Communications and Control Engineering

Keyou You
Nan Xiao
Lihua Xie

Analysis and Design of Networked Control Systems

Springer

Communications and Control Engineering

More information about this series at http://www.springer.com/series/61

Keyou You · Nan Xiao · Lihua Xie

Analysis and Design of Networked Control Systems

Springer

Keyou You
Department of Automation
Tsinghua University
Beijing
China

Nan Xiao
Future Urban Mobility Interdisciplinary
Research Group
Singapore-MIT Alliance for Research
and Technology Centre
Singapore
Singapore

Lihua Xie
School of Electrical and Electronic
Engineering
Nanyang Technological University
Singapore
Singapore

ISSN 0178-5354 ISSN 2197-7119 (electronic)
Communications and Control Engineering
ISBN 978-1-4471-6614-6 ISBN 978-1-4471-6615-3 (eBook)
DOI 10.1007/978-1-4471-6615-3

Library of Congress Control Number: 2014955622

Springer London Heidelberg New York Dordrecht

Printed on acid-free paper

Springer-Verlag London Ltd. is part of Springer Science+Business Media (www.springer.com)

Preface

With the rapid advances in sensing, communication, and networking technologies, research on networked control has gained momentum in recent years due to the wide application potential of networked control systems (NCSs) in intelligent transportation, industrial automation, advanced manufacturing, health care, national defense, etc. The defining feature of NCSs is that information is exchanged between sensors, controllers, and actuators over a shared communication network.

Traditionally, control theory is mainly concerned with using "ideal" information in a feedback loop to achieve some control performance objective, whereas communication theory focuses on the reliable transmission of information over "imperfect" channels, and is relatively indifferent to the specific purpose of the transmitted information. Recent studies show that the insertion of a shared network in the feedback loop will significantly affect the performance of NCSs. In fact, the analysis and design of such control systems has motivated the development of a new control paradigm by incorporating ideas from both control and communication theories.

The primary objective of this book is to present a new perspective of unifying communication, estimation, and control to investigate NCSs. Specifically, the book is devoted to characterizing the effect of communication networks on the stability and performance of NCSs. Toward this, we restrict ourselves to several fundamental problems of control and estimation over communication networks. By integrating control/estimation and communication theory, we are able to present a number of important and interesting results concerning the minimum data rate for stabilization of linear systems over noisy digital channels, the minimum network requirement for stabilization of linear systems over fading channels, conditions for stability of Kalman filtering with intermittent observations, etc. The results reveal a fundamental link between the topological entropy of linear dynamical systems and quality of communication channels. In addition, the design of quantizer for stabilization of linear systems under various network environments is extensively discussed. Many problems of Kalman filtering under information constraints are considered and solved. We anticipate that the techniques and results presented in this book will be useful in the analysis and design of NCSs.

Some of the materials contained herein arose from the joint work with our collaborators and the book would not have been possible without their efforts and support. In particular, we are indebted to Prof. Minyue Fu at The University of Newcastle, Australia, for fruitful discussions on many research problems in NCSs and the first author would like to thank his kind invitation for visiting The University of Newcastle during his Ph.D. candidature. We are also indebted to Prof. Li Qiu from The Hong kong University of Science and Technology for thoughtful discussions and insightful comments. We wish to thank our colleagues, including Shuai Liu, Tao Li, Yuqian Guo, and Yanlong Zhao at the Nanyang Technological University, for helpful discussions. We acknowledge IEEE, Elsevier, and SIAM for granting us the permission to reuse materials copyrighted by these publishers in this book.

We are pleased to thank the support of the National Science Foundation of Singapore under the competitive research program NRF-CRP8-2011-3, the National Natural Science Foundation of China under grants 61304038 and 61304044, and the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and the Natural Science Foundation of Shanghai City under grant 13ZR1453500.

Beijing — Keyou You
Singapore — Nan Xiao
Singapore — Lihua Xie

Contents

Chapter 1
Overview of Networked Control Systems

The primary objective of this chapter is to give an overview of networked control systems (NCSs) and the organization of this book. In contrast to the traditional control systems, the feedback loop of an NCS is closed via a communication network. The constraints in communication and computation and the interaction between them render the classical approaches that deal separately with control and communication not suitable for the analysis and design of an NCS and require a new paradigm. The research on NCSs has thus attracted considerable attention in the past fifteen years.

The chapter is organized as follows. In Sect. 1.1, we begin with an introduction to NCSs, their distinct features and historical development. In Sect. 1.2, highlights on each chapter are provided.

1.1 Introduction and Motivation

NCSs are spatially distributed systems wherein the control loops are closed through a communication network as shown in Fig. 1.1. The use of a shared network to connect spatially distributed components, e.g., *actuators, sensors* and *controllers*, makes it possible to design large-scale systems, and offers advantages such as low cost of installation, flexibility in system implementation, and the ease of maintenance, over conventional control systems. Examples of practical significance of NCSs include sensor networks, industrial control networks, multiple vehicle coordination, and Micro-Electro-Mechanical systems (MEMS), where their aims are to control one or more dynamical systems by deploying a shared network for data exchange.

