and the covariance matrix of estimation errors, that is, $\mathbf{Var}\,(\tilde{x}(k))$. For this purpose, define $L(j)$ with $j = 0$ as $L(0) = [I_n \;\; 0 \;\; \cdots \;\; 0]$. Then from the definitions of the matrices $\Gamma(k)$ and $L(k)$ it is obvious that

$$\Gamma(k) = \mathbf{col}\{CL(j)|_{j=0}^{k}\}. \tag{11.83}$$

Let e_j denote the jth canonical basis of the n-dimensional Euclidean space $\mathcal{R}^n$, $j = 1, 2, \cdots, n$, that is, e_j is an n-dimensional column vector with its jth row element 1 and all other elements 0. Moreover, for each $i = 1, 2, \cdots, n$, define the matrix

$$I(i) = e_i e_i^T.$$

Furthermore, let $s(i)$ with $i = 1, 2, \cdots, n$, be the indicator function defined as

$$s(i) = \begin{cases} 1 & \text{if the } i\text{th state of the system is directly measured by a sensor,} \\ 0 & \text{if the } i\text{th state of the system is not directly measured by any sensor.} \end{cases}$$

By Assumption 11.1 the output matrix C of the networked system described by Eqs. (11.69) and (11.70) can be expressed as

$$C = \mathbf{col}\{e_{j(i)}^T|_{i=1}^{p}\}, \tag{11.84}$$

where $j(i)$ is the position of the nonzero element in the ith row of the matrix C, $i = 1, 2, \cdots, p$. Moreover, the following equalities are valid:

$$C^T C = \sum_{m=1}^{p} e_{j(m)} e_{j(m)}^T = \sum_{m\in\mathcal{S}} I(m) = \sum_{m=1}^{n} s(m)I(m). \tag{11.85}$$

To clarify dependence of the matrices $\Gamma(k)$ and $\mathbf{Var}(\tilde{x}(k))$ on the sensor set $\mathcal{S}$, we further reexpress them respectively as $\Gamma(k, \mathcal{S})$ and $\mathbf{Var}(\tilde{x}(k), \mathcal{S})$, where the sensor set $\mathcal{S}$ is explicitly included. By Eqs. (11.83), (11.84), and (11.85) we have that

$$\begin{aligned}\Gamma^T(k,\mathcal{S})\Gamma(k,\mathcal{S}) &= \sum_{j=0}^{k} L^T(j)C^TCL(j)\\ &= \sum_{j=0}^{k}\left(L^T(j)\sum_{m=1}^{n} s(m)I(m)L(j)\right)\\ &= \sum_{m=1}^{n} s(m)\left(\sum_{j=0}^{k} L^T(j)I(m)L(j)\right)\\ &= \sum_{m\in\mathcal{S}} \mathbf{O}(k,m), \qquad (11.86)\end{aligned}$$

where

$$\mathbf{O}(k,m) = \sum_{j=0}^{k} L^T(j)I(m)L(j) = \sum_{j=0}^{k}(e_m^T L(j))^T(e^T(m)L(j)),$$
$$k = 0, 1, \cdots, \quad m = 1, 2, \cdots, n.$$

From its definition it is clear that for all $k = 0, 1, \cdots$ and $m = 1, 2, \cdots, n$, the matrix $\mathbf{O}(k, m)$ is independent of either the number of sensors in the system or the positions of the sensors. In addition, this matrix is at least semipositive definite.

Now, assume that there are two sensor sets $\mathcal{S}_1$ and $\mathcal{S}_2$ satisfying

$$\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \{1, 2, \cdots, n\}.$$

Then by Eq. (11.86) the following inequality is immediate:

$$\begin{aligned}\Gamma^T(k,\mathcal{S}_2)\Gamma(k,\mathcal{S}_2) &= \sum_{m\in\mathcal{S}_2} \mathbf{O}(k,m)\\ &= \sum_{m\in\mathcal{S}_1} \mathbf{O}(k,m) + \sum_{m\in\mathcal{S}_2\backslash\mathcal{S}_1} \mathbf{O}(k,m)\\ &\geq \sum_{m\in\mathcal{S}_1} \mathbf{O}(k,m)\\ &= \Gamma^T(k,\mathcal{S}_1)\Gamma(k,\mathcal{S}_1). \qquad (11.87)\end{aligned}$$

We can therefore declare from Lemma 2.1 that

$$\left[\frac{1}{\sigma^2}\Gamma^T(k,\mathcal{S}_2)\Gamma(k,\mathcal{S}_2)+\mathbf{Var}^{-1}(Z(k-1))\right]^{-1}$$
$$\leq\left[\frac{1}{\sigma^2}\Gamma^T(k,\mathcal{S}_1)\Gamma(k,\mathcal{S}_1)+\mathbf{Var}^{-1}(Z(k-1))\right]^{-1}. \tag{11.88}$$

By Eqs. (11.78) and (11.88) and by Lemma 2.1 we further have that

$$\mathbf{Var}(\tilde{x}(k),\mathcal{S}_2)\leq\mathbf{Var}(\tilde{x}(k),\mathcal{S}_1) \tag{11.89}$$

Eqs. (11.89) and (11.81) mean that, with the addition of some sensors, both the mean squares estimation errors of the Kalman filter and the covariance matrix of its estimation errors monotonically decrease. This is consistent with engineering intuition, as a sensor addition results in more measurement data, which provide more information about the plant state vector and therefore lead to a more accurate state estimate.

However, when the number of sensors is fixed, it is still not very easy to find the optimal sensor positions that minimize either the mean squares estimation errors or the covariance matrix of the state estimation errors [15,20,21].

On the other hand, discussions in the previous section show that observability of a networked system is necessary for the existence of a residual filter that is able to detect/identify an attack. It can be shown, however, that even the problem of finding the best sensor positions that lead to an observable networked system is NP-hard.

In particular, similarly to the proof of Theorem 3.2, we can also show that to recover the state vectors of the networked system described by Eqs. (11.69) and (11.70) from its output measurements $y(j)|_{j=0}^{k}$, it is necessary and sufficient that the following matrix $\mathcal{O}(k,\mathcal{S})$ w is of full column rank:

$$\mathcal{O}(k,\mathcal{S})=\left[\begin{array}{cccccc}C^T & A^T(0)C^T & (A(1)A(0))^TC^T & \cdots & \left(\prod_{j=k}^{0}A(j)\right)^T & C^T\end{array}\right]^T.$$

Note that the matrix $\mathcal{O}(k,\mathcal{S})$ is of full column rank if and only if the matrix $\mathcal{O}^T(k,\mathcal{S})\mathcal{O}(k,\mathcal{S})$ is of full rank. In fact, the ranks of these two matrices equal each other [27,28]. When sensor positions are selected to guarantee the observability of a networked system, it is reasonable from these observations to investigate the problem of maximizing the rank of the matrix $\mathcal{O}^T(k,\mathcal{S})\mathcal{O}(k,\mathcal{S})$ under the restrictions that $\mathcal{S}\subseteq\{1,2,\cdots,n\}$ and $|\mathcal{S}|\leq q$.

Concerning the set function involved in this optimization problem, the following conclusions are obtained.

Theorem 11.7. *Assume that the dynamics of a networked system is described by Eqs. (11.69) and (11.70). Moreover, assume that Assumption 11.1 is satisfied by its output matrix $C(k)$. Define the set function $f(\mathcal{S}) : 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ as*

$$f(k, \mathcal{S}) = \mathrm{rank}(\mathcal{O}^T(k, \mathcal{S})\mathcal{O}(k, \mathcal{S})), \tag{11.90}$$

where $\mathcal{V} = \{1, 2, \cdots, n\}$. Then this set function is submodular and increasing for each $k =$ $0, 1, 2, \cdots$.

The proof of this theorem is deferred to the appendix of this chapter.

From this theorem and the results on optimization of the set function given in Chapter 2 we can declare that optimal sensor placements for maximizing the rank of the observability matrix $\mathcal{O}(k, \mathcal{S})$ are in general NP-hard. On the other hand, the greedy heuristic method of that chapter usually gives a good approximation result. But it is worth noting that when the scale of a networked system is large, some computation issues may still arise for this greedy heuristic method, noting that in this case, the matrix $\bar{\mathbf{O}}(k, i)$ of Eq. (11.A.1) often has a high dimension. These are different significantly by the results of Chapter 3, given by Corollary 3.1 and Theorem 3.10. The output matrix there is allowed to take an arbitrary real value for a networked system, and an explicit parameterization is given for all output matrices that lead to an observable system.

When the state transition matrix of the networked system described by Eqs. (11.69) and (11.70) is also time invariant, arguments similar to those of Chapter 3 show that when the observability matrix $\mathcal{O}(k, \mathcal{S})$ is involved in an optimization problem, it is sufficient to only consider the case $k = n - 1$.

11.6.3 Actuator Placements

In the previous sections, it has been argued that an attacker usually intends to give destructive damages to a system in a stealthy way with small efforts. In this subsection, we investigate selection of input positions for the system described by Eqs. (11.69) and (11.70) such that some metrics on the input efforts can be minimized. For this purpose, we further adopt the following assumption, which is in a dual form of Assumption 11.1 adopted in the previous subsection for studying sensor placements.

Assumption 11.2. *The matrix $B(k)$ is time invariant. Moreover, each of its columns has a nonzero element* 1*. Furthermore, each of its rows has at most one nonzero element.*

When this assumption is satisfied, the input matrix $B(k)$ is abbreviated as B for simplicity.

Concerning the system described by Eqs. (11.69) and (11.70), its controllability Gramian at the time instant k, denoted $W_c(k)$, is defined as

$$W_c(k) = \sum_{j=0}^{k} \Phi(j,0) B B^T \Phi^T(j,0), \tag{11.91}$$

where $\Phi(j,0) = \prod_{i=j}^{0} A(i)$. Note that inputs to a system are usually energy restricted. Moreover, some functions of the controllability Gramian $W_c(k)$ have been argued to be nice quantities for measuring energies required to maneuver the system state vector from the zero-valued initial conditions to a prescribed value in the time interval $[0,\ k]$. Depending on whether the average energy required to maneuver the system state vector, a worst-case energy required to maneuver the system state vector or the volume of the states that can be reached through one unit or less of input energies, $\mathrm{tr}(W_c^{-1}(k))$, $\sigma_{\max}(W_c^{-1}(k))$ and $\log\det(W_c(k))$ are utilized respectively [16,17].

For each $i = 1, 2, \cdots, n$, let $a(i)$ with $i = 1, 2, \cdots, n$, be an indicator function defined as

$$a(i) = \begin{cases} 1 & \text{if the } i\text{th state of the system is directly maneuvered by an actuator,} \\ 0 & \text{if the } i\text{th state of the system is not directly maneuvered by any actuator.} \end{cases}$$

By Assumption 11.2 the input matrix B of the networked system can be expressed as

$$B = \left[\, e_{j(1)},\ e_{j(2)},\ \cdots,\ e_{j(q)} \,\right], \tag{11.92}$$

where $j(i)$ is the position of the nonzero element in the ith column of the input matrix B, $i = 1, 2, \cdots, q$. Moreover, similarly to Eq. (11.85), we can obtain the following relation:

$$B B^T = \sum_{m=1}^{n} a(m) I(m), \tag{11.93}$$

where the matrix $I(m)$ is defined as that in the previous subsection.

Define the set

$$\mathcal{A} = \left\{ i \;\middle|\; \begin{array}{c} B = \left[b_{jl}|_{j=1,l=1}^{j=n,l=q} \right] \text{ satisfies Assumption 11.2, and} \\ \text{there exists } \ j \in \{1, 2, \cdots, q\} \ \text{ such that } \ b_{ij} = 1 \end{array} \right\}.$$

Then it is clear that this set indicates all the states in the networked system described by Eqs. (11.69) and (11.70) that are directly controlled by an actuator. Similarly to the sensor set $\mathcal{S}$, in this subsection, we call this set the actuator set for that networked system.

To clarify dependence of the controllability Gramian $W_c(k)$ on the actuator set $\mathcal{A}$, we further reexpress it as $W_c(k, \mathcal{A})$, so that the sensor set $\mathcal{S}$ is explicitly included. Using completely the same arguments as those in the derivation of Eq. (11.86), from Eqs. (11.91) and (11.93) we obtain the following relation:

$$W_c(k, \mathcal{A}) = \sum_{m\in\mathcal{A}} a(m)\mathbf{C}(k, m), \tag{11.94}$$

where

$$\mathbf{C}(k, m) = \sum_{j=0}^{k} \Phi(j, 0)I(m)\Phi^T(j, 0), \qquad k = 0, 1, \cdots, \quad m = 1, 2, \cdots, n.$$

It is clear from its definition that, for all $k = 0, 1, \cdots$ and $m = 1, 2, \cdots, n$, the matrix $\mathbf{C}(k, m)$ is independent of either the number of actuators in the system or the positions of the actuators. In addition, this matrix is at least semipositive definite.

Now, assume that there are two actuator sets $\mathcal{A}_1$ and $\mathcal{A}_2$ satisfying

$$\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \{1, 2, \cdots, n\}.$$

Then by Eq. (11.94) we can obtain the following inequality by the same arguments as those in the derivation of Eq. (11.87):

$$W_c(k, \mathcal{A}_2) \geq W_c(k, \mathcal{A}_1). \tag{11.95}$$

We can therefore declare from Lemma 2.1 that when both matrices $W_c(k, \mathcal{A}_1)$ and $W_c(k, \mathcal{A}_2)$ are invertible, the following inequalities are valid:

$$\mathrm{tr}(W_c^{-1}(k, \mathcal{A}_2)) \leq \mathrm{tr}(W_c^{-1}(k, \mathcal{A}_1)), \tag{11.96}$$

$$\sigma_{\max}(W_c^{-1}(k, \mathcal{A}_2)) \leq \sigma_{\max}(W_c^{-1}(k, \mathcal{A}_1)), \tag{11.97}$$

$$\log\det(W_c(k, \mathcal{A}_2)) \geq \log\det(W_c(k, \mathcal{A}_1)), \tag{11.98}$$

that is, through adding some other actuators, both the average energy required to maneuver the system state vector and the worst-case energy required to maneuver the system state vector decrease monotonically, whereas the volume of the states that can be reached through one unit or less of input energies increases monotonically. Once again, this is in a good agreement with engineering intuition.

However, optimization of these metrics is generally mathematically difficult. In particular, we have the following results.

Theorem 11.8. *Let the networked system be described by Eqs. (11.69) and (11.70). Define the set function $f(\mathcal{A}) : 2^{\mathcal{V}} \longrightarrow \mathcal{R}$ as*

$$f(\mathcal{A}) = \log\det(W_c(k, \mathcal{A})), \tag{11.99}$$

where $\mathcal{V} = \{1, 2, \cdots, n\}$. Then this set function is submodular and increasing.

Proof. From Eq. (11.98) we can directly declare that this function is increasing.

To prove its submodularity, assume that ξ is an arbitrary element of the set $\mathcal{V}$. Moreover, assume that $\mathcal{A}_1$ and $\mathcal{A}_2$ are two sets satisfying $\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \mathcal{V}\backslash\{\xi\}$. Define the matrix $W_c(\gamma)$ and the function $g(\gamma)$ respectively as

$$\begin{aligned} W_c(\gamma) &= W_c(k, \mathcal{A}_1) + \gamma\left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right], \\ g(\gamma) &= \log\det\left[W_c(\gamma) + W_c(k, \{\xi\})\right] - \log\det(W_c(\gamma)), \end{aligned}$$

where $\gamma \in [0,\ 1]$.

By Eq. (11.95) we have that

$$W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1) \geq 0. \tag{11.100}$$

On the other hand, noting that by its definition the matrix $W_c(k, \{\xi\})$ is at least positive semidefinite, it is obvious that

$$W_c(\gamma) + W_c(k, \{\xi\}) \geq W_c(\gamma). \tag{11.101}$$

We can therefore declare from Lemma 2.1 that

$$\begin{aligned} \frac{dg(\gamma)}{d\gamma} &= \frac{d}{d\gamma}\log\det(W_c(\gamma) + W_c(k, \{\xi\})) - \frac{d}{d\gamma}\log\det(W_c(\gamma)) \\ &= \operatorname{tr}\left\{\left[W_c(\gamma) + W_c(k, \{\xi\})\right]^{-1}\left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right]\right\} \\ &\quad - \operatorname{tr}\left\{W_c^{-1}(\gamma)\left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right]\right\} \\ &= \operatorname{tr}\left\{\left[(W_c(\gamma) + W_c(k, \{\xi\}))^{-1} - W_c^{-1}(\gamma)\right]\left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right]\right\} \\ &= \operatorname{tr}\left\{\left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right]^{\frac{1}{2}}\left[(W_c(\gamma) + W_c(k, \{\xi\}))^{-1} - W_c^{-1}(\gamma)\right]\right. \\ &\quad \left. \times \left[W_c(k, \mathcal{A}_2) - W_c(k, \mathcal{A}_1)\right]^{\frac{1}{2}}\right\} \\ &\leq 0, \end{aligned} \tag{11.102}$$

which implies that $g(0) \geq g(1)$.[2]

On the other hand, from the definition of the function $g(\gamma)$ it is obvious that

$$g(0) = \log\det(W_c(k, \mathcal{A}_1) + W_c(k, \{\xi\})) - \log\det(W_c(k, \mathcal{A}_1)), \tag{11.103}$$

$$g(1) = \log\det(W_c(k, \mathcal{A}_2) + W_c(k, \{\xi\})) - \log\det(W_c(k, \mathcal{A}_2)). \tag{11.104}$$

Note that from Eq. (11.94) we immediately have that

$$W_c\left(k, \mathcal{A}\bigcup\{\xi\}\right) = W_c(k, \mathcal{A}) + W_c(k, \{\xi\})$$

for all $\mathcal{A} \subseteq \mathcal{V}\backslash\{\xi\}$ and $\xi \in \mathcal{V}$. Hence we can claim that $g(0)$ and $g(1)$ are in fact equal to the values of the following set function $f_\xi(\mathcal{A})$ respectively at $\mathcal{A} = \mathcal{A}_1$ and $\mathcal{A} = \mathcal{A}_2$:

$$f_\xi(\mathcal{A}) = \log\det\left[W_c\left(k, \mathcal{A}\bigcup\{\xi\}\right)\right] - \log\det(W_c(k, \mathcal{A})), \qquad \mathcal{A} \subseteq \mathcal{V}\backslash\{\xi\}.$$

The proof can now be completed by an application of Lemma 2.8. □

In addition to the set function logdet$(W_c(k, \mathcal{A}))$, it has also been proven in [16,17] that some other set functions, such as rank$(W_c(k, \mathcal{A}))$, $-$tr$(W_c^{-1}(k, \mathcal{A}))$, etc., are also monotone increasing and submodular. This means that maximization of these set functions is in general NP-hard, but the greedy heuristic algorithm described in Chapter 2 usually works well approximately. Once again, as in the sensor placement problem, computation issues may arise for a large-scale system in the application of that greedy heuristic algorithm, noting that the associated matrix in this case still has a great dimension, which is not very convenient for numerical computations.

On the other hand, from the duality between controllability and observability of a system, similiar results can be established for sensor placements using the associated observability Gramian.

In the above investigations on sensor/actuator placements, the output/input matrix of the system is restricted by Assumption 11.1/11.2. The associated results can be extended to a more general situation in which each row of the output matrix can take values from a set of some candidate rows, or each column of the input matrix can take values from a set of some candidate columns. The major required modification in these extensions is a redefinition of the matrix $I(i)$ adopted in the aforementioned investigations. Details can be found in the related literature, such as [16,17] and the references therein.

[2] In the above derivations, we applied the formula $\frac{d}{d\gamma}\log\det(X(\gamma)) = \mathrm{tr}\left\{X^{-1}(\gamma)\frac{dX(\gamma)}{d\gamma}\right\}$. To guarantee the validness of this formula, it is necessary that the matrix $X(\gamma)$ is invertible [27,29]. When this condition is not satisfied, replace the matrix $W_c(k, \mathcal{A})$ with the matrix $W_c(k, \mathcal{A}) + \varepsilon I$ in the definition of the set function $f(\mathcal{A})$, where ε is an arbitrary positive number. Then, the matrix $W_c(k, \mathcal{A}) + \varepsilon I$ is positive definite and therefore invertible. The conclusions of Theorem 11.8 can be obtained by letting ε approach zero.

11.7 Concluding Remarks

Importance of safety is now widely recognized for a networked system, especially with the ambitious introduction of public communication channels into a control system. In this chapter, we investigated some basic issues in system designs closely related to attack detection and identification, which reveals some relations among attack detectability/identifiability, system controllability/observability, and actuator/sensor placement. To develop an algorithm applicable to actual engineering problems, some other factors, such as robustness against modeling errors, stochastic properties of external disturbances, and measurement errors, must also be taken into account.

11.8 Bibliographic Notes

Network safety has been extensively investigated in power systems. [6] summarizes various important aspects of this problem and some interesting recent advancements. Recent interest in this problem are mainly stimulated by the prospective introduction of public communication channels into control systems, which brings some new characteristics into this field. Many essential issues are still in their primary stages toward a complete settlement. [2,7] are some recent surveys about this problem.

Basic concepts and results for combinatorial optimizations can be found in [26,30]. Concerning with convex optimization, [25,31] provides an excellent introduction on its essential motivations and results.

The notion of a system matrix is due to Rosenbrock, who also introduced zeros of linear multivariable systems in terms of the Smith–McMillan form of the system transfer matrix in [10,13]. The concept of weakly unobservable subspace appears to be firstly introduced for a discrete-time system by Silverman [32]. A summary and some detailed discussions can be found in [11]. Connections between the spectral properties of a weakly unobservable subspace and the system zeros are also dealt with in [33], and [34] has studied connections between the system zeros and the state space geometric structure of the system.

Appendix 11.A

11.A.1 Proof of Theorem 11.7

From the definition of the matrix $\mathcal{O}(k, \mathcal{S})$ and Eq. (11.85) it is obvious that

$$\begin{aligned}
\mathcal{O}^T(k,\mathcal{S})\mathcal{O}(k,\mathcal{S}) &= \left[C^T \;\; A(0)^T C^T \;\; (A(1)A(0))^T C^T \;\; \cdots \;\; \left(\prod_{j=k}^{0} A(j)\right)^T C^T\right] \\
&\quad \times \left[C^T \;\; A(0)^T C^T \;\; (A(1)A(0))^T C^T \;\; \cdots \;\; \left(\prod_{j=k}^{0} A(j)\right)^T C^T\right]^T \\
&= \sum_{l=0}^{k} \left\{ \left(\prod_{j=l}^{0} A(j)\right)^T C^T C \left(\prod_{j=l}^{0} A(j)\right) \right\} \\
&= \sum_{l=0}^{k} \left\{ \left(\prod_{j=l}^{0} A(j)\right)^T \left(\sum_{i \in \mathcal{S}} I(i)\right) \left(\prod_{j=l}^{0} A(j)\right) \right\} \\
&= \sum_{i \in \mathcal{S}} \sum_{l=0}^{k} \left\{ \left(\prod_{j=l}^{0} A(j)\right)^T I(i) \left(\prod_{j=l}^{0} A(j)\right) \right\} \\
&= \sum_{i \in \mathcal{S}} \bar{\mathbf{O}}(k,i), \qquad (11.A.1)
\end{aligned}$$

where

$$\bar{\mathbf{O}}(k,i) = \sum_{l=0}^{k} \left\{ \left(\prod_{j=l}^{0} A(j)\right)^T I(i) \left(\prod_{j=l}^{0} A(j)\right) \right\}, \qquad i = 1, 2, \cdots, n.$$

It is clear that, for each $i = 1, 2, \cdots, n$, the matrix $\bar{\mathbf{O}}(k,i)$ is at least positive semidefinite. Hence, for arbitrary sensor sets $\mathcal{S}_1$ and $\mathcal{S}_2$ satisfying $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \{1, 2, \cdots, n\}$, the following relations are valid:

$$\begin{aligned}
\mathcal{O}^T(k,\mathcal{S}_2)\mathcal{O}(k,\mathcal{S}_2) &= \sum_{m \in \mathcal{S}_2} \bar{\mathbf{O}}(k,m) \\
&= \sum_{m \in \mathcal{S}_1} \bar{\mathbf{O}}(k,m) + \sum_{m \in \mathcal{S}_2 \backslash \mathcal{S}_1} \bar{\mathbf{O}}(k,m) \\
&\geq \sum_{m \in \mathcal{S}_1} \bar{\mathbf{O}}(k,m) \\
&= \mathcal{O}^T(k,\mathcal{S}_1)\mathcal{O}(k,\mathcal{S}_1). \qquad (11.A.2)
\end{aligned}$$

Let $\lambda_i(\cdot)$ denote the ith eigenvalue of an $n \times n$-dimensional symmetric matrix that satisfies $\lambda_1(\cdot) \leq \lambda_2(\cdot) \leq \cdots \leq \lambda_n(\cdot)$. By Lemma 2.1 and the last equation we have that, for each $i = 1, 2, \cdots, n$,

$$\lambda_i \left\{\mathcal{O}^T(k, \mathcal{S}_2)\mathcal{O}(k, \mathcal{S}_2)\right\} \geq \lambda_i \left\{\mathcal{O}^T(k, \mathcal{S}_1)\mathcal{O}(k, \mathcal{S}_1)\right\}. \tag{11.A.3}$$

Note that, for a positive semidefinite matrix, its rank is equal to the number of its nonzero eigenvalues. Moreover, each of its eigenvalues is not smaller than 0. We can therefore declare from Eq. (11.A.3) that

$$\text{rank}\left\{\mathcal{O}^T(k, \mathcal{S}_2)\mathcal{O}(k, \mathcal{S}_2)\right\} \geq \text{rank}\left\{\mathcal{O}^T(k, \mathcal{S}_1)\mathcal{O}(k, \mathcal{S}_1)\right\}, \tag{11.A.4}$$

that is, the set function $f(k, \mathcal{S})$ of Eq. (11.90) is increasing.