The incorporation of a communication network in the feedback loop makes the analysis and design of an NCS complicated since in most problems, estimation/control interacts with communication in various ways. For example, there exists a critical positive data rate below which there does not exist *any* quantization and control scheme able to stabilize an unstable plant [1, 2]. In addition, there exists a critical packet loss rate above which the mean state estimation error covariance matrices of the Kalman filter will diverge [3]. These phenomena strongly imply that the network capacity has significant effects on the control/filtering performance.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_1

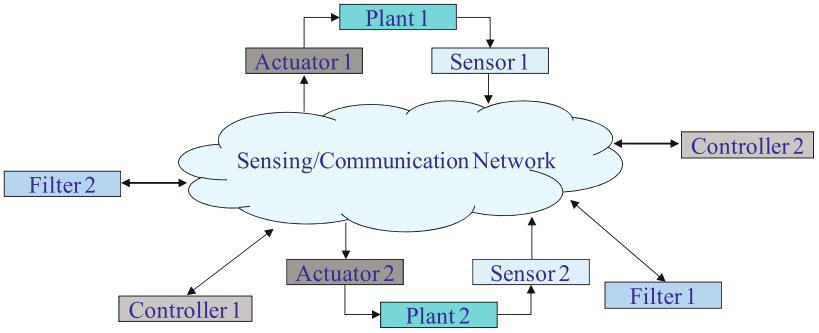

Fig. 1.1 General NCS architecture

While the conventional control theory is usually established under the assumption that data is exchanged without distortion and the information theory is relatively indifferent to the specific purpose of the transmitted information, the study in NCSs should pay attention to the interplay among communication, computation and control. This raises new fundamental challenges in investigating NCSs from the perspective of unifying communication, estimation, and control, and *control over networks* has been identified as one of the key directions for control research [4].

1.1.1 Components of NCS

The advantages of NCSs over the traditional control systems with wired point-to-point connections mainly result from the separation of the system components, which are individually described below.

(a) *Limited capacity network*. Current candidate networks for NCSs include DeviceNet [5], Ethernet [6], and FireWire [7], to name a few. Each network has its own protocols that are designed for a specific range of applications and the operation of an NCS largely depends on the performance parameters of the underlying network, which include transmission rate, delays, packet losses, and so on. These physical limitations of the network require us to reexamine the classical control theory.

We note that any communication channel is only able to carry information with a finite number of bits per unit time, which poses significant constraints on the operation of NCSs due to the possible low resolution of the transmitted information. For instance, in the current and future generations of MEMS arrays, there can be as many as 10^4–10^6 actuators or sensors on a single chip, and it is of central interest in closing feedback loops through wireless links such as Bluetooth^{TM} or IEEE 802.11(b) and using feedback network protocols such as CAN [8]. Although the total amount of information in bits per unit time may be large, each component is effectively allocated only a small portion. The low resolution feedback can only supply very limited information, and might induce large information loss.

To transmit a signal over a digital network, the signal must be sampled, encoded in a digital format, transmitted over the network and finally the data must be decoded at the receiver side, which will induce delays in the feedback loop. The variable network conditions such as congestion and channel quality may also induce variable delays between sampling and decoding at the receiver due to the network access delays (the time the network takes to accept data) and the transmission delays (the time during which data is in transmit inside the network) [9]. This happens in resource limited wireless sensor networks (WSNs) where communications between devices are power constrained and therefore limited in range and reliability. In feedback control systems delays are of primary concern.

On the other hand, changes in the environment such as the random presence of large metal objects will inevitably affect the propagation properties of the communication channels, or even block the communication links. Thus, data may be lost while in transit through the network, which is far more common in wireless than in wired networks. Long transmission delay sometimes may amount to a packet loss if the receiver discards outdated arrival data. This essentially means that the reliable transmission protocols such as TCP are not always be appropriate for NCSs since data that are retransmitted are outdated and may not be useful.

Other issues associated with NCSs include security and transmission errors. Those factors will inevitably lead to safety, performance degradation or even loss of stability of NCSs. Thus, it is fundamentally important to investigate the effect of network that closes the feedback on NCSs.

(b) *Sensors*. Sensors measure multiple physical quantities of a system and they include electronic sensors, biosensors, and chemical sensors. Sensor nodes are the simplest devices in the network. Since their number is usually much larger than the number of actuators or sinks, they have to be cheap, and hence operate on an extremely frugal energy budget. As such, they are usually operated with limited sensing, computing and communication capabilities, and in a distributed manner. These limitations should be taken into account in the analysis and design of NCSs.

(c) *Controllers*. In NCSs, sensors transmit the system output to the controller by putting the sensor measurement into a frame or a packet. The control signal is encapsulated in a frame or a packet and sent to the plant via the network as well. Under this setting, the traditional feedback control theory has to be revisited due to the availability of limited information. Thus, control over networks requires new design paradigms beyond classical control, and a deeper understanding on the nature of interactions between the cyber and physical worlds. The central theme of this book is to look for protocols, algorithms and design tools to bridge the gap between classical control and networked control.