To prove its submodularity, define the derived set function $f_i(k, \mathcal{S})$ for each $i \in \{1, 2, \cdots, n\}$ as

$$\begin{aligned} f_i(k, \mathcal{S}) = \text{rank}\left(\mathcal{O}^T\left(k, \mathcal{S}\bigcup\{i\}\right)\mathcal{O}\left(k, \mathcal{S}\bigcup\{i\}\right)\right) - \text{rank}(\mathcal{O}^T(k, \mathcal{S})\mathcal{O}(k, \mathcal{S})),\\ \mathcal{S} \subseteq \{1, 2, \cdots, n\}\backslash\{i\}. \end{aligned} \tag{11.A.5}$$

Note that, for arbitrary $n \times n$-dimensional real matrices X and Y,

$$\text{rank}(X + Y) = \text{rank}(X) + \text{rank}(Y) - \dim\left(\mathcal{S}pan(X)\bigcap\mathcal{S}pan(Y)\right),$$

where $\dim(\cdot)$ stands for the dimension of a subspace [27,28]. From this property, the definition of the set function $f_i(k, \mathcal{S})$, and from Eq. (11.A.1) we have that

$$\begin{aligned} f_i(k, \mathcal{S}) &= \text{rank}\left(\sum_{j\in\mathcal{S}\bigcup\{i\}} \bar{\mathbf{O}}(k, j)\right) - \text{rank}\left(\sum_{j\in\mathcal{S}} \bar{\mathbf{O}}(k, j)\right)\\ &= \text{rank}\left(\bar{\mathbf{O}}(k, i) + \sum_{j\in\mathcal{S}} \bar{\mathbf{O}}(k, j)\right) - \text{rank}\left(\sum_{j\in\mathcal{S}} \bar{\mathbf{O}}(k, j)\right)\\ &= \text{rank}\left(\bar{\mathbf{O}}(k, i)\right) - \dim\left(\mathcal{S}pan\left(\bar{\mathbf{O}}(k, i)\right)\bigcap\mathcal{S}pan\left(\sum_{j\in\mathcal{S}} \bar{\mathbf{O}}(k, j)\right)\right). \end{aligned} \tag{11.A.6}$$

Assume now that $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \{1, 2, \cdots, n\}\backslash\{i\}$. Then from the relation

$$\sum_{m\in\mathcal{S}_2} \bar{\mathbf{O}}(k, m) = \sum_{m\in\mathcal{S}_1} \bar{\mathbf{O}}(k, m) + \sum_{m\in\mathcal{S}_2\backslash\mathcal{S}_1} \bar{\mathbf{O}}(k, m)$$

and the definition of the matrix $\bar{\mathbf{O}}(k, m)$, $m = 1, 2, \cdots, n$, the following relation can be established by some algebraic manipulations:

$$\mathcal{S}pan\left(\sum_{j\in\mathcal{S}_2}\bar{\mathbf{O}}(k, j)\right) \supseteq \mathcal{S}pan\left(\sum_{j\in\mathcal{S}_1}\bar{\mathbf{O}}(k, j)\right) \tag{11.A.7}$$

Therefore

$$\begin{aligned}&\left\{\mathcal{S}pan\left(\bar{\mathbf{O}}(k, i)\right)\bigcap\mathcal{S}pan\left(\sum_{j\in\mathcal{S}_2}\bar{\mathbf{O}}(k, j)\right)\right\}\\ &\quad\supseteq\left\{\mathcal{S}pan\left(\bar{\mathbf{O}}(k, i)\right)\bigcap\mathcal{S}pan\left(\sum_{j\in\mathcal{S}_1}\bar{\mathbf{O}}(k, j)\right)\right\},\end{aligned} \tag{11.A.8}$$

which further leads the following inequality:

$$\begin{aligned}&\dim\left\{\mathcal{S}pan\left(\bar{\mathbf{O}}(k, i)\right)\bigcap\mathcal{S}pan\left(\sum_{j\in\mathcal{S}_2}\bar{\mathbf{O}}(k, j)\right)\right\}\\ &\quad\geq\dim\left\{\mathcal{S}pan\left(\bar{\mathbf{O}}(k, i)\right)\bigcap\mathcal{S}pan\left(\sum_{j\in\mathcal{S}_1}\bar{\mathbf{O}}(k, j)\right)\right\}.\end{aligned} \tag{11.A.9}$$

Substituting this inequality into Eq. (11.A.6), we have that, under the condition that $i \in \{1, 2, \cdots, n\}$ and $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \{1, 2, \cdots, n\}\backslash\{i\}$, the following inequality is true:

$$f_i(k, \mathcal{S}_2) \leq f_i(k, \mathcal{S}_1), \tag{11.A.10}$$

that is, this derived set function is decreasing.

We can therefore declare from Lemma 2.8 that the set function $f(k, \mathcal{S})$ is submodular. This completes the proof. □

References

[1] G. Richards, Hackers vs slackers, Engineering Technology 3 (2008).
[2] J.P. Farwell, R. Rohozinski, Stuxnet and the future of cyber war, Survival 53 (2011).
[3] J. Slay, M. Miller, Lessons learned from the Maroochy water breach, Critical Infrastructure Protection 253 (2007) 73–82.
[4] S. Kuvshinkova, SQL slammer worm: lessons learned for consideration by the electricity sector, North American Electric Reliability Council (2003).
[5] J.P. Conti, The day the samba stopped, Engineering Technology 5 (2010) 46–47.

[6] U. Hager, C. Rehtanz, N. Voropai (Eds.), Monitoring, Control and Protection of Interconnected Power Systems, Springer-Verlag, Berlin, Heidelberg, 2014.

[7] H. Sandberg, S. Amin, K.H. Johansson (Eds.), Special Issue on Cyberphysical Security, IEEE Control Systems Magazine 36 (2015) 20–127.

[8] F. Pasqualetti, F. Dörfler, F. Bullo, Attack detection and identification in cyber-physical systems, IEEE Transactions on Automatic Control 58 (2013) 2715–2729.

[9] H. Fawzi, P. Tabuada, S. Diggavi, Secure estimation and control for cyber-physical systems under adversarial attacks, IEEE Transactions on Automatic Control 59 (2014) 1454–1467.

[10] K.M. Zhou, J.C. Doyle, K. Glover, Robust and Optimal Control, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[11] H.L. Trentelman, A.A. Stoorvogel, M. Hautus, Control Theory for Linear Systems, Springer, London, UK, 2002.

[12] T. Geerts, Invariant subspaces and invertibility properties for singular systems: the general case, Linear Algebra and Its Applications 183 (1993) 61–88.

[13] H.H. Rosenbrock, State-Space and Multivariable Theory, John-Wiley, New York, USA, 1970.

[14] J.M. Dion, C. Commault, J. der Woude, Generic properties and control of linear structured systems: a survey, Automatica 39 (2003) 1125–1144.

[15] V. Tzoumas, A. Jadbabaie, G.J. Pappas, Sensor placement for optimal Kalman filtering: fundamental limits, submodularity, and algorithms, arXiv:1509.08146v3 [math.OC], 2016.

[16] T.H. Summers, F.L. Cortesi, J. Lygeros, On submodularity and controllability in complex dynamical networks, IEEE Transactions on Control of Network Systems 3 (2016) 91–101.

[17] V. Tzoumas, M.A. Rahimian, G.J. Pappas, A. Jadbabaie, Minimal actuator placement with bounds on control effort, arXiv:1409.3289v5 [math.OC], 2016.

[18] A.E. Bryson, Y.C. Ho, Applied Optimal Control: Optimization, Estimation and Control, Taylor & Francis, New York, USA, 1975.

[19] T. Kailath, A.H. Sayed, B. Hassibi, Linear Estimation, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[20] L.T. Ye, S. Roy, S. Sundaram, On the complexity and approximability of optimal sensor selection for Kalman filtering, arXiv:1711.01920v1 [math.OC], 2017.

[21] H.T. Zhang, R. Ayoub, S. Sundaram, Sensor selection for Kalman filtering of linear dynamic systems: complexity, limitations and greedy algorithms, Automatica 78 (2017) 202–210.

[22] D. Simon, Optimal State Prediction: Kalman, H_∞ and Nonlinear Approaches, Wiley-Interscience, A John Wiley & Sons, Inc., Publication, Hoboken, New Jersey, USA, 2006.

[23] Y.S. Chow, H. Teicher, Probability Theory: Independence, Interchangeability, Martingales, 3rd edition, Springer-Verlag, New York, USA, 1997.

[24] G.R. Grimmett, D.R. Stirzaker, Probability and Random Processes, Oxford University Press, USA, 2001.

[25] D.P. Bertsekas, Convex Optimization Theory, Athena Scientific, Boston, USA, 2009.

[26] L.A. Wolsey, Integer Programming, John Wiley & Sons, USA, 1988.

[27] F.Z. Zhang, Matrix Theory: Basic Results and Techniques, Springer, New York, 1999.

[28] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.

[29] P. Lancaster, M.T. Tismenetsky, The Theory of Matrices: With Applications, Academic, New York, 1985.

[30] L. Lovasz, Submodular functions and convexity, in: A. Bachem, M. Grotschel, B. Korte (Eds.), Mathematical Programming: The State of the Art, Springer, Bonn, Germany, 1983.

[31] J.B. Hiriart-Urruty, C. Lemarechal, Fundamentals of Convex Analysis, Springer, Berlin, Germany, 2001.

[32] L. Silverman, Discrete Riccati equations: alternative algorithms, asymptotic properties, and system theory interpretations, Control and Dynamical Systems 12 (1976) 313–386.

[33] W.M. Wonham, Linear Multivariable Control: A Geometric Approach, 3rd edition, Springer Verlag, New York, USA, 1985.

[34] A.S. Morse, Structural invariants of linear multivariable systems, SIAM Journal on Control and Optimization 11 (1973) 446–465.

CHAPTER 12

Some Related Issues

12.1 Introduction

Networked multiagent systems are a class of typical distributed systems that have attracted much attention due to the multidisciplinary nature among the areas of detecting technologies, communication networks, and computer engineering. Nowadays, multiagent systems and related technologies play an irreplaceable role in modern society, such as information war, smart home, cosmic discoveries, nanometer network, intelligent manufacturing, precision agriculture, and so on [1–3].

The concept of “agent” was first proposed by the computer science community. In the computer science and engineering, along with the scales of the problems, need to be solved become larger and larger, and the complexity becomes higher and higher, so the centralized computing system cannot fulfill the requirement for solving practical problems. Since the mid-1970s, distributed computing and distributed artificial intelligence have grown up rapidly. In the early study of DAI, the researchers mainly focused on the distributed problem solving, which divides a big problem into several subtasks, and each task is assigned with a subsystem to fulfill the task. In 1986, Minsky proposed the concept of agent in his book *Society of Mind*. An agent is an independent computing entity, which integrates the capacity of computing, decision-making, and communication. The agent is usually a piece of program or can be associated with a hardware entity, such as a robot. Agents cooperate with each other by exchanging information through mutual communication to accomplish group task and to optimize individual and group performances and costs. In the cooperation, the agents also compete for limited computation and communication resources. The multiagent systems can be viewed as a model for the human society.

In biological physics, the collective behaviors of biological group have drawn much attention from the scientists. In biological swarms, each individual also have integrated capacity of sensing, decision making, and communication. They aggregate as a whole group in lots of social behaviors such as foraging, mating, and escaping for maximizing interests of individual and the whole group.

The concept of multiagent systems also appears in the control engineering, and the cooperative control of multiagent systems has drawn much attention from the systems and control community. In recent years, along with the rapid development of advanced sensing, computing, and communication technologies, the architecture and operation mode of control systems

Estimation and Control of Large-Scale Networked Control Systems
DOI: 10.1016/B978-0-12-805311-9.00012-9

have been essentially changed. The units of control systems are not separate sensors, plants, controllers, and actuators but become independent entities with integrated sensing, computing, control, and communication. The agents collaborate and communicate with each other to achieve the goal of the whole network. In modern networked control systems, the control and decision-making process are performed in a distributed and hierarchical way. Communication and cooperation become crucial factors of control systems. On one hand, the research on multiagent systems has wide applications in cooperative information processing of sensor networks, multirobot cooperation, and formation control of UAVs. On the other hand, the research on multiagent systems provides a new angle of view for control systems. In the traditional control theory, a control system consists of plant, sensor, controller and actuator. The plant is controlled by the control signal passively. The information channel from the controller to the plant is called the control channel and from the controller to the plant is called the feedback channel. The controller and plant are in unequal status, whereas in the multiagent system theory, the system consists of agents on different levels. The agents on the same levels are in equal status. The information transmission from agent Bob to agent Alice is the control signal to Alice from Bob's angle of view, and there is also the information feedback from Alice's angle of view. Every agent is controlling others while is controlled by others. It depends on the angle of view whether an agent is a decision-maker or controlled object. Here the control and feedback are both parts of communication among agents. The communication network and control system are fully integrated together.

12.2 Cooperation Over Communications

12.2.1 Time Synchronization

Time synchronization, also called clock synchronization, is the premise for networked multiagent systems to achieve effective applications and services. Professor Lamport, who won the Turing Award in 1978, pointed out that clock synchronization is a critical foundation of distributed systems [4]. In networked systems, equipment localization, transmission scheduling, event sequencing, data fusion, and control actuation all require a synchronized clock with high accuracy. In the field of information science and engineering, clock synchronization means designing a synchronization algorithm to calibrate each local clock of electrical and electronics physical units such that different physical units finally have a common clock reference. Here local clock sources generally refer to oscillation circuit modules that produce stable pulse, such as crystal, digital-controlled oscillator, atomic clock, and so on [5].

Networked clock synchronization is a typical kind of state estimation problems over communication networks. Generally, the physical clock of each unit, that is, unit i, consists of an oscillator and a counter. The oscillator generates the standard unit of frequency f_i, which is a stochastic process. The counter keeps tracking of the number of oscillations occurred. After being set initially, the clock i provides its reading $c_i(t)$ of the natural time t by adding up the number of oscillations [6]:

$$c_i(t) = f_i(t - t_0) + c_i(t_0), \tag{12.1}$$

where t_0 is the initial time, and $c_i(t_0)$ is the hardware clock reading at t_0. In fact, Eq. (12.1) is a discrete-time version of the continuous integral model of physical clocks. Since f_i is slowly time varying, the estimated value $\hat{\tau}_i(t)$ of hardware clock reading can be given by

$$\hat{\tau}_i(t) = (c_i(t) - c_i(t_0))/\hat{f}_i + \tau_i(t_0), \tag{12.2}$$

where $\hat{f}_i$ is the estimation of nominal frequency, and $\tau_i(t_0)$ is the local clock reading at initial time. Notice that the oscillation frequency is almost fixed during a small period of time in stable physical–chemical environment.

Roughly speaking, there are two kinds of clock synchronization algorithms, hierarchical configuration and distributed configuration. In the former, there is a logic master–slave relationship, that is, the clock values of low-level equipments are endowed by high-level equipments. In the latter, the structure is totally distributed. For large-scale wireless sensor networks, distributed clock synchronization algorithms are of extremely important significance [7].

Recently, consensus-based distributed clock synchronization algorithms have been developed and studied extensively [8–15]. The objective of state consensus algorithms is to achieve the same global value of all nodes by local information exchange [16–20], which coincide with that of clock synchronization. An excellent clock synchronization algorithm should fit for the characteristics of wireless networks, such as fully distributed structure, limited power and memory, asynchronous information exchange, self-adapting, and easy implementation, and should be robust against random node/link failures and recreation and packet dropouts. Fortunately, there are systematic studies of these issues in multiagent state consensus. Especially, consensus-based clock synchronization algorithms have particular advantages. Firstly, the algorithm is totally distributed, independent of infrastructures for special network topologies, and much robust against link failures and time-varying topologies. Secondly, synchronization among neighboring nodes can get much higher precision with less computation overhead. Thirdly, the algorithm can uniformly compensate skew and offset asynchronously and relax the master–slave relationship. From framework of state consensus theory, the clock synchronization problem comes down to the high-order heterogeneous multiagent state consensus.

12.2.2 State Consensus

State consensus is one of the most fundamental problems of distributed coordination of multiagent networks, which, roughly speaking, means designing a network protocol such that as time goes on, all agents asymptotically reach an agreement. For some cases, the agreement is a common value, which may be the average of the initial states of the system and is often called average-consensus. Besides networked time synchronization, consensus and average-consensus have wide application background in the area such as formation control [21], distributed filtering [22], multisensor data fusion [23], and distributed computation [24].

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_{\mathcal{G}}, \mathcal{A}_{\mathcal{G}}\}$ be a weighted digraph, where $\mathcal{V} = \{1, 2, \cdots, N\}$ is the set of nodes with node i representing the ith agent, $\mathcal{E}_{\mathcal{G}}$ is the set of edges, and $\mathcal{A}_{\mathcal{G}} = [a_{ij}] \in \mathcal{R}^{N\times N}$ is the weighted adjacency matrix of $\mathcal{G}$. An edge in $\mathcal{G}$ is denoted by an ordered pair (j, i), and $(j, i) \in \mathcal{E}_{\mathcal{G}}$ if and only if the jth agent can directly send information to the ith agent. The neighborhood of the ith agent is denoted by $N_i = \{j \in \mathcal{V} \mid (j, i) \in \mathcal{E}_{\mathcal{G}}\}$. An element of N_i is called a neighbor of i. The ith node is called a source if it has no neighbors but is a neighbor of another node in $\mathcal{V}$. A node is called an isolated node if it has no neighbor and is not a neighbor of any other node. For any $i, j \in \mathcal{V}$, $a_{ij} \geq 0$, and $a_{ij} > 0 \Leftrightarrow j \in N_i$. The in-degree of i is defined as $deg_{in}(i) = \sum_{j=1}^{N} a_{ij}$, and the out-degree of i is defined as $deg_{out}(i) = \sum_{j=1}^{N} a_{ji}$. The Laplacian matrix of $\mathcal{G}$ is defined as $L_{\mathcal{G}} = \mathcal{D}_{\mathcal{G}} - \mathcal{A}_{\mathcal{G}}$, where $\mathcal{D}_{\mathcal{G}} = diag(deg_{in}(1), \cdots, deg_{in}(N))$. $\mathcal{G}$ is called a balanced digraph if $deg_{in}(i) = deg_{out}(i)$, $i = 1, 2, ..., N$. $\mathcal{G}$ is called an undirected graph if $\mathcal{A}_{\mathcal{G}}$ is a symmetric matrix. It is easily shown that an undirected graph must be a balanced digraph. For a given positive integer k, the union of k digraphs $\mathcal{G}_1 = \{\mathcal{V}, \mathcal{E}_{\mathcal{G}_1}, \mathcal{A}_{\mathcal{G}_1}\}$, ..., $\mathcal{G}_k = \{\mathcal{V}, \mathcal{E}_{\mathcal{G}_k}, \mathcal{A}_{\mathcal{G}_k}\}$ is denoted by $\sum_{j=1}^{k} \mathcal{G}_j = \{\mathcal{V}, \cup_{j=1}^{k} \mathcal{E}_{\mathcal{G}_j}, \sum_{j=1}^{k} \mathcal{A}_{\mathcal{G}_j}\}$. By the definition of a Laplacian matrix we know that $L_{\sum_{j=1}^{k} \mathcal{G}_j} = \sum_{j=1}^{k} L_{\mathcal{G}_j}$. A sequence (i_1, i_2), (i_2, i_3), ..., (i_{k-1}, i_k) of edges is called a directed path from node i_1 to node i_k. $\mathcal{G}$ is called a strongly connected digraph if for any $i, j \in \mathcal{V}$, there is a directed path from i to j. A strongly connected undirected graph is also called a connected graph. A directed tree is a digraph where every node except the root has exactly one neighbor and the root is a source. A spanning tree of $\mathcal{G}$ is a directed tree whose node set is $\mathcal{V}$ and whose edge set is a subset of $\mathcal{E}_{\mathcal{G}}$. For a balanced digraph, containing a spanning tree is equivalent to being strongly connected. We call $\{\mathcal{G}_1, ..., \mathcal{G}_k\}$ jointly-containing-spanning-tree if $\sum_{j=1}^{k} \mathcal{G}_j$ has a spanning tree. Especially, if $\mathcal{G}_j$, $j = 1, 2, ..., k$, are all undirected graphs and $\{\mathcal{G}_1, ..., \mathcal{G}_k\}$ jointly contains a spanning tree, then $\sum_{j=1}^{k} \mathcal{G}_j$ is connected. In this case, $\{\mathcal{G}_1, ..., \mathcal{G}_k\}$ is called jointly-connected.

Now we state a basic theorem on Laplacian matrices.

Theorem 12.1. *([16], [25]) If $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_\mathcal{G}, \mathcal{A}_\mathcal{G}\}$ is an undirected graph, then $L_\mathcal{G}$ is a symmetric matrix and has N real eigenvalues in ascending order:*

$$0 = \lambda_1(L_\mathcal{G}) \leq \lambda_2(L_\mathcal{G}) \leq ... \leq \lambda_N(L_\mathcal{G}),$$

and

$$\min_{x \neq 0, \mathbf{1}^T x = 0} \frac{x^T L_\mathcal{G} x}{\|x\|_2^2} = \lambda_2(L_\mathcal{G}),$$

where $\lambda_2(L_\mathcal{G})$ is called the algebraic connectivity of $\mathcal{G}$. Particularly, if $\mathcal{G}$ is connected, then $\lambda_2(L_\mathcal{G}) > 0$.

We consider the average-consensus control for a network of discrete-time first-order agents with the dynamics

$$x_i(t+1) = x_i(t) + u_i(t), \ t = 0, 1, ..., \ i = 1, ..., N, \tag{12.3}$$

where $x_i(t)$ and $u_i(t)$ are the state and control of the ith agent. Here for simplicity, we suppose that $x_i(t)$ and $u_i(t)$ are scalars and the initial state $x_i(0)$ is deterministic.

The ith agent can receive information from its neighbors:

$$y_{ji}(t) = x_j(t) + w_{ji}(t), \ j \in N_i, \tag{12.4}$$

where $y_{ji}(t)$ denotes the measurement of the jth agent's state $x_j(t)$ by the ith agent, and $\{w_{ji}(t), i, j = 1, 2, ..., N\}$ are the communication noises. The graph $\mathcal{G}$ shows the structure of the information flow in system (12.3), called the information flow graph or network topology graph of system (12.3). Denote $X(t) = [x_1(t), ..., x_N(t)]^T$; $(\mathcal{G}, X)$ is usually called a dynamic network [16].

We call the group of controls $\mathcal{U} = \{u_i, i = 1, 2, ..., N\}$ a measurement-based distributed protocol if each $u_i(t)$ depends only on the state of the ith agent and the measurement of its neighbors' states, that is,

$$u_i(t) \in \sigma(\cup_{s=0}^t \sigma(x_i(s), y_{ji}(s), j \in N_i)), \ t = 0, 1, ..., \ i = 1, 2, ..., N.$$

The so-called consensus control means to design a measurement-based distributed protocol for the dynamic network $(\mathcal{G}, X)$ such that all agents achieve an agreement on their states in some sense as $t \to \infty$. The so-called average-consensus control means to design a distributed protocol for the dynamic network $(\mathcal{G}, X)$ such that, for any initial value $X(0)$, the states of all the agents converge to $\frac{1}{N}\sum_{j=1}^N x_j(0)$ as $t \to \infty$, that is, $\frac{1}{N}\sum_{j=1}^N x_j(0)$ can be computed in a distributed way. For this case, $\frac{1}{N}\sum_{j=1}^N x_j(0)$ is called the group decision value [16].

Applying the distributed protocol $\mathcal{U}$ to system (12.3)–(12.4), generally speaking, will lead to a stochastic closed-loop system, and $x_i(t)$, $i = 1, 2, ..., N$, are all stochastic processes. A distributed protocol $\mathcal{U}$ is called an asymptotically unbiased mean square average-consensus protocol if it renders system (12.3)–(12.4) has the following properties: for any given $X(0) \in \mathcal{R}^N$, there is a random variable x^* such that $E(x^*) = \frac{1}{N}\sum_{j=1}^N x_j(0)$, $\mathbf{Var}(x^*) < \infty$, and

$$\lim_{t\to\infty} \mathbf{E}[x_i(t) - x^*]^2 = 0, \ i = 1, 2, ..., N.$$

For the dynamic network $(\mathcal{G}, X)$, we use the distributed protocol

$$u_i(t) = a(t)\sum_{j\in N_i} a_{ij}(y_{ji}(t) - x_i(t)), \ t = 0, 1, ..., \tag{12.5}$$

where and whereafter $a(t) > 0$, called a consensus-gain function, and if $|N_i| = 0$, then the sum $\sum_{j\in N_i}(\cdot)$ is defined as zero.

We call (12.5) a distributed stochastic approximation type protocol. It intuitively means that each agent updates its state in the direction of the nonnegative gradient of its local Laplacian potential. If the digraph $\mathcal{G}$ is balanced, then $V_{\mathcal{G}} \overset{\triangle}{=} x^T L_{\mathcal{G}} x = \frac{1}{2}\sum_{i=1}^N \sum_{j\in N_i} a_{ij}(x_j - x_i)^2$ is called the Laplacian potential function associated with $\mathcal{G}$ [16], which represents the degree of deviation between different agents' states, $\frac{1}{2}\sum_{j\in N_i} a_{ij}(x_j - x_i)^2$ is called the local Laplacian potential of the ith agent, and the nonnegative gradient direction with respect to x_i is $\sum_{j\in N_i} a_{ij}(x_j - x_i)$. Due to the communication noises, in the protocol (12.5), the update direction of the ith agent at time t is $\sum_{j\in N_i} a_{ij}(y_{ji}(t) - x_i(t))$, and $a(t)$ is the step size. When there is no communication noise (i.e., $y_{ji}(t) = x_j(t)$) and $a(t) \equiv 1$, (12.5) degenerates to the protocol (A.1) of [16].

Fixed Topology Case

In this section, we prove that, under mild conditions, the control law (12.5) is an asymptotically unbiased mean square average-consensus and almost sure strong consensus protocol.

For conciseness of expression, we further use the following notation:

$$\mathcal{S} = \{\xi \mid \{\xi(t) \in \mathcal{R}^{N\times N}, \mathcal{F}^\xi(t)\} \text{ is a martingale difference, } \sigma_\xi \overset{\triangle}{=} \sup_{t\geq 0} \mathbf{E}\|\xi(t)\|_2^2 < \infty\},$$

$$\mathcal{S}' = \{\xi \mid \xi \in \mathcal{S}, \sup_{t\geq 0} \mathbf{E}(\|\xi(t)\|_2^2 | \mathcal{F}^\xi(t-1)) < \infty \text{ a.s.}\},$$

$$\widetilde{\mathcal{S}}' = \{\xi \mid \xi \in \mathcal{S}, \sup_{t\geq 0, m\geq 0} \mathbf{E}(\|\xi(t+m)\|_2^2 | \mathcal{F}^\xi(t)) < \infty \text{ a.s.}\},$$

where and whereafter $\mathcal{F}^\xi(t) = \sigma\{\xi(0), ..., \xi(t)\}$.

Obviously, $\widetilde{\mathcal{S}}' \subset \mathcal{S}' \subset \mathcal{S}$. If $\{\xi(t) \in \mathcal{R}^{N\times N}, t = 0, 1, ...\}$ is a sequence of independent r.v.s with zero mean and uniformly bounded second-order moments, then $\{\xi(t), t = 0, 1, ...\} \in \widetilde{\mathcal{S}}'$. So bounded and Gaussian white noises both belong to $\widetilde{\mathcal{S}}'$.

Substituting the protocol (12.5) into (12.3) leads to

$$X(t+1) = [I_N - a(t)L_{\mathcal{G}}]X(t) + a(t)D_{\mathcal{G}}W(t), \ t = 0, 1, ..., \tag{12.6}$$

where and whereafter $D_{\mathcal{G}} = diag(\alpha_1^T, ..., \alpha_N^T)$ is an $N \times N^2$-dimensional block diagonal matrix with α_i being the ith row of $\mathcal{A}_{\mathcal{G}}$, and $W(t) = [w_1^T(t), ..., w_N^T(t)]^T$ with $w_i(t) = [w_{1i}(t), ..., w_{Ni}(t)]^T$.

We need the following assumptions:

A1) $\mathcal{G}$ is a balanced digraph;

A2) $\mathcal{G}$ contains a spanning tree;

A3) $\sum_{t=0}^{\infty} a(t) = \infty$, $\sum_{t=0}^{\infty} a^2(t) < \infty$;

A3)$'$ $\sum_{t=0}^{\infty} a(t) = \infty$, $\lim_{t\to\infty} a(t) = 0$.

We have the following theorems.

Theorem 12.2. *Apply the protocol (12.5) to system (12.3)–(12.4) and suppose that A1), A2), and A3)$'$ hold. Then, for any $W \in \mathcal{S}$,*

$$\lim_{t\to\infty} \mathbf{E}[V(t)] = 0, \ \forall \ X(0) \in \mathcal{R}^N, \tag{12.7}$$

that is, (12.5) is a mean square weak consensus protocol. Here $V(t) = \left\| \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) X(t) \right\|_2^2$ is the energy function of the consensus error.

Proof. Denote $J = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ denotes the N-dimensional column vector with all elements 1, and

$$\delta(t) = (I_N - J)X(t). \tag{12.8}$$

Then $V(t) = \delta^T(t)\delta(t)$. Thus, from A1) and Theorem 6 of [16] we have $\mathbf{1}^T L_{\mathcal{G}} = 0$ and $JL_{\mathcal{G}} = 0$, which together with (12.6), leads to

$$\begin{aligned} \delta(t+1) &= X(t) - a(t)L_{\mathcal{G}}X(t) + a(t)D_{\mathcal{G}}W(t) - JX(t) - a(t)JD_{\mathcal{G}}W(t) \\ &= \delta(t) - a(t)L_{\mathcal{G}}X(t) + a(t)(I-J)D_{\mathcal{G}}W(t) \\ &= [I_N - a(t)L_{\mathcal{G}}]\delta(t) + a(t)(I-J)D_{\mathcal{G}}W(t) \end{aligned} \tag{12.9}$$

and

$$\begin{aligned} V(t+1) &= V(t) - 2a(t)\delta^T(t)\widehat{L}_{\mathcal{G}}\delta(t) + a^2(t)\delta^T(t)L_{\mathcal{G}}^T L_{\mathcal{G}}\delta(t) \\ &\quad + 2a(t)\delta^T(t)(I - a(t)L_{\mathcal{G}}^T)(I-J)D_{\mathcal{G}}W(t) \\ &\quad + a^2(t)W^T(t)D_{\mathcal{G}}^T(I-J)^2 D_{\mathcal{G}}W(t). \end{aligned} \tag{12.10}$$

From A1) and Theorem 7 of [16], $\widehat{L}_{\mathcal{G}} = \frac{L_{\mathcal{G}}+L_{\mathcal{G}}^T}{2}$ is the Laplacian matrix of the symmetrized graph[1] $\widehat{\mathcal{G}}$ of $\mathcal{G}$. From A2), noticing that $\widehat{\mathcal{G}}$ is undirected, we know that $\widehat{\mathcal{G}}$ is strongly connected, and hence, from Theorem 12.1, $\lambda_2(\widehat{L}_{\mathcal{G}}) > 0$. Therefore, from $\delta^T(t)\widehat{L}_{\mathcal{G}}\delta(t) \geq \lambda_2(\widehat{L}_{\mathcal{G}})V(t)$ and (12.10) we have

$$\begin{aligned} V(t+1) &\leq (1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) + a^2(t)\|L_{\mathcal{G}}\|_2^2)V(t) \\ &\quad + 2a(t)\delta^T(t)(I - a(t)L_{\mathcal{G}}^T)(I-J)D_{\mathcal{G}}W(t) \\ &\quad + a^2(t)W^T(t)D_{\mathcal{G}}^T(I-J)^2 D_{\mathcal{G}}W(t). \end{aligned} \tag{12.11}$$

Noticing that $\delta(t) \in \mathcal{F}^W(t-1)$ and $W \in \mathcal{S}$, taking mathematical expectation on both sides of this inequality, we have

$$\mathbf{E}[V(t+1)] \leq (1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) + a^2(t)\|L_{\mathcal{G}}\|_2^2)\mathbf{E}[V(t)] + a^2(t)\|D_{\mathcal{G}}\|_2^2\|I-J\|_2^2\sigma_W. \tag{12.12}$$

Noticing that $\lambda_2(\widehat{L}_{\mathcal{G}}) > 0$ and $a(t) \to 0$, $t \to \infty$, we get that there is $t_0 > 0$ such that $a(t)\|L_{\mathcal{G}}\|_2^2 < \lambda_2(\widehat{L}_{\mathcal{G}})$ and $2a(t)\lambda_2(\widehat{L}_{\mathcal{G}}) \leq 1$, $\forall\, t \geq t_0$. Thus

$$0 \leq 1 - 2a(t)\lambda_2(\widehat{L}_{\mathcal{G}}) + a^2(t)\|L_{\mathcal{G}}\|_2^2 < 1, \ \forall\, t \geq t_0. \tag{12.13}$$

Then by A3)′ we have

$$\sum_{t=t_0}^{\infty}(2a(t)\lambda_2(\widehat{L}_{\mathcal{G}}) - a^2(t)\|L_{\mathcal{G}}\|_2^2) \geq \lambda_2(\widehat{L}_{\mathcal{G}})\sum_{t=t_0}^{\infty}a(t) = \infty \tag{12.14}$$

and

$$\frac{a^2(t)}{2a(t)\lambda_2(\widehat{L}_{\mathcal{G}}) - a^2(t)\|L_{\mathcal{G}}\|_2^2} \to 0, \ t \to \infty, \tag{12.15}$$

which, together with (12.13), (12.14), and Lemma 12.A.1 in Appendix 12.A, leads to (12.7). □

[1] The definition of the symmetrized graph of a digraph is referred to Definition 2 of [16].

In the protocol (12.5), an agent-independent consensus gain $a(t)$ is used. This requires some coordination of the consensus gain across the agents. It is interesting to investigate the case with agent-dependent consensus gains. For instance, in practical applications, there may be a small error between the actual consensus gain $a_i(t)$ of the ith agent and the designed consensus gain $a(t)$. For this case, the protocol (12.5) becomes

$$u_i(t) = a_i(t)\sum_{j=1}^{N} a_{ij}(y_{ji}(t) - x_i(t)),\ t = 0, 1, \ldots. \tag{12.16}$$

We have the following theorem.

Theorem 12.3. *Apply the protocol (12.16) to system (12.3)–(12.4). If Assumptions A1)–A2) hold and*

$$\sum_{t=0}^{\infty} a_j(t) = \infty,\ j = 1, 2, ..., N, \tag{12.17}$$

$$\lim_{t\to\infty} a_j(t) = 0,\ j = 1, 2, ..., N, \tag{12.18}$$

$$\max_{1\le i,j\le N} |a_i(t) - a_j(t)| = o(\sum_{j=1}^{N} a_j(t)),\ t \to \infty, \tag{12.19}$$

then, for any $W \in \mathcal{S}$,

$$\lim_{t\to\infty} \mathbf{E}[V(t)] = 0,\ \forall\ X(0) \in \mathcal{R}^N. \tag{12.20}$$

Proof. Denote

$$\bar{a}(t) = \frac{1}{N}\sum_{j=1}^{N} a_j(t),\ \Delta(t) = diag(\Delta_1(t), ..., \Delta_N(t)),$$

where $\Delta_i(t) = \bar{a}(t) - a_i(t)$. Substituting the protocol (12.16) into system (12.3)–(12.4), similarly to (12.10), we have

$$\begin{aligned} V(t+1) \;\; = \;\; & V(t) - 2\bar{a}(t)\delta^T(t)\widehat{L_G}\delta(t) + \bar{a}^2(t)\delta^T(t)L_G^T L_G\delta(t) \\ & + 2\delta^T(t)(I - \bar{a}(t)L_G^T)\Delta(t)L_G\delta(t) + \delta^T(t)L_G^T\Delta^2(t)L_G\delta(t) \\ & + 2\bar{a}(t)\delta^T(t)L_G^T\Delta(t)(I - J)D_G W(t) \\ & + 2\bar{a}(t)\delta^T(t)(I - \bar{a}(t)L_G^T)(I - J)D_G W(t) \\ & + \bar{a}^2(t)W^T(t)D_G^T(I - J)^2 D_G W(t). \end{aligned} \tag{12.21}$$

From the above, noting that $\max_j \sup_{t\geq 0} a_j(t) < \infty$, similarly to (12.12), we have

$$\mathbf{E}[V(t+1)] \quad \leq \quad (1-q(t))\mathbf{E}[V(t)] + \bar{a}^2(t)\|D_{\mathcal{G}}\|_2^2\|I-J\|_2^2\sigma_W, \tag{12.22}$$

where

$$q(t) \quad = \quad 2\lambda_2(\widehat{L}_{\mathcal{G}})\bar{a}(t) - 2(1+\alpha_0\|L_{\mathcal{G}}\|_2)\|L_{\mathcal{G}}\|_2\|\Delta(t)\|_2 - \|\Delta(t)\|_2^2\|L_{\mathcal{G}}\|_2^2$$

and $\alpha_0 = \max_j \sup_{t\geq 0} a_j(t)$. By (12.18) we know that

$$\lim_{t\to\infty} \bar{a}(t) = 0. \tag{12.23}$$

By (12.17) we have

$$\sum_{t=0}^{\infty} \bar{a}(t) = \infty. \tag{12.24}$$

Noting that $\|\Delta(t)\|_2 \leq \max_{1\leq i,j\leq N} |a_i(t) - a_j(t)|$, by (12.19), we get that

$$\|\Delta(t)\|_2 = o(\bar{a}(t)), \ t \to \infty.$$

Then by (12.23) and (12.24), similarly to (12.13)–(12.15), we get that there exists $t_1 > 0$ such that

$$0 < q(t) \leq 1, \ \forall\, t \geq t_1, \tag{12.25}$$

$$\sum_{t=t_1}^{\infty} q(t) = \infty, \tag{12.26}$$

and

$$\frac{\bar{a}^2(t)}{q(t)} \to 0, \ t \to \infty. \tag{12.27}$$

Then by (12.25), (12.26), (12.27), and Lemma 12.A.1 in Appendix 12.A we have (12.20). □

For the sufficient conditions ensuring (12.5) to be an asymptotically unbiased mean square average-consensus protocol, we have the following theorem.

Theorem 12.4. *Apply the protocol (12.5) to system (12.3)–(12.4). If Assumptions A1)–A3) hold, then, for any $W \in \mathcal{S}$,*

$$\lim_{t\to\infty} \mathbf{E}[x_i(t) - x^*]^2 = 0,\ i = 1, 2, ..., N,\ \forall\ X(0) \in \mathcal{R}^N, \tag{12.28}$$

where x^* is a r.v. depending on W and $X(0)$ and satisfying

$$\begin{aligned} \mathbf{E}(x^*) &= \frac{1}{N}\sum_{j=1}^{N} x_j(0), \\ \mathbf{Var}(x^*) &\leq \frac{\sigma_W^* |\mathcal{E}_\mathcal{G}| \sum_{i=1}^N \sum_{j\in N_i} a_{ij}^2}{N^2} \sum_{t=0}^{\infty} a^2(t). \end{aligned}$$

In particular, if $\{w_{ji}(t), t = 0, 1, ...\}$, $i = 1, 2, ..., N$, $j \in N_i$, are mutually independent, then $V_* \leq \mathbf{Var}(x^*) \leq V^*$, where

$$V^* = \frac{\sigma_W^* |\mathcal{E}_\mathcal{G}| \max_{1\leq i<j\leq N} a_{ij}^2}{N^2} \sum_{t=0}^{\infty} a^2(t),\ V_* = \frac{\sigma_{W_*} |\mathcal{E}_\mathcal{G}| \min_{1\leq i<j\leq N} a_{ij}^2}{N^2} \sum_{t=0}^{\infty} a^2(t),$$
$$\sigma_W^* = \max_{(j,i)\in\mathcal{E}_\mathcal{G}} \sup_{t\geq 0} \mathbf{E}[w_{ji}(t)]^2,\ \sigma_{W_*} = \min_{(j,i)\in\mathcal{E}_\mathcal{G}} \inf_{t\geq 0} \mathbf{E}[w_{ji}(t)]^2,$$

that is, (12.5) is an asymptotically unbiased mean square average-consensus protocol.

Proof. For all $W \in \mathcal{S}$, from (12.6) and $\mathbf{1}^T L_\mathcal{G} = 0$ it follows that

$$\frac{1}{N}\sum_{j=1}^{N} x_j(t+1) = \frac{1}{N}\sum_{j=1}^{N} x_j(t) + a(t)\frac{1}{N}\mathbf{1}^T D_\mathcal{G} W(t).$$

Taking summation for both sides of the above equations from $t = 0$ to $t = n - 1$ leads to

$$\frac{1}{N}\sum_{j=1}^{N} x_j(n) = \frac{1}{N}\sum_{j=1}^{N} x_j(0) + \frac{1}{N}\mathbf{1}^T D_\mathcal{G} \sum_{t=0}^{n-1} a(t)W(t). \tag{12.29}$$

Since $W \in \mathcal{S}$ and $\sum_{t=0}^{\infty} a^2(t) < \infty$, we get that $(\sum_{t=0}^{n} a(t)W(t), \mathcal{F}^W(n))$ is a martingale with

$$\sup_{n\geq 0} \mathbf{E}\left\| \sum_{t=0}^{n} a(t)W(t) \right\|_2^2 < \infty.$$

Then by Theorem 7.6.10 of [26] it follows that $\sum_{t=0}^{n} a(t)W(t)$ converges in mean square as $n \to \infty$. Denote the limit by $\sum_{t=0}^{\infty} a(t)W(t)$. Then (12.28) follows from Theorem 12.2 with

$$x^* = \frac{1}{N}\sum_{j=1}^{N} x_j(0) + \frac{1}{N}\mathbf{1}^T D_{\mathcal{G}} \sum_{t=0}^{\infty} a(t)W(t).$$

By Corollary 4.2.5 of [27] we have

$$\begin{aligned}
\mathbf{E}(x^*) &= \frac{1}{N}\sum_{j=1}^{N} x_j(0), \\
\mathbf{Var}(x^*) &= \lim_{n\to\infty} \mathbf{E}\Big(\frac{1}{N}\mathbf{1}^T D_{\mathcal{G}} \sum_{t=0}^{n} a(t)W(t)\Big)^2 \\
&= \frac{1}{N^2}\sum_{t=0}^{\infty}\Big\{a^2(t)\mathbf{E}(\sum_{i,j} a_{ij} w_{ji}(t))^2\Big\}. \qquad (12.30)
\end{aligned}$$

This, together with the Cauchy inequality, gives

$$\begin{aligned}
\mathbf{Var}(x^*) &\le \frac{\sum_{i=1}^{N}|N_i|}{N^2} \lim_{n\to\infty}\sum_{t=0}^{n}\Big\{a^2(t)\sum_{i,j} a_{ij}^2 \mathbf{E}(w_{ji}(t))^2\Big\} \\
&\le \frac{\sigma_W^* |\mathcal{E}_{\mathcal{G}}| \sum_{i,j} a_{ij}^2}{N^2}\sum_{t=0}^{\infty} a^2(t).
\end{aligned}$$

When $\{w_{ji}(t), t = 0, 1, \ldots\}$, $i = 1, 2, \ldots, N$, $j \in N_i$, are independent, by (12.30) we have

$$\begin{aligned}
\mathbf{Var}(x^*) &= \frac{1}{N^2}\lim_{n\to\infty}\sum_{t=0}^{n}\Big\{a^2(t)\sum_{i,j} a_{ij}^2 \mathbf{E}(w_{ji}(t))^2\Big\} \\
&\le \frac{\sigma_W^* |\mathcal{E}_{\mathcal{G}}| \max_{1\le i<j\le N} a_{ij}^2}{N^2}\sum_{t=0}^{\infty} a^2(t), \\
\mathbf{Var}(x^*) &\ge \frac{\sigma_{W_*} |\mathcal{E}_{\mathcal{G}}| \min_{1\le i<j\le N} a_{ij}^2}{N^2}\sum_{t=0}^{\infty} a^2(t).
\end{aligned}$$

This completes the proof of the theorem. □

Theorems 12.2–12.4 indicate that, for fixed topologies, Assumptions A1)–3) are a sufficient condition for the protocol (12.5) to ensure mean square weak consensus and asymptotically unbiased mean square average-consensus. Assumption A2) is to ensure the connectivity of the network to some extent, that is, the algebraic connectivity $\lambda_2(\widehat{L}_{\mathcal{G}}) > 0$, so that different agents may asymptotically agree on their states; Assumption A1) is to ensure the state average evolve around $\frac{1}{N}\sum_{j=1}^{N} x_j(0)$ so that an average-consensus can be achieved.

Assumption A3) is the step rule of standard stochastic approximation. From the proof of Theorem 12.2 we can see that the condition $\sum_{t=0}^{\infty} a(t) = \infty$ is to ensure the consensus error converges to zero with a certain rate. From the proof of Theorem 12.4 we can see the important role played by the condition $\sum_{t=0}^{\infty} a^2(t) < \infty$: when there are communication noises, by (12.29) the state average of the closed-loop system is not a constant anymore, and $\sum_{t=0}^{\infty} a^2(t) < \infty$ ensures the convergence of the state average of the closed-loop system.