1.1.2 Brief History of NCS

The research of NCSs is primarily fueled by the type of networks used in the feedback loop. The traditional communication architecture for control systems, which

has been successfully implemented in industry for decades, is point-to-point. In such systems the components are connected via hardwired connections and the systems are designed to bring all the information from the sensors to a central location where the conditions are monitored and decisions are taken on how to act [10]. However, this centralized point-to-point control system is not suitable to meet new application requirements in terms of modularity, decentralization of control, integrated diagnostics, quick and easy maintenance and low cost [11].

After extensive research and development, several network protocols for industrial control have been released. For example, Controller Area Network (CAN) was originally developed in 1983 by the German company Robert Bosch for use in car industries. Another example of industrial networks is Profibus developed by six German companies and five German institutes in 1987. Many other industrial network protocols including Foundation Fieldbus and DeviceNet were also developed about the same time period. These architectures can improve the efficiency, flexibility and reliability of NCSs through reduced wiring and distributed intelligence, and so reduce the installation, reconfiguration and maintenance time and costs. Today virtually all cars manufactured in Europe include embedded systems integrated through CAN.

With the rapid development of the wireless communication technology, it has become a trend to integrate devices through wireless rather than wired communication channels. With wireless networks deployed in control systems, the reliability and the time determinism of data transmissions are more difficult to guarantee. The assumption that the data collected are accurate, timely, and lossless is no longer valid for such systems. The shift from wired to wireless communication channels has highlighted important potential application advantages as well as several challenging problems for current research. In this book, most research problems on NCSs result from the use of wireless communications.

1.1.3 Challenges in NCS

The existing results in communication and information theory cannot provide immediate answers to many problems arising from control over networks in which the issues such as delay and causality are of fundamental importance. Shannon's capacity is a universal upper bound for the classical communication schemes, which usually does not consider the use of the transmitted message, and it is achievable under several restrictive assumptions.

- The capacity-achieving code is sufficiently long which results in significant delay. It is not very practical for real-time feedback control systems where the long delayed messages might not be useful.
- The information to be transmitted satisfies the asymptotic equipartition property [12], which usually requires that the information source satisfies some statistical properties, such as ergodicity and stationarity. Apparently, the information generated by a typical dynamic system can easily violate this property.

- The causality is not taken into consideration. Note that we are only concerned with causal systems where the output depends on past and current inputs but not future inputs. Thus, non-causal codes can not be used for designing feedback control for causal systems.
- There is no constraint on coding complexity.

Therefore, it is expected that one may be unable to achieve Shannon's capacity provided in standard information theory in practice, due to the constraint on coding complexity and the requirement on real-timeness and causality in a control system. The recent progress on causal coding [13, 14] may shed some light on this research direction. On the other hand, there is a lack of physical interpretation of the results on networked control in the context of information theory. Till now, a series of conditions have been proved to be both necessary and sufficient for stabilizability of networked systems under different channel models. However, the implementation issue of communication networks is hardly mentioned by the control community. Instead of Shannon's capacity, new notions of capacity are needed in the area of control over communication networks, which should serve the purposes of communications in control. The anytime capacity proposed by Sahai and Mitter [15] seems to be promising along this research line.

In a word, one of the urgent tasks in NCSs is to fill in the gap between the classical information theory and control theory.

1.2 Preview of the Book

The rest of this book is organized as follows.

In Chap. 2, we introduce various entropies in information and feedback control theories, respectively. Similar to the entropy of a random variable in information theory, *topological entropy* is an important quantity linking feedback control to network requirements for stabilization of a networked linear time-invariant (LTI) system. The most relevant results with focus on linking this quantity to various channel capacities are reviewed. Some possible research problems are highlighted as well.

In Chap. 3, a data rate theorem for stabilization of an LTI system over noiseless channel is established. As analogous to Shannon's channel-source coding theorem, this result shows that to stabilize a networked LTI system, the information rate of the noiseless channel has to be greater than the topological entropy of the system.

In Chap. 4, data rate theorem for mean square stabilization of an LTI system over lossy channels is developed under the assumption that the packet loss process follows an i.i.d. Bernoulli process. For general single input systems, the minimum data rate is explicitly given in terms of unstable eigenvalues of the open loop matrix and the packet loss rate.

In Chap. 5, we continue to enrich data rate theorem for mean square stabilization over lossy channels. Inspired by the Gilbert-Elliott channel model, the packet loss process is now modeled by a binary Markov process, which is more realistic than the i.i.d. case. It turns out that the minimum data rate for scalar systems can be

explicitly given in terms of the magnitude of the unstable mode and the transition probabilities of the Markov chain. Necessary and sufficient conditions on data rate for mean square stabilization of vector systems are provided respectively and shown to be optimal under some special cases.

In Chap. 6, the stabilization problem for NCSs over fading channels is studied. Both state feedback and output feedback are considered, where necessary and sufficient conditions on the network are derived for mean square stabilizability. We also present an extension to the case with output fading channels and the application of the results to vehicle platooning. Then the effect of pre- and post-channel processing and channel feedback is discussed in terms of the network requirement for stabilizability. In addition, the feedback stabilization and performance of a SISO plant controlled over a single channel are considered, where the channel undergoes both fading and additive noises and its input power is bounded by a predefined level.

In Chap. 7, logarithmic quantization is shown to be the coarsest quantizer for quadratic stabilization of an LTI system with single input. A fundamental result establishes that the study of quantized feedback control systems under logarithic quantization is equivalent to a robust control problem.