From Theorem 12.4 we can see that, under the control of the protocol (12.5), there exists a static error between the final state of the closed-loop system and the average of the initial states. $\mathbf{Var}(x^*)$ describes the static error in the sense of mean square. In fact, we can show that if the conditions of Theorem 12.4 hold, then

$$\mathbf{Var}(x^*) = \limsup_{t\to\infty} \max_{1\le i\le N} \mathbf{E}\left[x_i(t) - \frac{1}{N}\sum_{j=1}^{N} x_j(0)\right]^2,$$

that is, $\mathbf{Var}(x^*)$ gives the static maximum mean square error between each individual state and the average of the initial states of the whole system.

In some applications of the information fusion of wireless sensor networks, the number N of network nodes is usually very large. Theorem 12.4 gives the analytic expression of the static maximum mean square error between each individual state and the average of the initial states of the whole system, from which we can see the impact of N on the accuracy of the information fusion. When the noises of different communication channels are mutually independent, $\mathbf{Var}(x^*)$ is proportional to $\frac{|\mathcal{E}|}{N^2}$. In particular, if $|\mathcal{E}| = O(N)$ and $\max_{1\le i\le j\le N} a_{ij} = O(1)$, then $\mathbf{Var}(x^*) = O(N^{-1})$, $N \to \infty$. This means that the more the network nodes, the better the effect of the information fusion. However, a large number of nodes will result in a high cost for running and maintenance of the whole network, so the choice of N is a trade-off between the fusion accuracy and the cost.

On necessity of Assumptions A1)–A3) for asymptotically unbiased mean square average-consensus, we have the following result.

Theorem 12.5. *Apply the protocol (12.5) to system (12.3)–(12.4). If (12.5) is an asymptotically unbiased mean square average-consensus protocol for any $W \in \mathcal{S}$, then Assumptions A1)–A3) hold.*

Proof. The proof is provided in Appendix 12.A. □

Remark 12.1. *In [28], sufficient conditions are given to ensure mean square weak consensus for undirected graphs with independent and identically distributed communication noises. In [29], a necessary and sufficient condition is given to ensure continuous-time mean square average-consensus for Gaussian noises. Here, from Theorems 12.4–12.5 we can see that A1)–A3) are necessary and sufficient conditions ensuring that (12.5) is a mean square average-consensus protocol for any communication noises that are martingale differences with bounded second-order moments.*

For the special case with no communication noise, a sufficient condition for the protocol (12.5) ensuring average-consensus is given by the following theorem.

Theorem 12.6. *Apply the protocol (12.5) to system (12.3)–(12.4) with $W(t) = 0$, $t = 0, 1, \ldots$. If A1)–A2) hold, $\sum_{t=0}^{\infty} a(t) = \infty$, and*

$$\limsup_{t\to\infty} a(t) < \mu, \tag{12.31}$$

where $\mu = \min\{\frac{2\lambda_2(\widehat{L}_{\mathcal{G}})}{\|L_{\mathcal{G}}\|_2^2}, \frac{1}{2\lambda_2(\widehat{L}_{\mathcal{G}})}\}$, then $\lim_{t\to\infty} \|X(t) - JX(0)\|_2 = 0$, $\forall\, X(0) \in \mathcal{R}^N$.

Proof. Noticing that $W(t) = 0$, $t = 0, 1, \ldots$, by A1) and A2), similarly to (12.12), we have

$$V(t+1) \le (1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) + a^2(t)\|L_{\mathcal{G}}\|_2^2)V(t) + a^2(t)\|D_{\mathcal{G}}\|_2^2\|I - J\|_2^2\sigma_W.$$

Take a constant $\epsilon_0 \in (0, \mu - \limsup_{t\to\infty} a(t))$. Then by (12.31) we have that there is $t_0 > 0$ such that

$$a(t) \le \frac{2\lambda_2(\widehat{L}_{\mathcal{G}})}{\|L_{\mathcal{G}}\|_2^2} - \epsilon_0,\ \forall\, t \ge t_0, \tag{12.32}$$

and

$$a(t) \le \frac{1}{2\lambda_2(\widehat{L}_{\mathcal{G}})} - \epsilon_0,\ \forall\, t \ge t_0. \tag{12.33}$$

By (12.32) we have

$$1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) + a^2(t)\|L_{\mathcal{G}}\|_2^2 \le 1 - \epsilon_0\|L_{\mathcal{G}}\|_2^2 a(t) < 1,\ \forall\, t \ge t_0. \tag{12.34}$$

By (12.33) we have

$$1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) + a^2(t)\|L_{\mathcal{G}}\|_2^2 \ge 1 - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) \ge 2\epsilon_0\lambda_2(\widehat{L}_{\mathcal{G}}) \ge 0,\ \forall\, t \ge t_0, \tag{12.35}$$

From (12.32) and $\sum_{t=0}^{\infty} a(t) = \infty$ we get

$$\sum_{t=t_0}^{\infty}(2\lambda_2(\widehat{L}_{\mathcal{G}})a(t) - a^2(t)\|L_{\mathcal{G}}\|_2^2) \geq \epsilon_0 \|L_{\mathcal{G}}\|_2^2 \sum_{t=t_0}^{\infty} a(t) = \infty,$$

which, together with (12.34), (12.35), and Lemma 12.A.1, leads to

$$\lim_{t\to\infty} V(t) = 0. \tag{12.36}$$

Similarly to (12.29), noticing that $W(t) = 0$, $t = 0, 1, ...,$ we have

$$\frac{1}{N}\sum_{j=1}^{N} x_j(t) = \frac{1}{N}\sum_{j=1}^{N} x_j(0), \; t = 1, 2, \ldots .$$

This, together with (12.36), leads to the conclusion of the theorem. □

From the following theorem we can see that, under Assumptions A1)–A3), for a class of communication noises, the protocol (12.5) can ensure the almost sure consensus as well.

Theorem 12.7. *Apply the protocol (12.5) to system (12.3)–(12.4). If Assumptions A1)–A3) hold, then, for any $W \in \mathcal{S}'$,*

$$\lim_{t\to\infty} x_i(t) = x^* \; a.s., \; i = 1, 2, ..., N, \forall \; X(0) \in \mathcal{R}^N, \tag{12.37}$$

that is, (12.5) is an almost sure strong consensus protocol [28]. Here x^* is given by Theorem 12.4. Furthermore, if $a(t) \downarrow 0$, $t \to \infty$, then

$$\frac{1}{n}\sum_{t=0}^{n} \|\delta(t)\|_2 = o\left(\frac{1}{\sqrt{a(n)n}}\right), \; n \to \infty, \; a.s., \tag{12.38}$$

where $\delta(t)$ is given by (12.8).

Proof. For all $W \in \mathcal{S}'$, from $\delta(t) \in \mathcal{F}^W(t-1)$ and (12.11) it follows that

$$\begin{aligned} \mathbf{E}(V(t+1)|\mathcal{F}^W(t-1)) \;\; \leq \;\; & (1 + a^2(t)\|L_{\mathcal{G}}\|_2^2)V(t) - 2\lambda_2(\widehat{L}_{\mathcal{G}})a(t)V(t) \\ & + a^2(t)tr(D_{\mathcal{G}}^T(I - J)^2 D_{\mathcal{G}}) \\ & \times \mathbf{E}(\|W(t)\|_2^2|\mathcal{F}^W(t-1)) \text{ a.s.} \end{aligned} \tag{12.39}$$

Noticing that $\sup_{t\geq 0}\mathbf{E}(\|W(t)\|_2^2|\mathcal{F}^W(t-1))<\infty$ a.s. and $\sum_{t=0}^{\infty}a^2(t)<\infty$, by the nonnegative supermartingale convergence theorem [30], [31] we have that $V(t)$ converges almost surely as $t\to\infty$ and

$$\sum_{t=0}^{\infty}a(t)V(t)<\infty \text{ a.s.} \tag{12.40}$$

Furthermore, by Theorem 12.2 and $\mathcal{S}'\subset\mathcal{S}$,

$$\lim_{t\to\infty}V(t)=0 \text{ a.s.} \tag{12.41}$$

Since $W\in\mathcal{S}'$ and $\sum_{t=0}^{\infty}a^2(t)<\infty$, it follows that $\{\sum_{t=0}^{n}a(t)W(t),\mathcal{F}^W(n)\}$ is a martingale with

$$\sup_{n\geq 0}\mathbf{E}\Big\|\sum_{t=0}^{n}a(t)W(t)\Big\|_2^2<\infty.$$

Then by Theorem 7.6.10 of [26] we get that $\sum_{t=0}^{n}a(t)W(t)$ converges both in mean square and almost surely as $n\to\infty$. Thus by (12.29) and (12.41) we get that $x_i(t)$ converges almost surely as $t\to\infty$, $i=1,2,\ldots,N$. This, together with Theorem 12.4, gives (12.37).

If $a(t)\downarrow 0$, $t\to\infty$, then by Kronecker lemma [27] and (12.40) we have

$$\lim_{n\to\infty}a(n)\sum_{t=0}^{n}V(t)=0 \text{ a.s.},$$

which, together with the Cauchy inequality, results in

$$\frac{1}{n}\sum_{t=0}^{n}\|\delta(t)\|_2\leq\left(\frac{1}{n}\sum_{t=0}^{n}V(t)\right)^{1/2}=o\left(\frac{1}{\sqrt{a(n)n}}\right) \text{ a.s.} \quad \square$$

Theorem 12.7 implies that, under the same conditions, the states of different agents converge asymptotically to a common random variable with probability one. Note that this random variable may not be precisely the average of the initial states, although its sample mean is. For this case, (12.38) gives a rough estimate for the convergence rate of n-step mean consensus error.

Time-Varying Topology Case

In this case, the distributed protocol is running above a flow of topology graphs $\{\mathcal{G}(t), t = 0, 1, ...\}$, where $\mathcal{G}(t) = \{\mathcal{V}, \mathcal{E}_{\mathcal{G}(t)}, \mathcal{A}_{\mathcal{G}(t)}\}$, $t = 0, 1, ...$, is a sequence of digraphs with the same vertex set. The edge sets and weighted adjacency matrices are time varying.

The networks with time-varying topologies can be found in many engineering, biological, social, and economic systems, such as the creation and failure of communication links, the loss of data packages, the variation of the channel parameters, and the evolvement and reconfiguration of formations in swarms and flocking [32]. For many cases, fixed topologies are only ideal models; even if the protocol is designed for a fixed topology, then it is necessary to consider the robustness of the protocol with respect to the time-variation of the topology. For the stability and consensus of time-varying networked systems without communication noises, the readers are referred to [33], [34].

Here we will consider two kinds of typical topology graph flows:

$$\Gamma_1 = \{\{\mathcal{H}(t), t = 0, 1, ...\} |\ \mathcal{H}(t) \text{ is a balanced digraph } \forall\, t \geq 0,\ \sup_{t\geq 0} \|\mathcal{A}_{\mathcal{H}(t)}\|_2 < \infty\},$$

which is a family of all sequences of balanced graphs with bounded weighted adjacent matrix (we can see that a sequence of undirected graphs with bounded weighted adjacent matrix belongs to Γ_1);

$$\Gamma_2 = \{\{\mathcal{H}(t), t = 0, 1, ...\} |\ \mathcal{H}(t) \text{ is a balanced digraph } \forall\, t \geq 0,\ |\{\mathcal{H}(t), t = 0, 1, ...\}| < \infty\},$$

which is a family of sequences of switching balanced graphs. Obviously, $\Gamma_2 \subset \Gamma_1$. If $\{\mathcal{H}(t), t = 0, 1, ...\} \in \Gamma_2$, then the set $\{\mathcal{H}(t), t = 0, 1, ...\}$ has only finitely many elements. The most common sequence of switching balanced graphs is the sequence of undirected graphs $\{\mathcal{H}(t) = \{\mathcal{V}, \mathcal{E}_{\mathcal{H}(t)}, \mathcal{A}_{\mathcal{H}(t)}\}, t = 0, 1, ...\}$ with weighted adjacent matrices $\mathcal{A}_{\mathcal{H}(t)} = [a_{ij}(t)]_{N\times N}$, the elements of which take only two kinds of values: when i and j are mutually neighbors, $a_{ij}(t) = a_{ji}(t) = a_{ij} > 0$; otherwise, $a_{ij}(t) = a_{ji}(t) = 0$, $i, j \in \mathcal{V}$. This kind of sequences of switching undirected graphs are widely involved in the synchronization of Vicsek models [35], [36].

For $\{\mathcal{G}(t), t = 0, 1, ...\}$, the distributed network protocol is given by

$$u_i(t) = a(t) \sum_{j\in N_i(t)} a_{ij}(t)(y_{ji}(t) - x_i(t)),\ \forall\, t = 0, 1, ..., \tag{12.42}$$

where $a_{ij}(t)$ is the element of ith row and jth column of $\mathcal{A}_{\mathcal{G}(t)}$, which is the weighted adjacency matrix at time t. Note that $a_{ij}(t) > 0 \Leftrightarrow (j, i) \in \mathcal{E}_{\mathcal{G}(t)}$. Let $N_i(t) = \{j \in \mathcal{V}|\ a_{ij}(t) > 0\}$. Here

$$y_{ji}(t) = x_j(t) + w_{ji}(t),\ \ j \in N_i(t),\ \ t = 0, 1, \ldots. \tag{12.43}$$

Since $u_i(t)$ is adapted to $\sigma(x_i(t), y_{ji}(t), j \in N_i(t)), t = 0, 1, ..., i = 1, 2, ..., N, \mathcal{U} = \{u_1, ..., u_N\}$ is a distributed protocol. Substituting the protocol (12.42) into (12.3) gives

$$X(t+1) = [I_N - a(t)L_{\mathcal{G}(t)}]X(t) + a(t)D_{\mathcal{G}(t)}W(t),\ t = 0, 1, \ldots. \tag{12.44}$$

In this section, we need the following assumption:

A4) $a(t+1) \le a(t)$, $t = 0, 1, ...,$ and $\limsup_{t\to\infty} \frac{a(t)}{a(t+1)} < \infty$.

If $a(t) = \frac{1}{t+1}$ or $a(t) = \frac{\ln(t+2)}{t+1}$, then both A3) and A4) hold. In fact, if there are $\beta_1 \le 1$, $\beta_2 > -0.5$, $\gamma_1 \le 1$, $\gamma_2 > 0.5$, $c_1 > 0$, and $c_2 > 0$ such that, for sufficiently large t, $\frac{c_1}{t^{\gamma_1}[\ln(t)]^{\beta_1}} \le a(t) \le \frac{c_2[\ln(t)]^{\beta_2}}{t^{\gamma_2}}$, then A3) holds. If $a(t)$ decreases monotonically and there are $\gamma \in (0.5, 1]$, $\beta \ge -1$, $c_3 > 0$, and $c_4 > 0$ such that, for sufficiently large t, $\frac{c_3[\ln(t)]^{\beta}}{t^{\gamma}} \le a(t) \le \frac{c_4[\ln(t)]^{\beta}}{t^{\gamma}}$, then both A3) and A4) hold.

For convenience of citation, we further denote $\lambda_k^h = \lambda_2(L_{\widehat{\mathcal{G}}_k^h})$, where $\mathcal{G}_k^h = \sum_{i=k}^{k+h-1} \mathcal{G}(i)$. The main results of this section are summarized in the following theorems.

Theorem 12.8. *Apply the protocol (12.42) to system (12.3), (12.43). For any given $\{\mathcal{G}(t), t = 0, 1, \ldots\} \in \Gamma_1$, if there is an integer $h > 0$ such that $\inf_{m\ge 0} \lambda_{mh}^h > 0$ and Assumptions A3)′–A4) hold, then, for any $W \in \mathcal{S}$,*

$$\lim_{t\to\infty} \mathbf{E}[V(t)] = 0,\ \forall\ X(0) \in \mathcal{R}^N, \tag{12.45}$$

that is, (12.42) is a mean square weak consensus protocol.

Proof. Noticing that $\mathcal{G}(t)$ is a balanced graph, similarly to (12.9), by (12.44) we have

$$\delta(t+1) = [I_N - a(t)L_{\mathcal{G}(t)}]\delta(t) + a(t)(I - J)D_{\mathcal{G}(t)}W(t), \tag{12.46}$$

and hence

$$\delta[(m+1)h] = \Phi((m+1)h, mh)\delta(mh) + \overline{W}_{mh}^h, \tag{12.47}$$

where $\Phi(n+1, i) = (I_N - a(n)L_{\mathcal{G}(n)})\Phi(n, i)$, $\Phi(i, i) = I_N$, $i = 0, 1, ..., n$, $n = 1, 2, ...,$ and $\overline{W}_k^h = \sum_{j=k}^{k+h-1} \Phi(k+h-1, j)a(j)(I - J)D_{\mathcal{G}(j)}W(j)$.

By Assumption A4) we know that there exist a constant $C_h > 0$ and a positive integer m_0 such that $a(mh) \le C_h a[(m+1)h]$ and $a(mh) \le 1$, $\forall\ m \ge m_0$. Then since $\sup_{t\ge 0} \|\mathcal{A}_{\mathcal{G}(t)}\|_2 < \infty$, noting that $a(t) \downarrow 0$, we have

$$
\begin{aligned}
&\left\|\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) - I - \sum_{i=mh}^{(m+1)h-1} a(i)\left(L_{\mathcal{G}(i)} + L_{\mathcal{G}^T(i)}\right)\right\|_2 \\
\le\ & a^2(mh)\sum_{l=2}^{2h}\left(P_h\left(\max\left\{\sup_{t\ge 0}\|L_{\mathcal{G}(t)}\|_2, 1\right\}\right)^{P_h}\right) \\
\le\ & a^2[(m+1)h]M_h,\ \forall\, m \ge m_0,
\end{aligned}
$$

where $P_h = 2^{2h} - 2h - 1$ and

$$M_h = C_h^2 P_h(2h-1)\left(\max\left\{\sup_{t\ge 0}\|L_{\mathcal{G}(t)}\|_2, 1\right\}\right)^{P_h}. \tag{12.48}$$

Thus, by the definition of $V(t)$ and (12.47) we have

$$
\begin{aligned}
& V[(m+1)h] \\
\le\ & V(mh) - 2\delta^T(mh)\left[\sum_{i=mh}^{(m+1)h-1} a(i)\hat{L}(i)\right]\delta(mh) + a^2((m+1)h)M_h V(mh) \\
& + \left(\overline{W}_{mh}^h\right)^T \overline{W}_{mh}^h + 2\delta^T(mh)(\Phi((m+1)h, mh))^T\overline{W}_{mh}^h \\
\le\ & V(mh) - 2a((m+1)h)\delta^T(mh)\left[\sum_{i=mh}^{(m+1)h-1} \hat{L}(i)\right]\delta(mh) + a^2((m+1)h)M_h V(mh) \\
& + \left(\overline{W}_{mh}^h\right)^T \overline{W}_{mh}^h + 2\delta^T(mh)(\Phi((m+1)h, mh))^T\overline{W}_{mh}^h, \\
& \qquad \forall\, m \ge m_0,
\end{aligned} \tag{12.49}
$$

where $\hat{L}(i) = \frac{L_{\mathcal{G}(i)} + L_{\mathcal{G}(i)}^T}{2}$. From $W \in \mathcal{S}$, $\delta(mh) \in \mathcal{F}^W(mh-1)$, and the definition of $\overline{W}_{mh}^h$ it follows that

$$\mathbf{E}[\delta^T(mh)(\Phi((m+1)h, mh))^T\overline{W}_{mh}^h | \mathcal{F}^W(mh-1)] = 0 \text{ a.s.},$$

which implies

$$\mathbf{E}\left[\delta^T(mh)(\Phi((m+1)h, mh))^T\overline{W}_{mh}^h\right] = 0. \tag{12.50}$$

Further, since $\sup_{t\ge 0}\|D_{\mathcal{G}(t)}\| < \infty$, there exists a constant $N_h > 0$ such that

$$\mathbf{E}\left[\overline{W}(mh)^T\overline{W}_{mh}^h\right] \le N_h \sum_{i=mh}^{(m+1)h-1} a^2(i). \tag{12.51}$$

Since $\mathcal{G}(i)$, $i = 0, 1, \ldots$, are balanced digraphs, by Theorem 7 of [16], $\hat{L}_i = L_{\widehat{\mathcal{G}(i)}}$, where $\widehat{\mathcal{G}} = \{\mathcal{V}, \mathcal{E}_{\widehat{\mathcal{G}}}, \mathcal{A}_{\widehat{\mathcal{G}}}\}$ denotes the symmetrized graph of $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_{\mathcal{G}}, \mathcal{A}_{\mathcal{G}}\}$. By the definition of the union graph of symmetrized graphs we have $\sum_{i=mh}^{(m+1)h-1} \widehat{\mathcal{G}}(i) = \widehat{\mathcal{G}}_{mh}^h$, which in turn gives

$$\sum_{i=mh}^{(m+1)h-1} \hat{L}(i) = \sum_{i=mh}^{(m+1)h-1} L_{\widehat{\mathcal{G}}(i)} = L_{\sum_{i=mh}^{(m+1)h-1} \widehat{\mathcal{G}}(i)} = L_{\widehat{\mathcal{G}}_{mh}^h}.$$

Thus, by (12.49), (12.50), (12.51), and Theorem 12.1 we have

$$\begin{aligned}
\mathbf{E}(V[(m+1)h]) \quad &\leq \quad (1 - 2\lambda_{mh}^h a((m+1)h) + a^2((m+1)h)M_h)\mathbf{E}[V(mh)] \\
&\quad + N_h \sum_{i=mh}^{(m+1)h-1} a^2(i) \\
&= \quad (1 - 2(\inf_{m\geq 0} \lambda_{mh}^h)a((m+1)h) + a^2((m+1)h)M_h)\mathbf{E}[V(mh)] \\
&\quad + N_h \sum_{i=mh}^{(m+1)h-1} a^2(i), \ \forall\, m \geq m_0.
\end{aligned} \tag{12.52}$$

Noticing that $\inf_{m\geq 0} \lambda_{mh}^h > 0$,

$$\sum_{m=0}^{\infty} a(mh) \geq \frac{1}{h} \sum_{m=0}^{\infty} \sum_{i=mh}^{(m+1)h-1} a(i) = \sum_{t=0}^{\infty} a(t) = \infty,$$

and

$$\sum_{i=mh}^{(m+1)h-1} a^2(i) \to 0, \ m \to \infty,$$

similarly to (12.13), (12.14), and (12.15), by (12.52) and Lemma 12.A.1 we get $E[V(mh)] \to 0$, $m \to \infty$.

Therefore, for any given $\epsilon > 0$, there is $m_1 > 0$ such that

$$\mathbf{E}[V(mh)] \leq \epsilon, \ \forall\, m \geq m_1, \tag{12.53}$$

and

$$a^2(t) < \epsilon, \ \forall\, t \geq m_1 h. \tag{12.54}$$

Let $m_t = \lfloor \frac{t}{h} \rfloor$. Then, for any given $t \geq m_1 h$, we have $m_t \geq m_1$ and

$$0 \leq t - m_t h \leq h. \tag{12.55}$$

From (12.46) and the definition of $V(t)$ we have

$$\mathbf{E}[V(n+1)] \quad \leq \quad \widetilde{\phi}(n+1,k)\mathbf{E}[V(k)] + K_h \sum_{i=k}^{n} \widetilde{\phi}(n,i)a^2(i), \ \forall\, k \geq 0, \tag{12.56}$$

where $K_h = \sup_{t\geq 0} \|D_{\mathcal{G}(t)}\|_2^2 \|I - J\|_2^2 \sigma_W$, $\widetilde{\phi}(n,i) = \prod_{j=i}^{n-1}(1 - 2\lambda_2(\widehat{L}_{\mathcal{G}(j)})a(j) + a^2(j)\|L_{\mathcal{G}(j)}\|_2^2)$, $i = 0, 1, ..., n-1$, $n = 1, 2, ...$; $\widetilde{\phi}(i,i) = 1$, $i = 0, 1, ...$ Thus there exists $\gamma \geq 1$ such that $|\widetilde{\phi}(n,i)| \leq \gamma^{n-i}$, $\forall n \geq i \geq 0$. This, together with (12.55), (12.53), (12.54), and (12.56), gives

$$\begin{aligned}
\mathbf{E}[V(t+1)] \quad &\leq \quad \widetilde{\phi}(t+1, m_t h)\mathbf{E}[V(m_t h)] + K_h \sum_{i=m_t h}^{t} \widetilde{\phi}(t,i)a^2(i) \\
&\leq \quad \gamma^h \epsilon + \gamma^h K_h \sum_{i=m_t h}^{t} a^2(i) \\
&\leq \quad \gamma^h (1 + K_h(h+1))\epsilon, \ \forall\, t \geq m_1 h.
\end{aligned}$$

Hence (12.45) follows from the arbitrariness of ϵ. □

Theorem 12.9. *Apply the protocol (12.42) to system (12.3), (12.43). For any given* $\{\mathcal{G}(t), t = 0, 1, ...\} \in \Gamma_1$, *if there is an integer* $h > 0$ *such that* $\inf_{m\geq 0} \lambda_{mh}^h > 0$ *and Assumptions A3)–A4) hold, then, for any* $W \in \mathcal{S}$,

$$\lim_{t\to\infty} \mathbf{E}[x_i(t) - \widetilde{x}_*]^2 = 0, \ i = 1, 2, ..., N, \ \forall\, X(0) \in \mathcal{R}^N, \tag{12.57}$$

where $\widetilde{x}_*$ *is a r.v. depending on* W, $X(0)$, *and* $\{\mathcal{G}(t), t = 0, 1, ...\}$ *and satisfying* $E(\widetilde{x}_*) = \frac{1}{N}\sum_{j=1}^{N} x_j(0)$ *and* $\mathbf{Var}(\widetilde{x}_*) < \infty$, *that is, (12.42) is an asymptotically unbiased mean square average-consensus protocol.*

Proof. By (12.44), similarly to (12.29), we have

$$\frac{1}{N}\sum_{j=1}^{N} x_j(n) = \frac{1}{N}\sum_{j=1}^{N} x_j(0) + \frac{1}{N}\mathbf{1}^T \sum_{t=0}^{n-1} a(t) D_{\mathcal{G}(t)} W(t). \tag{12.58}$$

Since $W \in \mathcal{S}$, $\sup_{t\geq 0} \|D_{\mathcal{G}(t)}\|_2^2 < \infty$, and $\sum_{t=0}^{\infty} a^2(t) < \infty$, $\sum_{t=0}^{n} a(t) D_{\mathcal{G}(t)} W(t)$ is convergent in mean square. Hence, by Theorem 12.8, similarly to the proof of Theorem 12.4, we have (12.57). □

Remark 12.2. *Differently from the randomly time-varying communication link failures considered in [37], here the network topology may change continuously, and to ensure mean square average-consensus, we do not need additional distribution conditions on the events of the link failures and creations.*

Theorem 12.10. *Apply the protocol (12.42) to system (12.3), (12.43). For any given* $\{\mathcal{G}(t), t = 0, 1, ...\} \in \Gamma_1$, *if there is an integer* $h > 0$ *such that* $\inf_{m\geq 0} \lambda_{mh}^{h} > 0$ *and Assumptions A3)–A4) hold, then, for any* $W \in \widetilde{\mathcal{S}}'$,

$$\lim_{t\to\infty} x_i(t) = \widetilde{x}_* \; a.s. \; i = 1, 2, ..., N, \; \forall \, X(0) \in \mathcal{R}^N, \tag{12.59}$$

where $\widetilde{x}_*$ *is given by Theorem 12.9, that is, (12.42) is an almost sure strong consensus protocol.* □

The proof of Theorem 12.10 needs the following two lemmas.

Lemma 12.1. *For a sequence of digraphs* $\{G(t), t = 0, 1, ...\}$, *the following three statement are equivalent:*

(i) There is an integer $h > 0$ *such that* $\inf_{m\geq 0} \lambda_{mh}^{h} > 0$.

(ii) There is an integer $h > 0$ *such that* $\inf_{k\geq 0} \lambda_{k}^{h} > 0$.

(iii) There are integers $h > 0$ *and* $k_0 > 0$ *such that* $\inf_{m\geq 0} \lambda_{k_0+mh}^{h} > 0$.

Proof. $(ii) \Rightarrow (i)$ and $(ii) \Rightarrow (iii)$ are straightforward. It suffices to show that $(i) \Rightarrow (ii)$ and $(iii) \Rightarrow (i)$.

$(i) \Rightarrow (ii)$. Suppose that $\inf_{m\geq 0} \lambda_{mh_0}^{h_0} > 0$ for some integer $h_0 > 0$. For any given $k \geq 0$, set $n_k = \lceil \frac{k}{h_0} \rceil$. Then $L_{\sum_{j=k}^{k+2h_0-1} \widehat{G}(j)} - L_{\sum_{i=n_k h_0}^{(n_k+1)h_0-1} \widehat{G}(i)}$ is the Laplacian matrix of the union of graphs $\sum_{j=k}^{n_k h_0-1} \widehat{G}(j)$ and $\sum_{i=(n_k+1)h_0}^{k+2h_0-1}$. Thus $L_{\sum_{j=k}^{k+2h_0-1} \widehat{G}(j)} - L_{\sum_{i=n_k h_0}^{(n_k+1)h_0-1} \widehat{G}(i)}$ is positive semidefinite, which, together with Theorem 12.1, gives

$$\lambda_{k}^{2h_0} \geq \lambda_{n_k h_0}^{h_0} \geq \inf_{m\geq 0} \lambda_{mh}^{h} > 0, \; k = 0, 1, \ldots.$$

Thus $\inf_{k\geq 0} \lambda_{k}^{2h_0} > 0$.

$(iii) \Rightarrow (i)$. Suppose that $\inf_{m\geq 0} \lambda_{k_0+mh_0}^{h_0} > 0$ for some integers $k_0 > 0$ and $h_0 > 0$. Let $\overline{h} = k_0 + 2h_0$, $n_m = \lceil \frac{m\overline{h}-k_0}{h_0} \rceil$, $m = 1, 2, \ldots$. Then $L_{\sum_{i=m\overline{h}}^{(m+1)\overline{h}-1} \widehat{G}(i)} - L_{\sum_{j=k_0+n_m h_0}^{k_0+(n_m+1)h_0-1} \widehat{G}(j)}$ is positive semidefinite, which, together with Theorem 12.1, gives

$$\lambda_{m\overline{h}}^{\overline{h}} \geq \lambda_{k_0+n_m h_0}^{h_0} \geq \inf_{m\geq 0} \lambda_{k_0+mh_0}^{h_0} > 0, \; m = 1, 2, \ldots.$$

Noticing that $\lambda_0^{\overline{h}} \geq \lambda_{k_0}^{h_0} \geq \inf_{m\geq 0} \lambda_{k_0+mh_0}^{h_0} > 0$, we have

$$\inf_{m\geq 0} \lambda_{m\overline{h}}^{\overline{h}} \geq \inf_{m\geq 0} \lambda_{k_0+mh_0}^{h_0} > 0. \quad \square$$

Lemma 12.2. *Apply the protocol (12.42) to system (12.3), (12.43). For any given* $\{\mathcal{G}(t), t = 0, 1, \ldots\} \in \Gamma_1$, *if there are integers* $h > 0$ *and* $k_0 \geq 0$ *such that* $\inf_{m\geq 0} \lambda_{k_0+mh}^{h} > 0$ *and Assumptions A3)–A4) hold, then, for any* $W \in \widetilde{\mathcal{S}}'$,

$$\lim_{m\to\infty} V(k_0 + mh) = 0 \ a.s.$$

Proof. First, by Lemma 12.1 and Theorem 12.8 we have

$$\lim_{t\to\infty} \mathbf{E}[V(t)] = 0. \tag{12.60}$$

By $W \in \widetilde{\mathcal{S}}'$ there is a constant $\overline{N}_h > 0$ such that

$$\begin{aligned}
&\sup_{m\geq 0} \mathbf{E}[\|\overline{W}_{k_0+mh}^{h}\|_2^2 | \mathcal{F}^W(k_0 + mh - 1)] \\
&\quad \leq \overline{N}_h \sup_{t\geq 0, m\geq 0} \mathbf{E}[\|W(t+m)\|_2^2 | \mathcal{F}^W(t)] \sum_{i=k_0+mh}^{k_0+(m+1)h-1} a^2(i).
\end{aligned} \tag{12.61}$$

By (12.45), similarly to (12.47) and (12.49), we have

$$\delta[k_0 + (m+1)h] = \Phi(k_0 + (m+1)h, k_0 + mh)\delta(k_0 + mh) + \overline{W}_{k_0+mh}^{h}$$

and

$$\begin{aligned}
V[k_0 + (m+1)h] \ \leq \ & (1 - 2\lambda_{k_0+mh}^{h} a(k_0 + (m+1)h) + a^2(k_0 + (m+1)h)M_h)V(k_0 + mh) \\
& + (\overline{W}_{k_0+mh}^{h})^T \overline{W}_{k_0+mh}^{h} \\
& + 2\delta^T(k_0 + mh)\Phi^T(k_0 + (m+1)h, k_0 + mh)\overline{W}_{k_0+mh}^{h}.
\end{aligned} \tag{12.62}$$

Notice that $\{(V(k_0 + mh), \mathcal{F}^W(k_0 + mh - 1)), m = 0, 1, \ldots\}$, is an adapted sequence. Then, from (12.62) and (12.61) it follows that

$$\begin{aligned}
& \mathbf{E}(V[k_0 + (m+1)h] | \mathcal{F}^W(k_0 + mh - 1)) \\
\leq \ & (1 + a^2(k_0 + (m+1)h)M_h)\mathbf{E}[V(k_0 + mh)] \\
& + \overline{N}_h \sup_{t\geq 0, m\geq 0} \mathbf{E}[\|W(t+m)\|_2^2 | \mathcal{F}^W(t)] \sum_{i=k_0+mh}^{k_0+(m+1)h-1} a^2(i) \ \text{a.s.},
\end{aligned} \tag{12.63}$$

where M_h is given by (12.48). Since $\sum_{m=1}^{\infty} a^2(k_0 + mh) < \infty$ and $\sum_{m=0}^{\infty}\sum_{i=k_0+mh}^{k_0+(m+1)h-1} a^2(i) < \infty$, by (12.61), (12.63), and the nonnegative supermartingale convergence theorem we get that $V(k_0 + mh)$ converges a.s. as $m \to \infty$. Furthermore, by (12.60) we have $\lim_{m\to\infty} V(k_0 + mh) = 0$ a.s. □

Proof of Theorem 12.10. By Lemma 12.1 there exists $\widetilde{h} > 0$ such that $\inf_{m\geq 0} \lambda_{l+m\widetilde{h}}^{\widetilde{h}} \geq \inf_{k\geq 0} \lambda_k^{\widetilde{h}} > 0$, $l = 0, 1, ..., \widetilde{h} - 1$. Thus it follows from 12.2 that

$$\lim_{m\to\infty} V(l + m\widetilde{h}) = 0 \text{ a.s.}, \ l = 0, 1, \ldots, \widetilde{h} - 1.$$

This implies

$$\lim_{t\to\infty} V(t) = 0 \text{ a.s.} \tag{12.64}$$

Since $\{\sum_{t=0}^{n} a(t)D_{\mathcal{G}(t)}W(t), \mathcal{F}^W(n)\}$ is a martingale with $\sup_{n\geq 0} \mathbf{E}\|\sum_{t=0}^{n} a(t)D_{\mathcal{G}(t)} \times W(t)\|_2^2 < \infty$, by Theorem 7.6.10 of [26] we see that $\sum_{t=0}^{n} a(t)D_{\mathcal{G}(t)}W(t)$ converges almost surely as $n \to \infty$. This, together with (12.64) and (12.58), implies that $x_i(t)$, $i = 1, 2, ..., N$, converges almost surely as $t \to \infty$. Thus by Theorem 12.9 we get (12.59). □

Corollary 12.1. *Apply the protocol (12.42) to system (12.3), (12.43). For any given $\{\mathcal{G}(t), t = 0, 1, ...\} \in \Gamma_2$, if there is an integer $h > 0$ such that, for any $m \geq 0$, $\sum_{i=mh}^{(m+1)h-1} \mathcal{G}(i)$ contains a spanning tree and Assumptions A3)–A4) hold, then, for any $W \in \mathcal{S}$,*

$$\lim_{t\to\infty} \mathbf{E}[x_i(t) - \widetilde{x}_*]^2 = 0, \ i = 1, 2, ..., N, \ \forall\, X(0) \in \mathcal{R}^N.$$

Proof. Since $\sum_{i=mh}^{(m+1)h-1} \mathcal{G}(i)$, $m = 0, 1, ...$, has a spanning tree, $\widehat{\mathcal{G}}_{mh}^{h}$, $m = 0, 1, ...$, is strongly connected, which, together with Theorem 12.1, implies $\lambda_{mh}^{h} > 0$, $m = 0, 1, ...$ Furthermore, since $\{\mathcal{G}(t), t = 0, 1, ...\} \in \Gamma_2$, $|\{\lambda_{mh}^{h}, m = 0, 1, ...\}| < \infty$, and hence $\inf_{m\geq 0} \lambda_{mh}^{h} = \min_{m\geq 0} \lambda_{mh}^{h} > 0$. This, together with Theorem 12.9 and $\Gamma_2 \subset \Gamma_1$, completes the proof. □

Corollary 12.2. *Apply the protocol (12.42) to system (12.3), (12.43). For any given $\{\mathcal{G}(t), t = 0, 1, ...\} \in \Gamma_2$, if there is an integer $h > 0$ such that, for any $m \geq 0$, $\sum_{i=mh}^{(m+1)h-1} \mathcal{G}(i)$ contains a spanning tree and A3)–A4) hold, then, for any $W \in \mathcal{S}'$,*

$$\lim_{t\to\infty} x_i(t) = \widetilde{x}_* \ a.s. \ i = 1, 2, ..., N, \ \forall\, X(0) \in \mathcal{R}^N,$$

where $\widetilde{x}_*$ is given by Theorem 12.9, that is, (12.42) is an almost sure strong consensus protocol.

Proof. Similarly to Corollary 12.1, we can get $\inf_{m\geq 0} \lambda_{mh}^{h} > 0$. This, together with Theorem 12.10 and $\Gamma_2 \subset \Gamma_1$, leads to the desired conclusion. □

Theorems 12.8–12.10 are for the case of time-varying graph flows, whereas Corollaries 12.1–12.2 are for the special cases of switching graph flows, where the network switches among a finite number of digraphs and the condition that there is $h > 0$ such that $\inf_{m\geq 0} \lambda_{mh}^{h} > 0$ is equivalent to that there is $h > 0$ such that, for any $m \geq 0$, $\sum_{i=mh}^{(m+1)h-1} \mathcal{G}(i)$ contains a spanning tree, that is, $\{\mathcal{G}(i), i = mh, mh + 1, ..., (m + 1)h - 1\}$, $m = 0, 1, ...,$ are all jointly-containing-spanning-tree.

12.3 Adaptive Mean-Field Games for Large Population Coupled ARX Systems With Unknown Coupling Strength

Introduction

The research on multiagent dynamic games has a long history in the control community. A good survey of noncooperative dynamic games can be found in Basar and Olsder [38]. In recent years, the dynamic game theory gets new inspiration and renews its vitality in network control and multiagent systems. In the framework of dynamic games, a lot of researchers considered flow control, routing control, and multiagent cooperation problems [39,40]. For distributed multiagent systems, generally speaking, there is no centralized control station, and each agent has only limited sensing and communication ability, so control design is always required to be decentralized. In a decentralized control framework, the control input of each agent can only use the local state or, under certain circumstances, include those of others in its sensing/communication neighborhood.

Recently, Huang, Caines, and Malhamé did a pioneering work on decentralized stochastic games for a kind of individual-population interacting multiagent systems with mean-field coupling [41,42], which have wide application background in biological, engineering, and economic systems [43–46]. In this kind of systems, the number of agents is quite large. Each agent is driven by stochastic noises and interacts with all other agents via the population state average (PSA). The interactions between individual states and the PSA exist in both the dynamic equation and the cost function of every agent. For a given agent, the impact of any other single agent is so small that can be neglected; however, that of the overall population is significant enough for its evolution. Though the agents are coupled with the PSA, the PSA cannot be used for the individual control design, since it is unknown for any given agent. This is an essential difficulty of the decentralized control design for decentralized mean-field games. To overcome this difficulty, Huang, Caines, and Malhamé proposed the methodology called the Nash certainty equivalence (NCE) principle. In the NCE principle, the PSA is properly approximated by its mean-field approximation, a deterministic signal, which is then used for the individual control design instead of the PSA. This principle is similar in spirit to the well-known the certainty equivalence (CE) principle adopted in adaptive control, where the unknown parameters are estimated, and the estimates are used as the true parameters to construct the control laws.

For decentralized mean-field games, most of the relevant literature assumes precise dynamic models of agents. However, in real systems, there may be parametric uncertainties or unmodeled dynamics in agents' models due to various kinds of unknown or uncertain factors in the environment. Generally speaking, the parametric uncertainties can be divided into two categories: unknown local parameters, which contain the information of local environment, and unknown global parameters, which are shared by all agents. In this paper, we assume that the local dynamics of each agent is precisely known but the common coupling strength between the individual state and the PSA, which is a global parameter, is unknown. To eliminate the model uncertainties, each agent can exploit its learning ability to perfect its dynamic model by measured data step by step. By using an individual online learner or identifier, each agent uses its estimate for the coupling strength to construct its individual control law, which aims at optimizing its cost function. Therefore, the overall system emerges as a large population decentralized adaptive game. In this kind of adaptive games, there are two estimation processes. One is the estimation for the PSA, and the other is the identification for the unknown coupling strength. A key difficulty lies in that there is a product term of the unknown PSA and the unknown coupling strength in each agent's dynamic equation. So, if traditional identification algorithms were used, then the regression vector would contain the PSA as a component in each agent's identification algorithm. However, we know that the PSA is unavailable for each individual. Intuitively, the estimation signal for the PSA can be used to construct the identification algorithms instead of the PSA. Unfortunately, this may result in the coupling between the estimation process for the PSA and that for the unknown coupling strength. Decentralized adaptive games for individual–population interacting systems are considered firstly in Huang, Malhamé, and Caines [42] and Kizilkale and Caines [47]. In Kizilkale and Caines [47], the dynamic equations of agents are uncoupled, and the local dynamic parameters are unknown, whereas in Huang, Malhamé, and Caines [42], the dynamics of agents are coupled, but the precise value of the PSA is used in the identification algorithm. In brief, the coupling between the two estimation processes, which is a key difficulty in decentralized mean-field adaptive games, does not exist in Huang, Malhamé, and Caines [42] and Kizilkale and Caines [47]. To our best knowledge, up to now there is no relevant literature which involves the case where both the PSA and the coupling strength are unknown.

For decentralized adaptive mean-field games, there are some fundamental problems that have to be studied.

(1) Is the closed-loop system stable, that is, are the states of all agents kept bounded as time goes on? And if the answer is affirmative, can the stability be retained as the number N of agents increases to infinity?

(2) Is the estimate of the PSA strongly consistent or does the estimation error for the PSA converge to zero with respect to some metric almost surely as N tends to infinity? If the answer is affirmative, what is the convergence rate?

(3) Is the identification algorithm for the coupling strength strongly consistent or are the estimation errors bounded? If the estimation errors are bounded, can we ensure that the bound converges to zero as N tends to infinity and get the convergence rate?

(4) Is the designed decentralized control law asymptotically optimal almost surely, or is there an almost sure asymptotic Nash equilibrium? If the answer is affirmative, what is the convergence rate of the suboptimal cost function of each agent to the optimal value as N tends to infinity?

The large population decentralized adaptive mean-field game is essentially different from traditional adaptive control for single-agent systems [48,49], and the solutions to the convergence problems (1)–(4) cannot be found in the existing theoretical framework.

In this section, we consider the decentralized adaptive mean-field game for individual–population interacting stochastic multiagent systems. The dynamic equation of each agent is described by a discrete-time ARX model and coupled by terms of the PSA with unknown coupling strength. Each agent has a group tracking type cost function, also coupled by the PSA. Firstly, based on the NCE principle, the PSA is estimated by some deterministic signal. Secondly, the estimation of the PSA is used to construct the decentralized least square (LS) identification algorithm for the coupling strength. Finally, the estimates of the PSA and the coupling strength are both used to construct the decentralized control law based on the NCE and CE principles. By the stochastic Lyapunov method we analyze the decentralized LS algorithm, and then by probability limit theory, under mild conditions, we get the following convergence results of the closed-loop system: (i) The closed-loop system is stable almost surely, and the states of agents retain bounded as N tends to infinity. (ii) As N tends to infinity, the estimation error for the PSA converges to zero with rate $O(1/N)$ almost surely. (iii) As N tends to infinity, the identification error for the unknown coupling strength converges to zero with rate $O(1/\sqrt{N})$. (iv) The decentralized control law designed is an almost sure asymptotic Nash equilibrium, and the cost function of each agent is almost surely asymptotically optimal with convergence rate $O(1/N)$, given that all other agents also employ the strategy specified by the asymptotic Nash equilibrium.

We will use the following notation. For a family $\{\xi_\lambda, \lambda \in \Lambda\}$ of real-valued r.v.s, $\sigma(\xi_\lambda, \lambda \in \Lambda)$ denotes the σ-algebra $\sigma(\{\xi_\lambda \in B\}, B \in \mathcal{B}, \lambda \in \Lambda)$, where $\mathcal{B}$ denotes the one-dimensional Borel sets. For a sequence $\{\mathcal{F}_t, t \geq 0\}$ of nondecreasing σ-algebras and a sequence $\{\xi(t), t \geq 0\}$ of r.v.s, we say that $\xi(t)$ is adapted to $\mathcal{F}_t$ or that $\{\xi(t), \mathcal{F}_t\}$ is an adapted sequence if $\xi(t)$ is $\mathcal{F}_t$-measurable for all $t \geq 0$.

Problem Formulation

We consider a system of N agents denoted by $\mathbf{S}^N$. The dynamic equation of agent i is given by

$$x_i^N(t+1) = g_i(x_i^N(t), t) + u_i^N(t) + \alpha \overline{x}_N(t) + \omega_i(t+1), \ t = 0, 1, ..., \ 1 \le i \le N, \tag{12.65}$$

where $x_i^N \in \mathcal{R}$ and $u_i^N \in \mathcal{R}$ are the state and control input, respectively, $\overline{x}_N(t) \triangleq \frac{1}{N} \times \sum_{j=1}^N x_j^N(t)$ is the PSA, $\omega_i(t) \in \mathcal{R}$ is the random noise, $g_i(\cdot, \cdot) : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ is a known Borel-measurable function, and $\alpha \in \mathcal{R}$ is the unknown coupling parameter satisfying $|\alpha| < 1$. Note that model (12.65) is just the scalar version of the dynamic model considered in [50], but here the coupling strength α is unknown.

For model (12.65), we have the following assumptions:

A1) $\{\{\omega_i(t), \mathcal{F}_t^i\}, 1 \le i \le N, N \ge 1\}$ is a family of independent martingale difference sequences defined on a probability space $(\Omega, \mathcal{F}, P)$ with the following properties: there exist constants $\sigma > 0$ and $\beta > 2$ such that

$$\sup_{t \ge 0} \mathbf{E}[|\omega_i(t)|^\beta | \mathcal{F}_{t-1}^i] < \infty \text{ a.s.},$$

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^T [\omega_i(t)]^2 = \sigma^2 \text{ a.s.},$$

where $\mathcal{F}_t^i \triangleq \sigma(\omega_i(s), 0 \le s \le t)$.