In Chap. 8, we first study the stabilization of uncertain linear systems via finite level logarithmic quantizers. We show that a dynamic finite-level logarithmic quantizer with sufficient number of quantization levels can stabilize the uncertain system. Next, the attainability of the minimum average data rate for stabilization of linear systems via logarithmic quantization is confirmed. We derive explicit finite-level logarithmic quantizers and the corresponding controllers to approach the minimum average data rate under two basic network configurations.

In Chap. 9, the stabilization of Markov jump linear systems (MJLSs) via mode dependent quantized feedback is addressed. A mode estimation algorithm is given when the mode process is not directly observed by the controller. The above results are applied to the quantized stabilization of NCSs over lossy channels, and an extension to the output feedback case is also included.

In Chap. 10, we develop a general multi-level quantized filter for linear stochastic systems, which has almost the same computational complexity as that of Kalman filter. It is demonstrated via simulations that under a moderate number of bits quantization, its performance comes comparable to Kalman filter.

In Chap. 11, we generalize the quantized innovations Kalman filter to a symmetric digital channel, and apply it to design the LQG controller for discrete-time stochastic systems.

In Chap. 12, we consider the stability of the Kalman filter over a network, where the unreliable network experiences both fading and additive noises described by white noise processes. The channel fading contains transmission failure and signal fluctuation simultaneously. We study the network requirement for the stability of the mean error covariance matrix of a remote Kalman filter and provide bounds for the mean error covariance matrix.

In Chap. 13, the behavior of the state estimation error covariance of Kalman filtering with Markovian packet loss is analyzed. For second-order and certain classes of higher-order systems, necessary and sufficient conditions for stability of the mean

estimation error covariance matrices are provided. All stability criteria are expressed by simple inequalities in terms of the largest open loop pole and transition probabilities of the Markov process.

In Chap. 14, we propose an estimation framework with scheduled measurements for linear discrete-time stochastic systems. Two types of scheduling algorithms are devised. Specifically, the first one is an iterative version, which sequentially decides the sending of each element of the vector measurement to a remote estimator by using the innovation. The second one is simpler and the sending of the measurements is driven by a random process. Under both schedulers, the (approximate) MMSE estimators are derived.

In Chap. 15, we study the tradeoff between the communication load and estimation performance under scheduled measurements in the framework of parameter estimation of linear systems. In particular, it turns out that the innovation based scheduler is an efficient approach to reduce the communication cost and maintain a good estimation performance.

References

1. G. Nair, R. Evans, Exponential stabilisability of finite-dimensional linear systems with limited data rates. Automatica **39**(4), 585–593 (2003)
2. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)
3. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
4. R. Murray, K. Astrom, S. Boyd, R. Brockett, G. Stein, Future directions in control in an information-rich world. IEEE Control Syst. Mag. **23**(2), 20–33 (2003)
5. W. Lawrenz, *CAN System Engineering: From Theory to Practical Applications*, vol. 1 (Springer, New York, 1997)
6. A. Tanenbaum, *Computer Networks*, 4th edn. (Prentice Hall, New Jersey, 2003)
7. D. Anderson, *FireWire System Architecture: IEEE 1394a* (Addison-Wesley Longman Publishing Co., Reading, 1999)
8. F. Lian, J. Moyne, D. Tilbury, Performance evaluation of control networks: ethernet, controlnet, and devicenet. IEEE Control Syst. **21**(1), 66–83 (2001)
9. J. Hespanha, P. Naghshtabrizi, Y. Xu, A survey of recent results in networked control systems. Proc. IEEE **95**(1), 138–162 (2007)
10. J. Baillieul, P. Antsaklis, Control and communication challenges in networked real-time systems. Proc. IEEE **95**(1), 9–28 (2007)
11. T. Yang, Networked control system: a brief survey. IET Control Theory Appl. **153**(4), 403–412 (2006)
12. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
13. M. Gastpar, Causal coding and feedback in Gaussian sensor networks. *Advances in Control, Communication Networks, and Transportation Systems* (Birkhäuser, Boston, 2005), pp. 91–110
14. T. Linder, R. Zamir, Causal coding of stationary sources and individual sequences with high resolution. IEEE Trans. Inf. Theory **52**(2), 662–680 (2006)
15. A. Sahai, S. Mitter, The necessity and sufficiency of anytime capacity for control over a noisy communication link. Part I: scalar systems. IEEE Trans. Inf. Theory **52**(8), 3369–3395 (2006)

Chapter 2
Entropies and Capacities in Networked Control Systems

In this chapter, we introduce some basic concepts and results in communication and information theories.

The chapter is organized as follows. In Sect. 2.1, the Shannon's entropy of a random variable and its importance in information theory are briefly reviewed. For control over networks, we are more concerned with the topological entropy of an LTI system. In Sect. 2.2, we give the definition of Shannon's channel capacity and exemplify it by several types of channels. In Sects. 2.3 and 2.4, the state-of-art techniques in NCSs are reviewed. Some open problems on NCSs are highlighted in Sect. 2.5.