A2) $\{x_i^N(0), 1 \le i \le N, N \ge 1\}$ is independent of $\{\{\omega_i(t), \mathcal{F}_t^i\}, i \ge 1\}$ with common mathematical expectation $x_0 \triangleq \mathbf{E}(x_1^N(0)) < \infty$.

The cost function of agent i is given by

$$J_i^N(u_i^N, u_{-i}^N) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^T [x_i^N(t+1) - \Phi(t, \overline{x}_N(t))]^2, \tag{12.66}$$

where $u_{-i}^N = (u_1^N, ..., u_{i-1}^N, u_{i+1}^N, \cdots, u_N^N)$, and $\Phi(t, x) : [0, \infty) \times \mathcal{R} \to \mathcal{R}$ is a Borel-measurable function.

With regards to the cost function, we involve the following assumptions in the closed-loop analysis.

A3) The solution of the nonlinear iteration $x(t+1) = \Phi(t, x(t))$ with $x(0) = x_0$ satisfies

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^T x^2(t) < \infty.$$

A4) The solution of the nonlinear iteration $x(t+1)=\Phi(t,x(t))$ with $x(0)=x_0$ satisfies

$$\lim_{T\to\infty}\frac{x^2(T)}{1+\sum_{t=0}^{T-1}x^2(t)}=0.$$

We can easily verify that if $\Phi(t,x)=x$, $t\in[0,\infty)$, $x\in\mathcal{R}$, then both A3) and A4) hold.

For convenience of citation, for agent i, we denote the global-measurement-based admissible control set by

$$\mathcal{U}_{g,i}^N \triangleq \{u_i^N \mid u_i^N(t) \text{ is adapted to } \sigma(\cup_{j=1}^N \sigma(x_j^N(s), 0\le s\le t))\},$$

the local-measurement-based admissible control set by

$$\mathcal{U}_{l,i}^N \triangleq \{u_i^N \mid u_i^N(t) \text{ is adapted to } \sigma(x_i^N(s), 0\le s\le t)\},$$

and the admissible control set by $\mathcal{U}_i^N$. The so-called decentralized game means that agent i synthesizes u_i^N only based on the local measurement (i.e. $\mathcal{U}_i^N=\mathcal{U}_{l,i}^N$) to minimize its cost function $J_i^N(u_i^N,u_{-i}^N)$. We denote a control group of the sequence $\mathbf{S}^N$ of systems by $\mathbf{U}^N=\{u_i^N, 1\le i\le N\}$ and its associated cost function group by $\mathbf{J}^N=\{J_i^N(u_i^N,u_{-i}^N), 1\le i\le N\}$. To characterize the asymptotic optimality of the decentralized control law with respect to the stochastic cost functions, we introduce the concept of almost sure asymptotic Nash equilibrium given in Li and Zhang [50].

Definition 12.1. *For system (12.65), a sequence of control groups $\{\mathbf{U}^N=\{u_i^N,1\le i\le N\}, N\ge 1\}$ is called an almost sure asymptotic Nash equilibrium with respect to the associated sequence of cost function groups $\{\mathbf{J}^N=\{J_i^N,1\le i\le N\}, N\ge 1\}$ if there exists a sequence of nonnegative r.v.s $\{\epsilon_N(\omega), N\ge 1\}$ on the probability space $(\Omega,\mathcal{F},P)$ such that $\epsilon_N\to 0$ a.s. as $N\to\infty$, and for sufficiently large N,*

$$J_i^N(u_i^N,u_{-i}^N)\le \inf_{v_i\in\mathcal{U}_{g,i}^N} J_i^N(v_i,u_{-i}^N)+\epsilon_N,\ a.s.,\ i=1,2,...,N. \tag{12.67}$$

By Theorem 2.1 of [50] we know that $\inf_{v_i\in\mathcal{U}_{g,i}^N} J_i^N(v_i,u_{-i}^N)=\sigma^2$. We will further design a decentralized control law $\{\mathbf{U}^N, N\ge 1\}$ such that the closed-loop system satisfies

$$J_i^N(u_i^N,u_{-i}^N)\le\sigma^2+o(1),\ N\to\infty,\ \text{a.s.},$$

that is, the sequence of control groups $\{\mathbf{U}^N, N\ge 1\}$ is an almost sure asymptotic Nash equilibrium.

Control Design

Firstly, we make a review of the results with known coupling strength.

For the centralized control law design, the control of agent i depends on the PSA $\overline{x}_N$, whereas for the design of the decentralized control law, the PSA is unknown. If the coupling strength α is known, then we may use the NCE principle to design the decentralized control law. Firstly, we construct an estimate $f(t)$ of the PSA with the following property: if every agent takes $f(t)$ as the estimate of the PSA and, according to $f(t)$, makes the optimal decision, then the expectation of the closed-loop PSA is just $f(t)$ or convergent to it as N increases to infinity. Secondly, if the $f(t)$ with the above property indeed exists, then we can construct the decentralized control law by using $f(t)$ instead of $\overline{x}_N(t)$.

Based on the NCE principle, we now design the decentralized control law.

The auxiliary equation of agent i is given by

$$\widehat{x}_i^N(t+1) = \quad g_i(\widehat{x}_i^N(t), t) + \widehat{u}_i^N(t) + \alpha f(t) + \omega_i(t+1), \ t \geq 0, \ i = 1, 2, ..., N, \tag{12.68}$$

with a tracking-type cost function

$$J_i^N(\widehat{u}_i^N) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} [\widehat{x}_i^N(t+1) - \Phi(t, f(t))]^2.$$

In this case, the optimal control obviously is

$$\widehat{u}_i^N(t) = \Phi(t, f(t)) - g_i(\widehat{x}_i^N(t), t) - \alpha f(t). \tag{12.69}$$

Substituting control (12.69) into the model (12.68), we have

$$\mathbf{E}(\widehat{x}_i^N(t+1)) = \Phi(t, f(t)), \ \mathbf{E}(\widehat{x}_i^N(0)) = x_0. \tag{12.70}$$

As mentioned before, the mathematical expectation of the closed-loop PSA ought to be $f(t)$, that is,

$$\frac{1}{N} \sum_{j=1}^{N} \mathbf{E}(\widehat{x}_j^N(t)) = f(t), \ t \geq 0. \tag{12.71}$$

Therefore, the unique solution of the auxiliary system (12.70) and (12.71) can be used as the estimate of the PSA. We denote it by $f^*(t)$, which is iteratively given by

$$f^*(t+1) = \Phi(t, f^*(t)), \ t \geq 0, \ f^*(0) = x_0. \tag{12.72}$$

By (12.69) and the NCE principle the control law for agent i can be taken as

$$u_i^0(t) = \Phi(t, f^*(t)) - g_i(x_i^N(t), t) - \alpha f^*(t). \tag{12.73}$$

Here and hereafter, we omit the superscript N of $u_i^{0^N}(t)$ for conciseness of expression. Comparing (12.73) with the centralized control law, we can see that $\overline{x}_N$ in (12.73) is replaced by f^* for control design.

As shown in [50], we can prove the asymptotic consistency of the estimate f^* for the PSA and the stability and asymptotic optimality of the closed-loop system under the control law (12.73). We have the following theorems [50].

Lemma 12.3. *For system (12.65), if Assumptions A1)–A2) hold, then under the control law (12.73), the closed-loop system has the following properties:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}[\xi_N(t)]^2 = \frac{\sigma^2}{N(1-\alpha^2)} \ a.s., \tag{12.74}$$

where

$$\xi_N(t) = \overline{x}_N(t) - f^*(t) \tag{12.75}$$

is the estimation error for the PSA.

Lemma 12.4. *For system (12.65), if Assumptions A1)–A3) hold, then under the control law (12.73), the closed-loop system satisfies*

$$\sup_{N\geq 1}\max_{1\leq i\leq N}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}[x_i^N(t)]^2 < \infty \ a.s.$$

Lemma 12.5. *For system (12.65) with cost function (12.66), if Assumptions A1)–A2) hold and there exists a constant $\gamma > 0$ such that $|\Phi(t,x) - \Phi(t,y)| \leq \gamma|x-y|$, $\forall\, x, y \in \mathcal{R}$, $t \geq 0$, then under the control law (12.73), the associated cost function group satisfies*

$$J_i^N(u_i^0, u_{-i}^0) \leq \sigma^2 + \epsilon_N \ a.s., \ i = 1, 2, ..., N, \tag{12.76}$$

where

$$\epsilon_N = \frac{2\sigma^2(\alpha^2+\gamma^2)}{N(1-\alpha^2)}.$$

If the coupling strength α is unknown, then the control law (12.73) is unavailable. Naturally, we might think that based on the model (12.65), agent i could use the following recursive LS algorithm to estimate α:

$$\begin{aligned}\overline{\alpha}_i(t+1) &= \overline{\alpha}_i(t) + (1+P(t)\overline{x}_N^2(t))^{-1}\{P(t)\overline{x}_N(t)[x_i^N(t+1) - u_i^N(t) - g_i(x_i^N(t), t)\\ &\quad - \overline{\alpha}_i(t)\overline{x}_N(t)]\}, \qquad (12.77)\\ P(t+1) &= P(t) - \frac{P^2(t)\overline{x}_N^2(t)}{1+\overline{x}_N^2(t)P(t)}. \qquad (12.78)\end{aligned}$$

Then by the CE principle, instead of α, the estimation $\overline{\alpha}_i(t)$ could be used to construct the control law

$$u_i^N(t) = \Phi(t, f^*(t)) - g_i(x_i^N(t), t) - \overline{\alpha}_i(t)f^*(t), \ i = 1, 2, ..., N. \qquad (12.79)$$

However, the control law (12.77)–(12.79) is not decentralized due to the use of the PSA in the identification algorithm. Since the PSA $\overline{x}_N(t)$ is unknown for agent i, we use $f^*(t)$, which is the estimation of $\overline{x}_N(t)$ based on the NCE principle, to construct the identification algorithm of agent i:

$$\begin{aligned}\alpha_i(t+1) &= \alpha_i(t) + (1+P(t)f^{*^2}(t))^{-1}\{P(t)f^*(t)[x_i^N(t+1) - u_i^N(t) - g_i(x_i^N(t), t)\\ &\quad - \alpha_i(t)f^*(t)]\}, \ \alpha_i(0) = \alpha_0, \qquad (12.80)\\ P(t+1) &= P(t) - \frac{P^2(t)f^{*^2}(t)}{1+f^{*^2}(t)P(t)}, \ P(0) = P_0, \qquad (12.81)\end{aligned}$$

where α_0 and $P(0) = P_0$ are initial conditions to be designed, and $f^*(t)$ is computed off-line by (12.72). The identification algorithm (12.80)–(12.81) is decentralized, since it only uses the local state and input of each agent. Then by the CE principle we use the estimate $\alpha_i(t)$ to construct the control law:

$$u_i^N(t) = \Phi(t, f^*(t)) - g_i(x_i^N(t), t) - \alpha_i(t)f^*(t), \ i = 1, 2, ..., N. \qquad (12.82)$$

We can see that the control law (12.80)–(12.82) is decentralized and designed based on both the NCE and the CE principles.

Remark 12.3. *Here, a decentralized two-level control scheme is used for adaptive mean-field adaptive games. On the high level, the PSA is estimated based on the NCE principle. On the low level, the coupling strength is identified based on the decentralized LS algorithms and the estimate of the PSA. The decentralized control law is constructed by combining the NCE and CE principles.*

Closed-Loop Analysis

In this section, we analyze the identification algorithm, stability and optimality of the closed-loop system, and the consistency of the estimates for the PSA and the coupling strength.

From the model (12.65) and (12.80), we get

$$\widetilde{\alpha}_i(t+1) = \left(1 - \frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}\right)\widetilde{\alpha}_i(t) - \frac{P(t)f^*(t)(\alpha\xi_N(t)+\omega_i(t+1))}{1+P(t)f^{*^2}(t)}, \tag{12.83}$$

where $\widetilde{\alpha}_i(t) \overset{\triangle}{=} \alpha - \alpha_i(t)$ is the estimation error for the coupling strength α, and $\xi_N(t)$ is the estimation error for the PSA given by (12.75).

Denote $V_i(t+1) = \widetilde{\alpha}_i^2(t+1)P^{-1}(t+1)$, $r(T) = e + \sum_{t=0}^{T} f^{*^2}(t)$. From (12.81) we know that

$$P^{-1}(t+1) = P^{-1}(t) + f^{*^2}(t). \tag{12.84}$$

Then summing this equation from both sides, we have

$$P^{-1}(t+1) = P_0^{-1} + \sum_{k=0}^{t} f^{*2}(k). \tag{12.85}$$

From (12.83) and (12.84) we have

$$\begin{aligned} V_i(t+1) \quad = \quad & V_i(t) - \frac{\widetilde{\alpha}_i^2(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)} - 2\frac{\widetilde{\alpha}_i(t)f^*(t)(\alpha\xi_N(t)+\omega_i(t+1))}{1+P(t)f^{*^2}(t)} \\ & + \frac{P(t)f^{*^2}(t)(\alpha\xi_N(t)+\omega_i(t+1))^2}{1+P(t)f^{*^2}(t)}. \end{aligned} \tag{12.86}$$

For the identification algorithm (12.80) and (12.81), we have the following results, which are important for the closed-loop analysis of the decentralized control law.

Theorem 12.11. *If Assumption A1) holds and*

$$\limsup_{T\to\infty} \frac{f^{*^2}(T)}{P_0^{-1} + \sum_{t=0}^{T-1} f^{*^2}(t)} < C \textit{ for some } C > 0, \tag{12.87}$$

then the identification algorithm (12.80)–(12.81) has the following properties:

(i)

$$V_i(T+1)+\sum_{t=0}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))^2}{1+P(t)f^{*^2}(t)} \le \alpha^2\frac{2C+1}{C+1}\sum_{t=0}^{T}\xi_N^2(t)+o\left(\sum_{t=0}^{T}\xi_N^2(t)\right)$$
$$+o\left(\sum_{t=0}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))^2}{1+P(t)f^{*^2}(t)}\right)$$
$$+O(\ln r(T))\ a.s., \tag{12.88}$$

(ii)

$$\widetilde{\alpha}_i^2(n+1)\le\alpha^2\frac{2C+1}{C+1}\frac{\sum_{t=0}^{T}\xi_N^2(t)}{r(T)}+o\left(\frac{\sum_{t=0}^{T}\xi_N^2(t)}{r(T)}\right)+O\left(\frac{\ln r(T)}{r(T)}\right)\ a.s.,$$
$$i=1,2,...,N, \tag{12.89}$$

(iii)

$$\sum_{t=0}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))^2}{1+P(t)f^{*^2}(t)}$$
$$\le \frac{\alpha^2(2C+1)}{(1-\delta)(C+1)}\sum_{t=0}^{T}\xi_N^2(t)$$
$$+o\left(\sum_{t=0}^{T}\xi_N^2(t)\right)+O(\ln r(T))\ a.s.,\ \forall\,\delta\in(0,1),\ i=1,2,...,N. \tag{12.90}$$

Proof. From (12.87) we have

$$\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}\le\frac{C}{1+C},\ \forall\, t\ge t_1,\ \text{for some } t_1>0. \tag{12.91}$$

Then summating the both sides of (12.86) from $t=t_1$ to $t=T$, we get

$$V_i(T+1)+\sum_{t=t_1}^{T}\frac{\widetilde{\alpha}_i^2(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}+2\sum_{t=t_1}^{T}\frac{\widetilde{\alpha}_i(t)f^*(t)\alpha\xi_N(t)}{1+P(t)f^{*^2}(t)}+\sum_{t=t_1}^{T}\frac{\alpha^2\xi_N^2(t)}{1+P(t)f^{*^2}(t)}$$
$$=V_i(t_1)+\sum_{t=t_1}^{T}\frac{\alpha^2\xi_N^2(t)}{1+P(t)f^{*^2}(t)}-2\sum_{t=t_1}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))\omega_i(t+1)}{1+P(t)f^{*^2}(t)}$$

$$+2\sum_{t=t_1}^{T}\frac{\alpha\xi_N(t)\omega_i(t+1)}{1+P(t)f^{*^2}(t)}+2\sum_{t=t_1}^{T}\frac{P(t)f^{*^2}(t)(\alpha\xi_N(t)+\omega_i(t+1))^2}{1+P(t)f^{*^2}(t)}. \tag{12.92}$$

From this equation, (12.87), and Lemma 12.A.4 we have

$$\begin{aligned}
&V_i(T+1)+\sum_{t=t_1}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))^2}{1+P(t)f^{*^2}(t)}\\
&\quad\le V_i(t_1)+\alpha^2\frac{2C+1}{C+1}\sum_{t=0}^{T}\xi_N^2(t)+O\left(\left(\sum_{t=t_1}^{T}\frac{(\widetilde{\alpha}_i(t)f^*(t)+\alpha\xi_N(t))^2}{1+P(t)f^{*^2}(t)}\right)^{1/2+\epsilon}\right)\\
&\quad\quad+O\left(\left(\sum_{t=0}^{T}\xi_N^2(t)\right)^{1/2+\epsilon}\right)+2\sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)\omega_i^2(t+1)}{1+P(t)f^{*^2}(t)}.
\end{aligned} \tag{12.93}$$

From (12.85) we have

$$\sum_{t=0}^{T}\frac{P(t)f^{*2}(t)}{1+P(t)f^{*^2}(t)} = \sum_{t=0}^{T}\frac{f^{*2}(t)}{P_0^{-1}+\sum_{k=0}^{t}f^{*^2}(k)}, \tag{12.94}$$

and then

$$\begin{aligned}
\sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)} &= \sum_{t=0}^{T}\frac{P^{-1}(t+1)-P^{-1}(t)}{P^{-1}(t+1)}\\
&\le \sum_{t=0}^{T}\int_{P^{-1}(t)}^{P^{-1}(t+1)}\frac{dx}{x}\\
&= \ln P^{-1}(T+1)+\ln P_0 = O(\ln r(T)).
\end{aligned} \tag{12.95}$$

Denote $\mathcal{F}_t=\sigma(\cup_{j=1}^{N}\mathcal{F}_t^j)$. For any given $\nu\in(2,\min\{\beta,4\}]$, by the Cramér–Rao inequality we have

$$\begin{aligned}
&\sup_{t\ge 0}\mathbf{E}[|\omega_i^2(t+1)-\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t)|^{\nu/2}|\mathcal{F}_t]\\
&\quad\le\sup_{t\ge 0}\mathbf{E}[|\omega_i(t+1)|^{\nu}|\mathcal{F}_t]+\sup_{t\ge 0}E[|\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t)|^{\nu/2}|\mathcal{F}_t],
\end{aligned}$$

which, together with Assumption (A1) and Lyapunov inequality, leads to

$$\begin{aligned}
&\sup_{t\geq 0}\mathbf{E}[|\omega_i^2(t+1)-\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t)|^{\nu/2}|\mathcal{F}_t]\\
\leq\ &\sup_{t\geq 0}\mathbf{E}[|\omega_i(t+1)|^{\nu}|\mathcal{F}_t]+\sup_{t\geq 0}|\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t)|^{\nu/2}\\
\leq\ &2\sup_{t\geq 0}\mathbf{E}[|\omega_i(t+1)|^{\nu}|\mathcal{F}_t]\\
\leq\ &2(\sup_{t\geq 0}\mathbf{E}[|\omega_i(t+1)|^{\beta}|\mathcal{F}_t^i])^{\nu/\beta}<\infty\text{ a.s.}
\end{aligned}\tag{12.96}$$

Then by Lemma 12.A.4, (12.94), and (12.95), for any given $\epsilon>0$, noting that $0\leq\frac{f^{*^2}(t)}{P_0^{-1}+\sum_{k=0}^{t}f^{*^2}(k)}\leq 1$, we have

$$\begin{aligned}
\sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)\omega_i^2(t+1)}{1+P(t)f^{*^2}(t)} &= \sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}(\omega_i^2(t+1)-\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t))\\
&\quad+\sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t)\\
&\leq \sup_{t\geq 0}\mathbf{E}(\omega_i^2(t+1)|\mathcal{F}_t^i)\sum_{t=0}^{T}\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}\\
&\quad+O\left(\left(\sum_{t=0}^{T}\left(\frac{P(t)f^{*^2}(t)}{1+P(t)f^{*^2}(t)}\right)^{\nu/2}\right)^{2/\nu+\epsilon}\right)\\
&= O\left(\sum_{t=0}^{T}\frac{f^{*^2}(t)}{P_0^{-1}+\sum_{k=0}^{t}f^{*^2}(k)}\right)+O(1)\\
&= O(\ln r(T)),
\end{aligned}\tag{12.97}$$

which, together with (12.93), leads to (i) and (iii). Combining (i) and (iii), we get (ii). □

Remark 12.4. *By (12.75) the model (12.65) can be rewritten as*

$$\widehat{x}_i^N(t+1)=g_i(\widehat{x}_i^N(t),t)+\widehat{u}_i^N(t)+\alpha f^*(t)+\alpha\xi_N(t)+\omega_i(t+1),\ t\geq 0,\ i=1,2,...,N.$$

So (12.80), (12.81), and (12.82) can be viewed as the identification algorithm and adaptive control law for the model

$$\widehat{x}_i^N(t+1)=g_i(\widehat{x}_i^N(t),t)+\widehat{u}_i^N(t)+\alpha f^*(t)+\omega_i(t+1),\ t\geq 0,\ i=1,2,...,N,$$

with $\alpha\xi_N(t)$ as the unmodeled dynamics. We can see that $\alpha\xi_N(t)$ contains the states of all other agents due to decentralized information pattern; the conditions on unmodeled dynamics used in robust adaptive control [51,52] can not be used here.

Substituting the control (12.82) into the model (12.65), we get the closed-loop equation of agent i

$$x_i^N(t+1) = \widetilde{\alpha}_i(t) f^*(t) + \alpha \xi_N(t) + \Phi(t, f^*(t)) + \omega_i(t+1). \tag{12.98}$$

Summing this equation for $i = 1, 2, ..., N$, by (12.72) we know that $\xi_N(t)$ satisfies the following recursive equation:

$$\xi_N(t+1) \quad = \quad \alpha \xi_N(t) + f^*(t) \frac{1}{N} \sum_{j=1}^{N} \widetilde{\alpha}_j(t) + \frac{1}{N} \sum_{j=1}^{N} \omega_j(t+1). \tag{12.99}$$

From (12.83) and (12.99) we can see that the dynamic equation (12.99) of the estimation error and the dynamic equation (12.83) of the identification error are coupled together. The main result of this paper is the following:

Theorem 12.12. *If Assumptions (A1)–(A4) hold, then for system (12.65), under the control (12.72), (12.80), (12.81), and (12.82), we have:*

(i) The estimate for PSA is asymptotically consistent:

$$\limsup_{T \to \infty} \|\xi_N\|_T^2 = O(1/N) \ a.s., \tag{12.100}$$

where $\|\xi_N\|_T = \sqrt{\frac{1}{T} \sum_{t=0}^{T} \xi_N^2(t)}$.