2.1 Entropies

2.1.1 Entropy in Information Theory

For any probability distribution, *entropy* is a quantity to capture the concept of information, which has many properties that agree with the intuitive notion of what a measure of information should be. It is a measure of uncertainty of a random variable.

Definition 2.1 The entropy of a discrete random variable X with distribution function $p(x)$ and sample space $\mathscr{X}$ is defined as

$$H(X) \triangleq -\sum_{x\in\mathscr{X}} p(x)\log p(x). \tag{2.1}$$

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. If the base of the log is e, entropy is measured in *nats*. 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, $H(X) = -\mathscr{E}[\log p(X)]$, where $\mathscr{E}[\cdot]$ is the mathematical expectation operator.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_2

For two discrete random variables X and Y with joint probability mass function $p(x, y)$, the *joint entropy* is defined by $H(X, Y) = -\mathscr{E}[\log p(X, Y)]$. Similarly, the *conditional entropy* is defined by $H(X|Y) = -\mathscr{E}[\log p(X|Y)]$, 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 $p(x, y)$, mutual information is defined as

$$I(X; Y) = \mathscr{E}\left[\log \frac{p(X, Y)}{p(X)p(Y)}\right].$$

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 2.1 If X is a continuous random variable, differential entropy is defined accordingly, see [1, Chap. 9] for details.

Example 2.1 Let $X = 1$ with probability p and $X = 0$ with probability $1 - p$. Then, $H(X) = -p \log p - (1 - p) \log(1 - p)$. One can easily verify that $H(X) = 0$ when $p = 0$ or 1. This makes sense because if $p = 0$ or 1, 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 2.2 The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$H(\mathscr{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n), \tag{2.2}$$

when the limit exists.

If $\{X_i\}$ is an independent and identically distributed (i.i.d.) process, then $H(\mathscr{X}) = H(X_1)$. If $\{X_i\}$ is a stationary stochastic process, it is easy to verify that the limit in (2.2) always exists [1].

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 [1]. While 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.

2.1.2 Topological Entropy in Feedback Theory

In information theory, 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. [2]. It is expected that the topological entropy will be important to the data rate problem for stabilization of dynamical systems.

Definition 2.3 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 [3] or the degree of instability [4] of the plant. The mathematician Kurt Mahler first introduced his measure to polynomials [3]. The Mahler measure of a monic polynomial $a(z) = \prod_{i=1}^{n}(z - a_i)$ is defined as

$$M(a) \triangleq \prod_{i=1}^{n} \max\{|a_i|, 1\}. \tag{2.3}$$

The Mahler measure of a square matrix $A \in \mathbb{R}^{n\times n}$ is given by that of its characteristic polynomial:

$$M(A) \triangleq M(\det(zI - A)) = \prod_i \max\{|\lambda_i|, 1\} = 2^{H_T(A)}. \tag{2.4}$$

Then, Mahler measure of an LTI plant with any detectable and stabilizable realization (A, B, C, D) can be defined as the Mahler measure of system matrix A. The degree of instability of a square matrix A is defined in the same way as in (2.4) [4].

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, the importance of the topological entropy or Mahler measure to NCSs will be revealed.

2.2 Channel Capacities

In 1948, communications and information theory was founded by Claude Shannon in his seminal work, "A mathematical theory of communication" [5]. The central problem in this theory is how to transmit information over channels. A general communication system is depicted in Fig. 2.1.

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 not be perfectly recovered. *Channel capacity* is the tightest

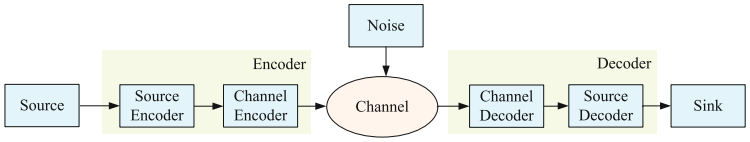

Fig. 2.1 General configuration of a communication system

upper bound on the "average" amount of information that can be transmitted over a communication channel. The following definitions are originally introduced in [5].

2.2.1 Noiseless Channels

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 will be received without error. The capacity $\mathscr{C}$ of a discrete noiseless channel is given by

$$\mathscr{C} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \log_2(\mu_k)}{n}, \tag{2.5}$$

when the limit exists.

Example 2.2 (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 $\mathscr{C} = 1$ bit.

2.2.2 Noisy Channels

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 $\mathscr{C}$ is defined by

$$\mathscr{C} = \max_{p(x)} I(X;Y), \tag{2.6}$$

where the maximum in (2.6) is taken over all possible input distributions $p(x)$.

The most fundamental results of communications and information theory are Shannon's source coding theorem and Shannon's channel coding theorem. The former states that the average number of bits (denoted by R) required to represent the result of a random event is given by its entropy (denoted by H), i.e., $R > H$, while the latter establishes that the probability of error could be made nearly zero for all communication rates (denoted by R) below channel capacity (denoted by $\mathscr{C}$),

i.e., $R < \mathscr{C}$. Similar to the well-known separation principle in control theory, the source-channel separation theorem further combines the source coding theorem and channel coding theorem. It is shown that if a source with entropy H satisfies the asymptotic equipartition property[1] and $H < \mathscr{C}$, then this source can be sent reliably over a discrete memoryless channel with capacity $\mathscr{C}$ [1].

The achievable channel capacity can be used to characterize diverse communication constraints depending on the underlying channel model and information pattern. Next, we highlight several practical issues of communication systems from information and communication point of view.

First of all, 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 [1]. In order 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.

Secondly, 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. The capacity constraint arises when a power bound on the channel input is introduced in order to avoid the interference to other communication users and/or due to the hardware limitations of the transmitter [1]. For an additive white Gaussian noise channel, the capacity can be computed simply from the noise characteristics of the channel as [1]

$$\mathscr{C} = \frac{1}{2}\log_2(1+\gamma), \tag{2.7}$$

where γ represents the signal-to-noise ratio (SNR) of the channel.

Thirdly, data loss or erasure is another common issue in general communication systems. It may result from transmission failure due to errors in wireless physical links, or from buffer overflows caused by congestion, or from packet reordering due to long transmission delays where the outdated signal is discarded in real-time applications [6]. The capacity of a binary erasure channel is [1]

$$\mathscr{C} = 1 - \alpha, \tag{2.8}$$

where α denotes the fraction of the bits that are erased,[2] and the receiver is assumed to know which bits have been lost.

Fourthly, the issue of fading phenomenon becomes increasingly important in wireless communication systems and has attracted recurring research interests from the communication community [8–10]. Fading describes the fluctuation experienced

[1] Intuitively, the asymptotic equipartition property indicates that "almost all events are almost equally surprising"; see, e.g., Chap. 3 of [1] for more details.

[2] We note that, for the erasure channel model considered in the control community, e.g., [7], the data is lost in the unit of packet (a collection of bits) rather than bit.

by transmitted signals due to the effects of multi-path and shadowing in wireless channels. A precise mathematical description of fading is either unavailable at present or too complex to deal with. Therefore, depending on the particular propagation environment and communication scenario, a range of simple statistical models have been proposed for fading channels (e.g., Rayleigh, Nakagami, Rician) [9]. Moreover, for non-coherent and ideal coherent modulations over fading channels, only the fading envelope statistics is required for analysis and design [10]. Results on the capacity of a fading channel can be found in [9, 11–13]. For example, if the channel state information (CSI) is known at the receiver, then the capacity of a fading channel is [9]

$$\mathscr{C} = \int_0^\infty \frac{1}{2} \log_2(1+\gamma) p(\gamma) d\gamma, \tag{2.9}$$

where γ denotes the instantaneous signal-to-noise ratio (SNR), and $p(\gamma)$ is the distribution function of γ.

Finally, MIMO communication system is needed in large-scale applications. The network information theory deals with the simultaneous rates of communication from multiple transmitters to multiple receivers in the presence of interference and noise [14], however a systematic theory is yet unavailable at the present time.

Shannon's capacity provided in standard information theory is a universal upper bound for all communication schemes, and it is achievable under the following restricted assumptions.

- The capacity-achieving code is sufficiently long.
- Causality is not taken into consideration.
- There is no constraint on coding complexity.

Therefore, it is expected that one may be unable to achieve Shannon's capacity in practice, due to the constraint on coding complexity and the requirement on real-timeness and causality. Although some progress has been made on causal coding [15, 16], the existing results in communications and information theory still cannot provide an immediate answer to control over networks in which the issues such as delay and causality are of fundamental importance. The so-called *data rate theorem* for stabilization to be presented later in this book is a reminiscent of the Shannon's source coding theorem.

2.3 Control Over Communication Networks

2.3.1 Quantized Control Over Noiseless Networks

Control using quantized feedback has been an important research area for a long time, even as early as in 1956 [17]. Most of the early work adopts the attitude that a quantized measurement of a real number is an *approximation* of that number and

models quantization error as an extra additive white noise [18]. The standard solutions of stochastic control are then applied. Although this approach would seem to be reasonable if the quantizer resolution is high, it is challenged in the new environment where only coarse information is allowed to propagate through the network due to limited network bandwidth or for the purpose of energy saving, e.g., in wireless sensor networks (WSNs). The change of view on quantization in the control community can be traced back to the seminal paper [19] where the author treats quantization as partial information of the quantized entity rather than its approximation, and demonstrates the significance of the historical values of the quantizer output. An important line of research that focuses on the interplay among coding, estimation and control was initiated by Wong and Brockett [20, 21]. Till now, various quantization methods for control 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 while a dynamic quantizer uses memory and is more complicated and potentially more powerful. In the same spirit of [19], Brockett and Liberzon [22] propose a dynamic finite-level uniform quantizer for stabilization, and point out that there exist a dynamic adjustment policy for the quantizer sensitivity and a quantized state feedback controller to asymptotically stabilize an LTI system. Those original works have motivated to raise a fundamental question: how much information needs to be communicated between the quantizer and the controller for stabilizing a discrete LTI system? Various authors have addressed this problem under different scenarios, e.g., [21, 23–27], leading to the appealing *data rate theorem* which is a reminiscent of the Shannon's source-channel coding theorem. To put precisely, the stabilization of an LTI system over a discrete noiseless channel can be achieved if and only if it satisfies the following inequality:

$$\mathscr{C} > H_T(A), \tag{2.10}$$

where $\mathscr{C}$ denotes the capacity of the noiseless channel and $H_T(A)$ is the topological entropy of the LTI system. This implies that if the plant is intuitively more unstable, a larger channel capacity is required for stabilization. To approach the minimum data rate, a dynamic quantizer is needed.