(ii) The closed-loop system is almost surely uniformly stable:

$$\sup_{N \geq 1} \max_{1 \leq i \leq N} \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} [x_i^N(t)]^2 < \infty \ a.s. \tag{12.101}$$

(iii) Furthermore, if there exists $\gamma > 0$ such that, for any $x, y \in \mathcal{R}$ and $t \geq 0$, we have $|\Phi(t,x) - \Phi(t,y)| \leq \gamma |x - y|$, then $\{U^N = \{u_i(t), 1 \leq i \leq N\}, N \geq 1\}$ is an almost sure asymptotic Nash equilibrium with respect to the associated sequence of cost function groups, and the cost function of each agent is almost surely asymptotically optimal with the convergence rate $O(N^{-1})$ given that all other agents also employ the strategy specified by the asymptotic Nash equilibrium:

$$\max_{1 \leq i \leq N} J_i^N(u_i^N, u_{-i}^N) \leq \sigma^2 + \frac{2\sigma^2(\gamma^2 + \alpha^2)}{N(1 - \alpha^2)} \ a.s. \tag{12.102}$$

Proof. Take positive real numbers $\epsilon \in (0, \frac{1-\alpha^2}{2\alpha^2})$ and $\delta \in (0, 1-\alpha^2(2\epsilon+1))$. From Assumption A4) we have that

$$\limsup_{T\to\infty} \frac{f^{*^2}(T)}{P_0^{-1} + \sum_{t=0}^{T-1} f^{*^2}(t)} < \epsilon, \tag{12.103}$$

and similarly to (12.91), we have

$$\frac{P(t) f^{*^2}(t)}{1 + P(t) f^{*^2}(t)} \le \frac{\epsilon}{1+\epsilon}, \ \forall\, t \ge t_\epsilon,\ t_\epsilon > 0. \tag{12.104}$$

From (12.72) and Assumption A3) we get that $r(T) = O(T)$, $n \to \infty$. Then by (12.103), (12.104), and (iii) of Theorem 12.11 we have

$$\sum_{t=0}^{T} (\widetilde{\alpha}_i(t) f^*(t) + \alpha \xi_N(t))^2 \le \frac{\alpha^2(2\epsilon+1)}{1-\delta} \sum_{t=0}^{T} \xi_N^2(t) + o\left(\sum_{t=0}^{T} \xi_N^2(t)\right) + o(T),\ i = 1, 2, ..., N, \tag{12.105}$$

which, together with (12.99), leads to

$$\sum_{t=0}^{T} \xi_N^2(t+1) \le \frac{\alpha^2(2\epsilon+1)}{1-\delta} \sum_{t=0}^{T} \xi_N^2(t) + o(\sum_{t=0}^{T} \xi_N^2(t)) + o(n) + \sum_{t=0}^{T} \left(\frac{1}{N} \sum_{j=1}^{N} \omega_j(t+1)\right)^2. \tag{12.106}$$

From this and from Assumption A1) we get

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \xi_N^2(t) \le \frac{\sigma^2}{N(1-\mu(\alpha,\epsilon,\delta))} \text{ a.s.}, \tag{12.107}$$

where $\mu(\alpha,\epsilon,\delta) \triangleq \frac{\alpha^2(2\epsilon+1)}{1-\delta}$. This, together with (12.105), leads to

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} (\widetilde{\alpha}_i(t) f^*(t) + \alpha \xi_N(t))^2 \le \frac{\sigma^2 \mu(\alpha,\delta)}{N(1-\mu(\alpha,\epsilon,\delta))} \text{ a.s.} \tag{12.108}$$

Furthermore, by (12.98), Assumption A3), and Lemma 12.A.4, we have (ii).

From (12.98), (12.66), and Assumption A1) it follows that

$$\begin{aligned} J_i^N(u_i^N, u_{-i}^N) &= \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} [\widetilde{\alpha}_i(t) f^*(t) + \alpha \xi_N(t) + \Phi(t, f^*(t)) - \Phi(t, \overline{x}_N(t)) + \omega_i(t+1)]^2 \\ &= I_1^N + I_2^N + \sigma^2, \end{aligned} \tag{12.109}$$

where

$$I_1^N = 2 \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} [\widetilde{\alpha}_i(t) f^*(t) + \alpha\xi_N(t) + \Phi(t, f^*(t)) - \Phi(t, \overline{x}_N(t))]\omega_i(t+1)$$

and

$$\begin{aligned} I_2^N &= \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} [\widetilde{\alpha}_i(t) f^*(t) + \alpha\xi_N(t) + \Phi(t, f^*(t)) - \Phi(t, \overline{x}_N(t))]^2 \\ &\le 2 \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} [\widetilde{\alpha}_i(t) f^*(t) + \alpha\xi_N(t)]^2 + 2\gamma^2 \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \xi_N^2(t), \end{aligned}$$

which, together with (12.109), (12.107), (12.108), and Lemma 12.A.4, leads to

$$\max_{1\le i\le N} J_i^N(u_i^N, u_{-i}^N) \le \sigma^2 + \frac{2\sigma^2(\gamma^2 + \mu(\alpha, \epsilon, \delta))}{N(1-\mu(\alpha, \epsilon, \delta))} \text{ a.s.} \tag{12.110}$$

Letting ϵ and δ go to zero in (12.110) and (12.107), we get (iii) and

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \xi_N^2(t) \le \frac{\sigma^2}{N(1-\alpha^2)} \text{ a.s.,} \tag{12.111}$$

which gives (i). □

Remark 12.5. *Comparing (12.100) and (12.102) with (12.74) and (12.76), it is shown that for the case with unknown coupling strength, under the adaptive control law designed, the convergence rates of the estimation error for PSA and the cost function of each agent to the best response value are the same as those for the case with known coupling strength.*

Remark 12.6. *From Theorem 12.12 we can see that to ensure the control law to be an asymptotic Nash equilibrium, the consistency of the identification for the coupling strength α is not necessary. This is similar to the case of LS-based adaptive tracker (Guo and Chen [53]).*

In the following theorem, under certain excitation condition on the nonlinear iteration, we get the asymptotic consistency of the identification algorithm, that is, the upper limit of the identification error vanishes as the number N of agents increases to infinity. We need the following assumption.

A5) The solution of the nonlinear iteration $x(t+1) = \Phi(t, x(t))$ with $x(0) = x_0$ satisfies

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} x^2(t) > 0.$$

Theorem 12.13. *If Assumptions (A1)–(A5) hold, then for system (12.65), under the control (12.72), (12.80), (12.81), and (12.82), the closed-loop system satisfies*

$$\limsup_{t\to\infty} \widetilde{\alpha}_i^2(t) \leq \frac{\alpha^2\sigma^2}{N(1-\alpha^2)\underline{f}}, \; i=1,2,\ldots,N, \; a.s.,$$

where $\underline{f} = \liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}(f^*(t))^2 > 0$.

Proof. By (ii) of Theorem 12.11 and Assumption A4) we have

$$\widetilde{\alpha}_i^2(T+1) \leq \alpha^2 \frac{2\epsilon+1}{\epsilon+1}\frac{\sum_{t=0}^{T}\xi_N^2(t)}{r(T)} + o\left(\frac{\sum_{t=0}^{T}\xi_N^2(t)}{r(T)}\right) + O\left(\frac{\ln r(T)}{r(T)}\right) \text{ a.s.,}$$
$$i=1,2,\ldots,N, \; \forall \epsilon > 0. \quad (12.112)$$

From Assumption A5) we have that there exists $c_0 > 0$ such that $r(T) \geq c_0 T$ for sufficiently large T, which, together with (12.112), (i) of Theorem 12.12, and Assumption A5), leads to

$$\limsup_{t\to\infty} \widetilde{\alpha}_i^2(t) \leq \frac{\alpha^2\sigma^2}{N(1-\alpha^2)\underline{f}}\frac{2\epsilon+1}{\epsilon+1} \text{ a.s., } i=1,2,\ldots,N.$$

Letting ϵ go to zero, we get the conclusion of the theorem. □

12.4 Other Topics and Theoretical Challenges

Formation and flocking control is an important research direction of multiagent systems, which means that multiple agents form a predefined geometric shape through team collaboration subjected to environmental constraints (such as obstacle avoidance, etc.) simultaneously. The research on formation control is initially inspired by the collective behavior of biological groups. For the study of multiagent formation and flocking, on one hand, researchers hope to reveal the inherent mechanism how a biological group forms an ordered mode on the macroscopic level through collaboration among individuals on the microscopic level, and, on the other hand, the self-organized collective behavior of biological groups can also inspire designing novel and practical formation control algorithms. In practice, compared to single-agent systems, multiagent systems can complete complex tasks with more efficiency and flexibility. Multiagent formation control has a wide range of applications in the military, industrial, aerospace, and other fields. Multiagent formation and flocking are often studied under the distributed control framework. Compared with centralized control systems, distributed control systems have more challenges in the control law design due to the complex interactions between multiple agents, intrinsic parallelism, high system dimensions, incomplete information and uncertainty, and so on. Up to now, the research directions of multiagent formation and flocking control mainly include formation generation, formation maintenance, formation reconfiguration, and so on. Formation control strategies mainly include the leader-following-based approach, the behavior-based approach, the virtual-structure-based method, and the artificial potential field approach.

For the leader-following based approach, there are one (or more agents) in the formation to act as the leader (or leaders), and the rest agents play as the followers, which track the position and direction of the leaders. Wang [54] discussed the formation generation problem and some navigation strategies for individual movements in the formation, such as the nearest-neighbor tracking, multineighbor tracking, etc. Also, sufficient conditions for the formation stability based on nearest-neighbor tracking strategy were developed. Kumar et al. [55–57] proposed two kinds of leader-following-based formation patterns and feedback control laws for nonholonomic mobile robots and established the asymptotic stability of the closed-loop systems. Das et al. [58] developed the bottom-up approach to various kinds of composite controllers and estimators and realized multiagent formation maintenance and switching based on the omnidirectional visual information of agents. Pereira et al. [59] proposed the cooperative leader–follower approach, where the motions between leaders and followers interact with each other. Consolini et al. [60] studied the formation control with control input constrains that restrict the possible trajectories of the leader and admissible positions of the followers relative to the leader. Dimarogonas et al. [61] designed distributed coordinated control of rotating rigid bodies by leader-following-based approach and graph theory. Defoort et al. [62] proposed a second-order sliding-mode controller to realize multiagent only without the knowledge of the absolute velocity of the leader. The prominent advantage of leader-following-based approach lies in the transformation of the formation problem into the tracking problem, whose closed-loop stability can be solved by the control theory. The drawback of leader-following-based approach is that the chain information structure leads to poor robustness against disturbances. Once leaders are damaged, catastrophic damage happens for the entire group. In addition, for most cases, there is no feedback to the leader's movement from the followers. If the followers cannot keep up with the leader's movement, then the group cannot form an effective formation due to the lack of certain feedback mechanisms.

For the behavior-based approach, simple basic actions of individual agents are first designed, and then more complex group movements are achieved by assembling these simple actions in a certain way. In 1987, for generating realistic and efficient bird flocking images, Reynolds [63] designed three basic rules of bird flocking: collision avoidance (attempting to avoid collisions with nearby birds), velocity matching (attempting to match the average velocity direction of nearby birds), and flock centering (attempting to fly to the average position of nearby birds). Through the combination of these three basic rules, the simulation of group aggregation of flocking birds is finally realized. Balch et al. made robots to achieve obstacle avoidance with formation maintenance by designing a series of simple actions, and simulation and experiments have demonstrated the effectiveness of behavior-based approach [64]. The behavior-based approach can naturally integrate multiple goals in a multiagent system. However, it is generally difficult to carry out quantitative mathematical analysis, such as the convergence speed and stability of the formation.

For the virtual structure method, the multiagent system is regarded as a virtual rigid structure, and the agents are treated as the points on the virtual structure with fixed relative position. When the virtual structure moves, the agents track the points on the virtual structure. Lewis et al. [65] introduced the concept of virtual structure in the formation control and adopted the bidirectional control strategy, where the robots move continuously to be kept in the virtual structure. At the same time, the virtual structure is continuously adapted to the positions of robots. Beard et al. [66] solved the spacecraft formation control problem based on the virtual structure method. Ogren et al. [67] used the Lyapunov function to define the formation error, which is fed back in the virtual structure method. Ren et al. [68] introduced the distributed control into the virtual structure method to overcome the shortcomings of the centralized virtual structure method and designed the distributed control strategy of the aircraft formation. Yoshioka et al. [69] designed the formation control algorithm for nonholonomic robots based on the virtual structure feedback linearization. For the virtual structure method, it is easy to define the cooperative behavior of the group and to maintain the group formation in the movement. The control law can be designed according to the formation errors. However, we have to keep the virtual structure consistent at all times, and it is hard to achieve frequent formation switching. In addition, the virtual structure method is not suitable for large-scale systems, since with the increasing of the number of agents, the restrictions among individuals become quite complex.

The artificial potential field approach models the agent movement space by a force field. There are both attractive and repulsive forces in the field, which ensure the aggregation and collision avoidance of agents. Leonard and Fiorelli [70] proposed a distributed control framework based on the artificial potential field and the virtual leader for the collaboration of multi-intelligent vehicles. Also, for the multi-intelligent vehicle collaboration, Ogren et al. [71] combined the virtual structure and artificial potential field to propose a stable control strategy, where each smart car is regarded as a mobile sensor, and the network as a mobile and self-organized sensor array. For the deployment of mobile sensors in unknown environment, Howard et al. [72] proposed a distributed strategy based on the artificial potential field method.

Although the research on multiagent formation has achieved fruitful results, the gap between theoretical methods and engineering application is still large. It is quite difficult to apply the theoretical results of formation control due to the challenges in practical applications, such as the changing in the external environment, the dynamical constraints of agents, and the limited capacity of communication and perception of agents. Most references assume that agents have ideal perception and communication capacity, however, there may be latency and limited bandwidth in the communication network and limitation on the sensors' perceived ranges. Therefore, it is challenging to study the formation control under nonideal communication and measurement environment.

12.5 Bibliographic Notes

Olfati-Saber and Murray [16] considered the average-consensus control for first-order integrator networks with fixed and switching topologies. They proved that, if at each time instant, the network is a strongly connected and balanced digraph, then the weighted average-type protocol can ensure average-consensus. Kingston and Beard [73] extended the results of [16] to the discrete-time models and weakened the condition of instantaneous strong connectivity. They proved that if at each time instant the topology graph is balanced and the union of graphs over every bounded time interval is strongly connected, then average-consensus can be achieved. Xiao and Boyd [74] considered first-order discrete-time average-consensus with fixed and undirected topologies. They designed the weighted adjacency matrix to optimize the convergence rate by semidefinite programming. In addition to these works, some researchers also considered the high-order dynamics [75,76], the topologies of random graphs [77–80] or control design based on individual performance optimization [81–84].

Real networks are often interfered by various kinds of noises during the sending, transmission, and receiving of information, such as thermal noise, channel fading, quantization effect during encoding and decoding [85], and so on. Consensus of dynamic networks with stochastic communication noises is a common problem in distributed systems [86] and has attracted the attention of some researchers [28,87–91]. Ren, Beard, and Kingston [88] and Kingston, Ren, and Beard [89] introduced time-varying consensus gains and designed consensus protocols based on a Kalman filter structure. They proved that, when there is no communication noise, the designed protocols can ensure consensus to be achieved asymptotically. Xiao, Boyd, and Kim [90] considered the first-order discrete-time average-consensus control with fixed topologies and additive input noises. They designed the optimal weighted adjacency matrix to minimize the static mean square consensus error. However, since the consensus gain and the adjacency matrix are time invariant, as time goes on, the state average of the system diverges with probability one, even if the noises are bounded. Huang and Manton [28] considered the first-order discrete-time consensus control with fixed topologies and communication noises. They introduced decreasing consensus gains $a(k)$ (where k is the discrete time instant) in the protocol to attenuate the noises. They proved that if $a(k)$ is of order $1/k^{\gamma}$, $k \to \infty$, $\gamma \in (0.5, 1]$, and the network is a strongly connected circulant graph, then the static mean square error between the individual state and the average of the initial states of all agents is in the same order as the variance of the noises; if $a(k)$ satisfy the step rule of standard approximation and the network is a connected undirected graph, then the designed protocol can ensure mean square weak consensus. Li and Zhang [29,91] considered the first-order continuous-time average-consensus control with fixed topologies and communication noises. They used time-varying consensus gains in the protocol and gave a necessary and sufficient condition for asymptotically unbiased mean square average-consensus.

The LQG mean-field games with scalar agent models and deterministic discounted cost functions are studied by Huang, Malhamé, and Caines [41–92]. Li and Zhang [50–93] introduced the concepts of asymptotic Nash equilibria in the probability sense and extended to the cases with state space or ARX dynamic models and stochastic ergodic cost functions. The mean-field method is also developed independently by Lasry and Lions [94] and Weintraub and Benkard [95,96] by using the concept of oblivious equilibrium. The mean-field control for Markov decision problems is considered in Tembine, Boudec, El-Azouzi, and Altman [97]. Now, decentralized mean-field games have been extended to nonlinear dynamic models and the case with heterogeneous agents [98–100].

Appendix 12.A

12.A.1 Proof of Theorem 12.5

Lemma 12.A.1. *([101]) Let* $\{u(k), k = 0, 1, \cdots\}$, $\{\alpha(k), k = 0, 1, \cdots\}$, *and* $\{q(k), k = 0, 1, \cdots\}$ *be real sequences satisfying* $0 < q(k) \le 1$, $\alpha(k) \ge 0$, $k = 0, 1, \cdots$, $\sum_{k=0}^{\infty} q(k) = \infty$, $\frac{\alpha(k)}{q(k)} \to 0$, $t \to \infty$, *and*

$$u(k+1) \le (1 - q(k))u(k) + \alpha(k).$$

Then $\limsup_{k\to\infty} u(k) \le 0$. *In particular, if* $u(k) \ge 0$, $k = 0, 1, \cdots$, *then* $u(k) \to 0$, $k \to \infty$.

Lemma 12.A.2. *Apply the protocol (12.5) to system (12.3)–(12.4). If* $W(t) = 0$, $t = 0, 1, \ldots$, *then*

$$\lim_{t\to\infty} \|X(t) - JX(0)\|_2 = 0, \ \forall \ X(0) \in \mathcal{R}^N, \tag{12.A.1}$$

only if A1)–A2) hold.

Proof. It suffices to show that if $\mathcal{G}$ is a nonbalanced graph or $\mathcal{G}$ contains no spanning tree, then (12.A.1) does not hold.

Step 1: Consider the case where $\mathcal{G}$ is a nonbalanced graph. In this case, since $L_{\mathcal{G}}$ is the Laplacian matrix, $L_{\mathcal{G}}$ has a zero eigenvalue. Hence, there exists an N-dimensional vector α,

$\alpha^T\mathbf{1} = 1$, such that $\alpha^T L_{\mathcal{G}} = 0$. Furthermore, by Theorem 6 of [16], $\alpha \neq \frac{1}{N}\mathbf{1}$. This, together with (12.6) and $W(t) \equiv 0$, implies that $\alpha^T X(t+1) = \alpha^T X(t)$, $t = 0, 1, \ldots$. Thus,

$$\alpha^T X(t) \equiv \alpha^T X(0), \ \forall X(0) \in \mathcal{R}^N. \tag{12.A.2}$$

When (12.A.1) holds, so does (12.7), and hence,

$$\lim_{t\to\infty} \alpha^T X(t) = \alpha^T J X(0) = \frac{1}{N}\mathbf{1}^T X(0), \quad \forall X(0) \in \mathcal{R}^N,$$

which, together with (12.A.2), leads to $\alpha = \frac{1}{N}\mathbf{1}$. This contradicts $\alpha \neq \frac{1}{N}\mathbf{1}$. Thus, (12.A.1) does not hold.

Step 2: Consider the case where $\mathcal{G}$ contains no spanning tree. In this case, there are only three possibilities [102]:

(I) $\mathcal{G}$ has at least one isolated node i_0. Applying protocol (12.5) results in

$$\begin{cases} x_{i_o}(t+1) = x_{i_o}(t), \\ \widetilde{X}(t+1) = (I_{N-1} - a(t)\widetilde{\mathcal{L}})\widetilde{X}(t), \ t = 0, 1, \ldots, \end{cases}$$

where $\widetilde{X}(t) = [x_1(t), \cdots, x_{i_o-1}(t), x_{i_o+1}(t), \cdots, x_N(t)]^T$, $\widetilde{\mathcal{L}}$ is the Laplacian matrix of the graph removing the isolated node i_0. Take $x_{i_o}(0) = 0$, $x_j(0) = 1, \forall\ j \neq i_0$. Then, $x_{i_o}(0) = 0$ implies $x_{i_o}(t) \equiv 0$. By $\widetilde{\mathcal{L}}\mathbf{1} = 0$ we have $x_j(t) \equiv 1$, $j \neq i_0$. Thus, (12.A.1) does not hold.

(II) $\mathcal{G}$ has no isolated node but has at least two source nodes i_1, i_2. Take $x_{i_1}(0) = 0$, $x_{i_2}(0) = 1$. Then, applying protocol (12.5), similarly to (I), we have $x_{i_1}(t) \equiv 0 \neq 1 \equiv x_{i_2}(t)$. Thus, (12.A.1) does not hold.

(III) $\mathcal{G}$ has no isolated node, has at most one source node, and can be divided into two subgraphs $\mathcal{G}_1 = \{\mathcal{V}_1, \mathcal{E}_1, \mathcal{A}_1\}$ and $\mathcal{G}_2 = \{\mathcal{V}_2, \mathcal{E}_2, \mathcal{A}_2\}$ satisfying $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$, $\mathcal{V}_1 \cap \mathcal{V}_2 = \Phi$, $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, $\mathcal{E}_1 \cap \mathcal{E}_2 = \Phi$. Without loss of generality, suppose that $\mathcal{V}_1 = \{1, 2, \ldots, |\mathcal{V}_1|\}$, $\mathcal{V}_2 = \{|\mathcal{V}_1| + 1, \ldots, |\mathcal{V}_1| + |\mathcal{V}_2|\}$, $\mathcal{A}_{\mathcal{G}} = diag(\mathcal{A}_1, \mathcal{A}_2)$ is a diagonal block matrix. Then applying protocol (12.5) leads to

$$\begin{cases} X_1(t+1) = (I_{|\mathcal{V}_1|} - a(t)\widetilde{\mathcal{L}}_1)X_1(t), \\ X_2(t+1) = (I_{|\mathcal{V}_2|} - a(t)\widetilde{\mathcal{L}}_2)X_2(t), \end{cases}$$

where $X_1(t)$, $X_2(t)$ are the states of the nodes in $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively, $\widetilde{\mathcal{L}}_1$ and $\widetilde{\mathcal{L}}_2$ are the Laplacian matrices of $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Take $x_i(0) = 0$, $i \in \mathcal{V}_1$, $x_j(0) = 1$, $j \in \mathcal{V}_2$. Then, similarly to (I), we have $x_i(t) \equiv 0 \neq 1 \equiv x_j(t)$, $i \in \mathcal{V}_1$, $j \in \mathcal{V}_2$. Thus, (12.A.1) does not hold. □

Lemma 12.A.3. *Apply the protocol (12.5) to system (12.3)–(12.4). If for any $M > 0$, there is $t_0 \geq M$ such that $\mathbf{Pr}\{\|(I - J)D_{\mathcal{G}}W(t_0)\|_2 > 0\} > 0$, then for any given $K \geq 0$, there is $t_1 \geq K$ such that $\mathbf{E}[V(t_1)] > 0$.*

Proof. By contradiction suppose that there is $K_0 > 0$ such that $\mathbf{E}[V(t)] = 0$ for all $t \geq K_0$. Then $\delta(t) = 0$ a.s., $\forall\, t \geq K_0$, which, together with (12.9), implies that $\mathbf{Pr}\{\|(I - J)D_{\mathcal{G}}W(t)\|_2 = 0\} = 1, \forall\, t \geq K_0$. This contradicts the condition of the lemma. Thus, the lemma is true. □

Proof of Theorem 12.5. We need only to show that none of the following four cases is true:

(I) Assumption A1) does not hold.

(II) Assumption A2) does not hold.

(III) Under Assumption A1), $\sum_{t=0}^{\infty} a^2(t) = \infty$.

(IV) Under Assumption A1),

$$\sum_{t=0}^{\infty} a^2(t) < \infty, \quad \sum_{t=0}^{\infty} a(t) < \infty.$$

Since $W = \{W(t) = 0, t = 0, 1, ...\} \in \mathcal{S}$, by Lemma 12.A.2 it is clear that neither (I) nor (II) is true. So, it suffices to show that neither (III) nor (IV) is true.

Step 1: Let us prove that (III) is not true.