Coarser quantization implies lower resolution data circulating between the controller and the plant. It is of interest to seek the minimum quantization density for quadratically stabilizing an unstable plant and the corresponding optimal quantizer. In [28, 29], a logarithmic quantizer is proved to give the coarsest quantization density for quadratic stabilization of an unstable single input LTI system and the minimum density ρ is again determined by the product of unstable open poles, i.e.,

$$\rho = \frac{M(A) - 1}{M(A) + 1}, \tag{2.11}$$

where $M(A)$ is the Mahler measure of the LTI system in (2.4). Similarly, the more unstable the plant, the higher the quantization density (more information flows) is

required for quadratic stabilization. By using a sector bound approach, Fu and Xie [29] show that the quadratic stabilization problem with a set of logarithmic quantizers for MIMO systems is the same as quadratic stabilization of an associated system with sector-bounded uncertainty. This reveals that the minimum quantization density for MIMO systems is extremely difficult to establish.

However, a static logarithmic quantizer in [28, 29] uses an infinite data rate to represent the quantizer output, which is impractical. In [30], it is shown that an unstable linear system can be stabilized by using a fixed-rate Finite-level logarithmic quantizer with a dynamic scaling. Furthermore, we have shown that logarithmic quantization is optimal in the sense of approaching the minimum average data rate given in (2.10) to stabilize an LTI system in [31]. Apart from the theoretical merit on its own, the practical importance of studying logarithmic quantization is obvious since floating-point quantization may be treated as logarithmic quantization. Currently, scientific calculations are almost exclusively implemented by using floating-point roundoff and more and more digital signal processors contain floating-point arithmetic [32].

Performance control via quantized feedback has also been considered. Clearly, the quantizer and controller/estimator should be jointly designed so as to achieve the optimal performance for the overall system. This problem is generally very challenging, not only because the quantizer and estimator/contoller are inter-related but also due to that for a different performance criterion, the optimal quantizer-estimator/controller will be substantially different. In [29], a sector bound approach with the logarithmic quantizer is used to address the linear quadratic regulation (LQR) and H_∞ performance problems. However, its optimality is unclear. Optimal control of partially observed linear Gaussian systems is considered under a quadratic cost in [33–39]. Unlike the classical LQG problem, the separation principle for the design of control and estimation does not hold in general with quantized feedback, which makes it ambitious to design an optimal quantizer and controller to minimize the quadratic cost. The optimality of a quantized stabilization strategy is analyzed in [40, 41], where the number of quantization levels used by the feedback and the convergence time of the closed loop system play a central role.

2.3.2 *Quantized Control Over Noisy Networks*

Tatikonda and Mitter [42] study the observability and stabilizability of an LTI plant observed/controlled over different types of communication channels: a lossy channel with a given loss rate, an error-free limited-data-rate channel, an additive noise channel with a power constraint, and a delayed transmission channel. By applying information theoretic tools of source coding and channel coding, a general necessary condition for almost sure asymptotic observability and stabilizability is

$$\mathscr{C} > H_T(A), \tag{2.12}$$

where $\mathscr{C}$ is the Shannon's capacity of the noisy channel. The authors also study the sufficiency of (2.12) for observability and stabilizability in the case of lossy channels.

Unfortunately, the above sufficiency is shown to be incorrect [43] for a system with non-vanishing disturbance. Here we are not planning to explore all possible situations in NCSs where the notion of capacity is applicable but mainly focus on the additive noise channel model with SNR constraints [44–51] and the fading channel model [7, 52–54].

In a practical networked system, such as a resource limited WSN, the issues of packet loss and limited bandwidth usually co-exist. Therefore, it is of theoretical and practical significance to investigate the minimum data rate for stabilization over lossy networks. Recently, much effort has been devoted to examining how the limited data rate of the communication channel and the randomness of channel variation affect the stabilizability of an LTI system [42, 55–61]. Intuitively, due to possible packet loss, additional bits are required to stabilize the system. Thus, one of the fundamental issues is to quantify these additional bits required for the stabilization of the system. The problem is further complicated by the fact that different data rate may be required under different notions of stabilization. For instance, a necessary and sufficient condition on the almost sure stabilization over an erasure channel for a certain class of linear systems turns out to be that the Shannon capacity of the channel should be strictly greater than the intrinsic entropy rate of the system [42, 58, 59], which unfortunately fails for the moment stabilization [60]. Data rate theorem for mean square stabilization over lossy feedback channels is established in [55–57, 61].