Suppose that there is at least one node that is not a source node, that is, there is $i_0 > 0$, $i_0 \neq j_0 > 0$, such that $a_{i_0 j_0} > 0$. Without loss of generality, suppose $i_0 = 1$ and $j_0 = 2$. Let $W = \{[0, \widetilde{w}_{21}(t), ..., 0, ..., 0]^T, t = 0, 1, ...\}$, where $\{\widetilde{w}_{21}(t), t = 0, 1, ...\}$ is a standard white noise sequence. Then we can see that $W \in \mathcal{S}$.

If $\sum_{t=0}^{\infty} a^2(t) = \infty$, then by (12.29) and the convergence of $x_i(t)$, $i = 1, 2, ..., N$, in mean square to a common random variable with finite second-order moment, we would have that $\frac{1}{N}\mathbf{1}^T D_{\mathcal{G}} \sum_{t=0}^{n-1} a(t)W(t)$ converges in mean square to a random variable with finite second-order moment x_w as $t \to \infty$. Furthermore, by Corollary 4.2.5 of [27] we get

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{1}{N}\mathbf{1}^T D_{\mathcal{G}} \sum_{t=0}^{n-1} a(t)W(t)\right)^2 = \mathbf{E}(x_w)^2 < \infty. \tag{12.A.3}$$

On the other hand,

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{1}{N}\mathbf{1}^T D_{\mathcal{G}} \sum_{t=0}^{n-1} a(t)W(t)\right)^2 = \frac{a_{12}^2}{N^2} \sum_{t=0}^{\infty} a^2(t) = \infty.$$

This contradicts (12.A.3). Thus, (III) is not true.

Step 2: Let us prove that (IV) is not true.

Similarly to Step 1, suppose that $a_{12} > 0$ and that W is the same as in Step 1. Since $\sum_{t=0}^{\infty} a^2(t) < \infty$, $a(t) \to 0$ as $t \to \infty$. Notice that $\mathbf{E}[\|(I-J)D_{\mathcal{G}}W(t)\|_2^2] = \frac{N-1}{N}a_{12}^2 > 0$, $\forall\, t \geq 0$. Then, by Lemma 12.A.3 there is $t_0 > 0$ such that

$$\mathbf{E}[V(t_0)] > 0, \tag{12.A.4}$$

$$0 \leq 2a(t)\lambda_{\max}(\widehat{L}_{\mathcal{G}}) < \frac{\ln 2}{2},\ \forall\, t \geq t_0. \tag{12.A.5}$$

By (12.10), similarly to (12.12), we have

$$\mathbf{E}[V(t+1)] \geq (1 - 2\lambda_{max}(\widehat{L}_{\mathcal{G}})a(t))\mathbf{E}[V(t)],\ \forall\, t \geq t_0.$$

This, together with (12.A.5) and the inequality $1 - x \geq e^{-2x}$, $x \in [0, \frac{\ln 2}{2})$, implies

$$\mathbf{E}[V(n)] \geq \exp\left\{-4\lambda_{\max}(\widehat{L}_{\mathcal{G}})\sum_{t=t_0}^{n-1} a(t)\right\}\mathbf{E}[V(t_0)],\ \forall\, n > t_0.$$

Thus, from (12.A.4) and $\sum_{t=0}^{\infty} a(t) < \infty$ we have

$$\liminf_{t\to\infty} \mathbf{E}[V(t)] \geq \exp\left\{-4\lambda_{\max}(\widehat{L}_{\mathcal{G}})\sum_{t=t_0}^{\infty} a(t)\right\}\mathbf{E}[V(t_0)] > 0.$$

This contradicts the fact that $x_i(t)$, $i = 1, 2, \ldots, N$, converges in mean square to a common random variable. Thus, (IV) is not true. □

Lemma 12.A.4. *([48]) Let $\{X(t), \mathcal{F}_t\}$ be a matrix martingale difference sequence, and let $\{M(t), \mathcal{F}_t\}$ be an adapted sequence of random matrices such that $\|M(t)\| < \infty$, $\forall\, t \geq 0$. If*

$$\sup_{t\geq 0} \mathbf{E}[\|X(t)\|^{\alpha} | \mathcal{F}_{t-1}] < \infty\ a.s.$$

for some $\alpha \in (0, 2]$, then, as $T \to \infty$,

$$\sum_{t=0}^{T} M(t)X(t+1) = O\left(s_T(\alpha)\ln^{1/\alpha+\eta}(s_T^{\alpha}(\alpha) + e)\right)\ a.s.,\ \forall\, \eta > 0,$$

where

$$s_T(\alpha) = \left(\sum_{t=0}^{T} \|M(t)\|^{\alpha}\right)^{1/\alpha}.$$

References

[1] N.M. Freris, H. Kowshik, P. Kumar, Fundamentals of large sensor networks: connectivity, capacity, clocks, and computation, Proceedings of the IEEE 98 (2010) 1828–1846.

[2] C. Lenzen, T. Locher, P. Sommer, R. Wattenhofer, Clock synchronization: open problems in theory and practice, in: International Conference on Current Trends in Theory and Practice of Computer Science, Springer, pp. 61–70.

[3] L. Lin, C. Yang, M. Ma, S. Ma, Diffusion-based clock synchronization for molecular communication under inverse Gaussian distribution, IEEE Sensors Journal 15 (2015) 4866–4874.

[4] L. Lamport, Time, clocks, and the ordering of events in a distributed system, Communications of the ACM 21 (1978) 558–565.

[5] I.-K. Rhee, J. Lee, J. Kim, E. Serpedin, Y.-C. Wu, Clock synchronization in wireless sensor networks: an overview, Sensors 9 (2009) 56–85.

[6] N.M. Freris, S.R. Graham, P. Kumar, Fundamental limits on synchronizing clocks over networks, IEEE Transactions on Automatic Control 56 (2011) 1352–1364.

[7] A.R. Swain, R. Hansdah, A model for the classification and survey of clock synchronization protocols in WSNs, Ad Hoc Networks 27 (2015) 219–241.

[8] R. Carli, A. Chiuso, L. Schenato, S. Zampieri, A PI consensus controller for networked clocks synchronization, IFAC Proceedings Volumes 41 (2008) 10289–10294.

[9] G. Xiong, S. Kishore, Analysis of distributed consensus time synchronization with Gaussian delay over wireless sensor networks, EURASIP Journal on Wireless Communications and Networking 2009 (2009) 1.

[10] G. Xiong, S. Kishore, Discrete-time second-order distributed consensus time synchronization algorithm for wireless sensor networks, EURASIP Journal on Wireless Communications and Networking 2009 (2008) 1.

[11] J. Chen, Q. Yu, Y. Zhang, H.-H. Chen, Y. Sun, Feedback-based clock synchronization in wireless sensor networks: a control theoretic approach, IEEE Transactions on Vehicular Technology 59 (2010) 2963–2973.

[12] L. Lin, S. Ma, M. Ma, A group neighborhood average clock synchronization protocol for wireless sensor networks, Sensors 14 (2014) 14744–14764.

[13] L. Schenato, F. Fiorentin, Average TimeSynch: a consensus-based protocol for clock synchronization in wireless sensor networks, Automatica 47 (2011) 1878–1886.

[14] G.S. Seyboth, F. Allgöwer, Clock synchronization over directed graphs, in: 52nd IEEE Conference on Decision and Control, IEEE, pp. 6105–6111.

[15] G.S. Seyboth, D.V. Dimarogonas, K.H. Johansson, P. Frasca, F. Allgöwer, On robust synchronization of heterogeneous linear multi-agent systems with static couplings, Automatica 53 (2015) 392–399.

[16] R. Olfati-Saber, R. Murray, Consensus problems in networks of agents with switching topology and time-delays, IEEE Transactions on Automatic Control 49 (2004) 1520–1533.

[17] H.-T. Zhang, M.Z. Chen, G.-B. Stan, Fast consensus via predictive pinning control, IEEE Transactions on Circuits and Systems I: Regular Papers 58 (2011) 2247–2258.

[18] Y. Chen, J. Lü, Z. Lin, Consensus of discrete-time multi-agent systems with transmission nonlinearity, Automatica 49 (2013) 1768–1775.

[19] H.-T. Zhang, Z. Chen, Consensus acceleration in a class of predictive networks, IEEE Transactions on Neural Networks and Learning Systems 25 (2014) 1921–1927.

[20] Z. Cheng, H.-T. Zhang, M.-C. Fan, G. Chen, Distributed consensus of multi-agent systems with input constraints: a model predictive control approach, IEEE Transactions on Circuits and Systems I: Regular Papers 62 (2015) 825–834.

[21] A. Sinha, D. Ghose, Generalization of linear cyclic pursuit with application to rendezvous of multiple autonomous agents, IEEE Transactions on Automatic Control 51 (2006) 1819–1824.

[22] R. Olfati-Saber, Distributed Kalman filter with embedded consensus filters, in: Proceedings of the 44th IEEE Conference on Decision and Control, IEEE, pp. 8179–8184.

[23] L. Xiao, S. Boyd, S. Lall, A scheme for robust distributed sensor fusion based on average consensus, in: IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, IEEE, 2005, pp. 63–70.

[24] N. Lynch, Distributed Algorithms, Morgan Kaufmann, 1996.

[25] C. Godsil, G.F. Royle, Algebraic Graph Theory, vol. 207, Springer Science & Business Media, 2013.

[26] R.M. Dudley, Real Analysis and Probability, vol. 74, Cambridge University Press, 2002.

[27] Y.S. Chow, H. Teicher, Probability Theory: Independence, Interchangeability, Martingales, Springer Science & Business Media, 2012.

[28] M. Huang, J.H. Manton, Coordination and consensus of networked agents with noisy measurements: stochastic algorithms and asymptotic behavior, SIAM Journal on Control and Optimization 48 (2009) 134–161.

[29] T. Li, J.-F. Zhang, Mean square average-consensus under measurement noises and fixed topologies: necessary and sufficient conditions, Automatica 45 (2009) 1929–1936.

[30] G.C. Goodwin, K.S. Sin, Adaptive Filtering Prediction and Control, Courier Corporation, 2014.

[31] J. Neveu, Discrete-Parameter Martingales, vol. 10, Elsevier, 1975.

[32] R. Olfati-Saber, J. Fax, R. Murray, Consensus and cooperation in networked multi-agent systems, Proceedings of the IEEE 95 (2007).

[33] L. Moreau, Stability of multiagent systems with time-dependent communication links, IEEE Transactions on Automatic Control 50 (2005) 169–182.

[34] L. Moreau, Stability of continuous-time distributed consensus algorithms, in: Decision and Control, 2004. CDC. 43rd IEEE Conference on, vol. 4, IEEE, 2004, pp. 3998–4003.

[35] A. Jadbabaie, J. Lin, A. Morse, Coordination of groups of mobile autonomous agents using nearest neighbor rules, IEEE Transactions on Automatic Control 48 (2003) 988–1001.

[36] M. Cao, D.A. Spielman, A.S. Morse, A lower bound on convergence of a distributed network consensus algorithm, in: Proceedings of the 44th IEEE Conference on Decision and Control, IEEE, pp. 2356–2361.

[37] M. Huang, J.H. Manton, Stochastic consensus seeking with measurement noise: convergence and asymptotic normality, in: 2008 American Control Conference, IEEE, 2008, pp. 1337–1342.

[38] T. Basar, G.J. Olsder, Dynamic Noncooperative Game Theory, vol. 23, SIAM, 1999.

[39] E. Altman, T. Basar, Multiuser rate-based flow control, IEEE Transactions on Communications 46 (1998) 940–949.

[40] D. Bauso, L. Giarré, R. Pesenti, Non-linear protocols for optimal distributed consensus in networks of dynamic agents, Systems & Control Letters 55 (2006) 918–928.

[41] M. Huang, P.E. Caines, R.P. Malhamé, Large-population cost-coupled LQG problems: generalizations to non-uniform individuals, in: Decision and Control, 2004. CDC. 43rd IEEE Conference on, vol. 4, IEEE, 2004, pp. 3453–3458.

[42] M. Huang, R.P. Malhamé, P.E. Caines, Nash strategies and adaptation for decentralized games involving weakly-coupled agents, in: Proceedings of the 44th IEEE Conference on Decision and Control, IEEE, pp. 1050–1055.

[43] M. Huang, P.E. Caines, R.P. Malhamé, Individual and mass behaviour in large population stochastic wireless power control problems: centralized and Nash equilibrium solutions, in: Decision and Control, 2003. Proceedings. 42nd IEEE Conference on, vol. 1, IEEE, 2003, pp. 98–103.

[44] J. McNamara, A. Houston, E. Collins, Optimality models in behavioral biology, SIAM Review 43 (2001) 413–466.

[45] R. Breban, R. Vardavas, S. Blower, Mean-field analysis of an inductive reasoning game: application to influenza vaccination, Physical Review E 76 (2007) 031127.

[46] H. Yin, P.G. Mehta, S.P. Meyn, U.V. Shanbhag, Learning in mean-field oscillator games, in: Proceedings of the 49th IEEE Conference on Decision and Control (CDC), 2010, pp. 3125–3132.

[47] A.C. Kizilkale, P.E. Caines, Mean field stochastic adaptive control, IEEE Transactions on Automatic Control 58 (2013) 905–920.

[48] H.-F. Chen, L. Guo, Identification and Stochastic Adaptive Control, Springer Science & Business Media, 2012.

[49] T.E. Duncan, L. Guo, B. Pasik-Duncan, Adaptive continuous-time linear quadratic Gaussian control, IEEE Transactions on Automatic Control 44 (1999) 1653–1662.

[50] T. Li, J.-F. Zhang, Decentralized tracking-type games for multi-agent systems with coupled ARX models: asymptotic Nash equilibria, Automatica 44 (2008) 713–725.

[51] H.-F. Chen, L. Guo, A robust adaptive controller, IEEE Transactions on Automatic Control 33 (1991) 1035–1043.

[52] L. Guo, Time-Varying Stochastic Systems, Ji Lin Science and Technology Press, 1993.

[53] L. Guo, H.-F. Chen, The Åström–Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers, IEEE Transactions on Automatic Control 36 (1991) 802–812.

[54] P.K. Wang, Navigation strategies for multiple autonomous mobile robots moving in formation, Journal of Robotic Systems 8 (1991) 177–195.

[55] J.P. Desai, J. Ostrowski, V. Kumar, Controlling formations of multiple mobile robots, in: Robotics and Automation, 1998. Proceedings. 1998 IEEE International Conference on, vol. 4, IEEE, 1998, pp. 2864–2869.

[56] J.P. Desai, J.P. Ostrowski, V. Kumar, Modeling and control of formations of nonholonomic mobile robots, IEEE Transactions on Robotics and Automation 17 (2001) 905–908.

[57] J.P. Desai, A graph theoretic approach for modeling mobile robot team formations, Journal of Robotic Systems 19 (2002) 511–525.

[58] A.K. Das, R. Fierro, V. Kumar, J.P. Ostrowski, J. Spletzer, C.J. Taylor, A vision-based formation control framework, IEEE Transactions on Robotics and Automation 18 (2002) 813–825.

[59] G.A. Pereira, A.K. Das, V. Kumar, M.F.M. Campos, Formation control with configuration space constraints, in: Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on, vol. 3, IEEE, 2003, pp. 2755–2760.

[60] L. Consolini, F. Morbidi, D. Prattichizzo, M. Tosques, Leader–follower formation control of nonholonomic mobile robots with input constraints, Automatica 44 (2008) 1343–1349.

[61] D.V. Dimarogonas, P. Tsiotras, K.J. Kyriakopoulos, Leader–follower cooperative attitude control of multiple rigid bodies, Systems & Control Letters 58 (2009) 429–435.

[62] M. Defoort, T. Floquet, A. Kokosy, W. Perruquetti, Sliding-mode formation control for cooperative autonomous mobile robots, IEEE Transactions on Industrial Electronics 55 (2008) 3944–3953.

[63] C.W. Reynolds, Flocks, herds and schools: a distributed behavioral model, ACM SIGGRAPH Computer Graphics 21 (1987) 25–34.

[64] T. Balch, R.C. Arkin, Behavior-based formation control for multirobot teams, IEEE Transactions on Robotics and Automation 14 (1998) 926–939.

[65] M.A. Lewis, K.-H. Tan, High precision formation control of mobile robots using virtual structures, Autonomous Robots 4 (1997) 387–403.

[66] R.W. Beard, J. Lawton, F.Y. Hadaegh, et al., A coordination architecture for spacecraft formation control, IEEE Transactions on Control Systems Technology 9 (2001) 777–790.

[67] P. Ogren, M. Egerstedt, X. Hu, A control Lyapunov function approach to multi-agent coordination, in: Decision and Control, 2001. Proceedings of the 40th IEEE Conference on, vol. 2, IEEE, 2001, pp. 1150–1155.

[68] W. Ren, R. Beard, Decentralized scheme for spacecraft formation flying via the virtual structure approach, Journal of Guidance, Control, and Dynamics 27 (2004) 73–82.

[69] C. Yoshioka, T. Namerikawa, Formation control of nonholonomic multi-vehicle systems based on virtual structure, IFAC Proceedings Volumes 41 (2008) 5149–5154.

[70] N.E. Leonard, E. Fiorelli, Virtual leaders, artificial potentials and coordinated control of groups, in: Decision and Control, 2001. Proceedings of the 40th IEEE Conference on, vol. 3, IEEE, 2001, pp. 2968–2973.

[71] P. Ogren, E. Fiorelli, N.E. Leonard, Cooperative control of mobile sensor networks: adaptive gradient climbing in a distributed environment, IEEE Transactions on Automatic Control 49 (2004) 1292–1302.

[72] A. Howard, M.J. Matarić, G.S. Sukhatme, Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem, in: Distributed Autonomous Robotic Systems 5, Springer, 2002, pp. 299–308.

[73] D.B. Kingston, R.W. Beard, Discrete-time average-consensus under switching network topologies, in: 2006 American Control Conference, IEEE, 2006, pp. 3551–3556.

[74] L. Xiao, S. Boyd, Fast linear iterations for distributed averaging, Systems & Control Letters 53 (2004) 65–78.

[75] G. Xie, L. Wang, Consensus control for a class of networks of dynamic agents, International Journal of Robust and Nonlinear Control 17 (2007) 941–959.

[76] W. Ren, E. Atkins, Distributed multi-vehicle coordinated control via local information exchange, International Journal of Robust and Nonlinear Control 17 (2007) 1002–1033.

[77] Y. Hatano, M. Mesbahi, Agreement over random networks, IEEE Transactions on Automatic Control 50 (2005) 1867–1872.

[78] C.W. Wu, Synchronization and convergence of linear dynamics in random directed networks, IEEE Transactions on Automatic Control 51 (2006) 1207–1210.

[79] F. Fagnani, S. Zampieri, Average consensus with packet drop communication, SIAM Journal on Control and Optimization 48 (2009) 102–133.

[80] A. Tahbaz-Salehi, A. Jadbabaie, A necessary and sufficient condition for consensus over random networks, IEEE Transactions on Automatic Control 53 (2008) 791–795.

[81] D. Bauso, L. Giarré, R. Pesenti, Non-linear protocols for optimal distributed consensus in networks of dynamic agents, Systems & Control Letters 55 (2006) 918–928.

[82] T. Li, J.-F. Zhang, Decentralized tracking-type games for multi-agent systems with coupled ARX models: asymptotic Nash equilibria, Automatica 44 (2008) 713–725.

[83] M. Huang, P.E. Caines, R.P. Malhamé, Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria, IEEE Transactions on Automatic Control 52 (2007) 1560–1571.

[84] T. Li, J.-F. Zhang, Asymptotically optimal decentralized control for large population stochastic multiagent systems, IEEE Transactions on Automatic Control 53 (2008) 1643–1660.

[85] A. Kashyap, T. Başar, R. Srikant, Quantized consensus, Automatica 43 (2007) 1192–1203.

[86] W. Ren, R.W. Beard, E.M. Atkins, A survey of consensus problems in multi-agent coordination, in: Proceedings of the 2005, American Control Conference, IEEE, 2005, pp. 1859–1864.

[87] R. Carli, F. Fagnani, A. Speranzon, S. Zampieri, Communication constraints in coordinated consensus problems, in: 2006 American Control Conference, IEEE, 2006, pp. 4189–4194.

[88] W. Ren, R.W. Beard, D.B. Kingston, Multi-agent Kalman consensus with relative uncertainty, in: Proceedings of the 2005, American Control Conference, IEEE, 2005, pp. 1865–1870.

[89] D.B. Kingston, W. Ren, R.W. Beard, Consensus algorithms are input-to-state stable, in: Proceedings of the 2005, American Control Conference, IEEE, 2005, pp. 1686–1690.

[90] L. Xiao, S. Boyd, S.-J. Kim, Distributed average consensus with least-mean-square deviation, Journal of Parallel and Distributed Computing 67 (2007) 33–46.

[91] T. Li, Asymptotically unbiased average consensus under measurement noises and fixed topologies, IFAC Proceedings Volumes 41 (2008) 2867–2873.

[92] M. Huang, R.P. Malhamé, P.E. Caines, Nash equilibria for large-population linear stochastic systems of weakly coupled agents, in: Analysis, Control and Optimization of Complex Dynamic Systems, Springer, 2005, pp. 215–252.

[93] T. Li, J.-F. Zhang, Asymptotically optimal decentralized control for large population stochastic multiagent systems, IEEE Transactions on Automatic Control 53 (2008) 1643–1660.

[94] J.-M. Lasry, P.-L. Lions, Mean field games, Japanese Journal of Mathematics 2 (2007) 229–260.

[95] G.Y. Weintraub, C.L. Benkard, B. Van Roy, Oblivious equilibrium: a mean field approximation for large-scale dynamic games, in: NIPS, pp. 1489–1496.

[96] G.Y. Weintraub, C.L. Benkard, B. Van Roy, Markov perfect industry dynamics with many firms, Econometrica 76 (2008) 1375–1411.

[97] H. Tembine, J.-Y. Le Boudec, R. El-Azouzi, E. Altman, Mean field asymptotics of Markov decision evolutionary games and teams, in: Game Theory for Networks, 2009. GameNets' 09. International Conference on, 2009, pp. 140–150.

[98] M. Huang, R.P. Malhamé, P.E. Caines, et al., Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle, Communications in Information & Systems 6 (2006) 221–252.

[99] M. Huang, P.E. Caines, R.P. Malhamé, An invariance principle in large population stochastic dynamic games, Journal of Systems Science and Complexity 20 (2007) 162–172.

[100] M. Huang, Large-population LQG games involving a major player: the Nash certainty equivalence principle, SIAM Journal on Control and Optimization 48 (2010) 3318–3353.

[101] B.T. Polyak, Introduction to Optimization, Optimization Software Inc., New York, 1987.

[102] W. Ren, R. Beard, Consensus seeking in multiagent systems under dynamically changing interaction topologies, IEEE Transactions on Automatic Control 50 (2005) 655–661.

Index

Estimation and Control of Large-Scale Networked Systems

Tong Zhou, Keyou You and Tao Li

Estimation and Control of Large-Scale Networked Systems is the first book that systematically summarizes results on large-scale networked systems. In addition, this book also summarizes most recent results on structure identification of a networked system, attack identification and prevention.

This book also provides necessary mathematical knowledge for studying large- scale networked systems, and a systematic description of the current status of this field, which includes features of these systems, difficulties in dealing with its state estimation and controller design, and major achievements.

Numerical examples in the chapters provide strong application backgrounds and/or are abstracted from actual engineering problems, such as gene regulation networks and electricity power systems. This book is an ideal resource for researchers in the field of systems and control engineering.

Key features

- Provides necessary mathematical knowledge for studying large-scale networked systems
- Introduces new features for filter and control design of networked control systems
- Summarizes most recent results on structure identification of a networked system, attack identification and prevention

Engineering

ISBN 978-0-12-805311-9

9 780128 053119

ELSEVIER

Butterworth-Heinemann
An imprint of Elsevier
elsevier.com/books-and-journals