The control community has also studied the additive noise model in the early work on mitigating quantization effects [39], and a brief review for recent results on networked control using additive noise model methodology can be found in [44]. In both continuous and discrete-time settings, Braslavsky et al. [46] show that there are limitations on the ability to stabilize an unstable SISO LTI plant over an SNR constrained channel using LTI feedback. For the discrete-time case, a necessary and sufficient condition for stabilizability via static state feedback is given by

$$\gamma > M(A)^2 - 1, \tag{2.13}$$

where γ is the SNR of the channel, and $M(A)$ is the Mahler measure of the system matrix A as defined in (2.4). The authors also provide a connection between their results and the data rate results in [24, 25]. Moreover, additional SNR would be required for stabilizing an non-minimum phase (NMP) plant when dynamic output feedback is concerned [46]. A series of results, e.g., [47–51] follow the same line of research in [45, 46]. The thesis [47] analyzes the effect of colored channel noise and bandwidth limitations, and it is shown that those constraints further increase the SNR requirement for stabilizability. The performance issue in terms of the sensitivity function is also studied in [47]. In [48], a minimum phase plant with relative degree one and a single unstable pole is assumed to be controlled over a first order moving average Gaussian channel. It is shown in [48] that the stabilization is possible precisely when the feedback capacity defined in [62] is greater than the topological entropy of the plant. The work [49] introduces a scaling factor to the network and addresses the output variance minimization problem for a minimum phase plant with relative degree one by using linear quadratic Gaussian (LQG) theory.

The networked architecture considered in [50] includes an LTI encoder-decoder pair and a channel feedback. It is proved in [50] that the availability of channel feedback plays a key role in reducing the minimum SNR for stabilizability. The tracking performance defined as the stationary variance of the tracking error is also investigated in [50]. The aforementioned results [45–50] all focus on the SISO case except for [50] where an extension to a simple two-by-two MIMO system is considered. The authors of [51] study the best achievable tracking performance of an MIMO LTI plant over parallel output SNR constrained channels, where the plant is assumed to be minimum phase.

From control point of view, the fading envelope can be considered as a random multiplicative gain that affects the transmitted signal. Elia [7] consider the mean square stabilization of an LTI plant over fading channels in the framework of robust control, where the randomness of the fading is interpreted as a stochastic model uncertainty. The additive noise of the channel is ignored in [7], since the main concern is the stabilization issue and the additive noise would not affect the results especially when there is no power constraint on the channel input. In this case, the controller design that allows the maximum variance of the channel uncertainty is posed as a nonconvex optimization problem and the D-K iteration procedure [63] can be used to solve the problem. In addition, it is proved in [7] that under stochastic channel fading, a necessary and sufficient condition for stabilizing a single-input unstable plant via state feedback in the mean square sense is

$$\mathscr{C}_{\mathrm{MS}} > H_T(A), \tag{2.14}$$

where $\mathscr{C}_{\mathrm{MS}}$ denotes the mean square capacity defined in terms of the fading envelope statistics, and $H_T(A)$ is the topological entropy of the system matrix A.

2.4 Estimation Over Communication Networks

A typical WSN usually consists of a large number of sensor nodes deployed in the area of interest and has found broad applications in environmental monitoring, intelligent buildings and transportation, logistics and national defense. In many cases, sensor nodes are equipped with limited power and communication resources, and have limited resolutions. These constraints have introduced challenges in the design of signal processing techniques and communication protocols to address problems associated with incomplete measurements. Recently, there is an increasing interest to develop various energy efficient algorithms for estimation problems with limited information, which is to be summarized below.

2.4.1 Quantized Estimation Over Noiseless Networks

The key problem in quantized estimation is the joint design of quantizer and the corresponding estimator to minimize the estimation error in an appropriate sense.

One of the main difficulties lies in that the unknown parameters are inaccessible to the quantizer design. For example, to estimate an unknown parameter θ under binary quantization of $y = \theta + v$, where v is a Gaussian random variable with zero mean, an optimal quantizer to minimize the mean square error is to simply place the quantizer threshold at θ [64, 65]. However, such a threshold selection is not implementable since θ is unavailable to the quantizer design. It is acknowledged that the estimation performance is very sensitive to the choice of the quantizer threshold [64, 65].

Motivated by this, an interesting quantizer threshold selection scheme is proposed in [65]. It consists in periodically applying a set of thresholds with equal frequencies, hoping that some thresholds are close to the unknown parameter. To asymptotically approach the minimum mean square error (MMSE), the authors in [66] construct an adaptive quantization involving delta modulation with a variable step size. The optimal step size is obtained through an online maximum likelihood estimation process, lacking a recursive form. This problem is resolved in [67], where a simple adaptive quantizer and a recursive estimation algorithm are designed to asymptotically approach the MMSE by exploiting the fact that quantizing innovations may require fewer bits than quantizing observations. The advantage of quantizing innovations is also extensively explored in [30, 68–70].

In fact, abundant quantization schemes have been developed in the context of WSN, e.g., [30, 39, 68, 69, 71–79]. Luo [78] studies the static parameter estimation under severe bandwidth constraints where each sensor's observation is quantized to one or a few bits. The resulting estimator turns out to exhibit a comparable variance that comes close to the variance of the optimal estimator which relies on un-quantized observations. To reduce the computational load of nonlinear filtering algorithms, [74] converts the integration problem into a finite summation using the quantization method. A quantized particle filter is established in [77] by the method of reconstructing the required probability density. Under a binary quantization, a dynamic quantization scheme based on feedback from the resource-sufficient estimation center is proposed for the state estimation of a hidden Markov model in [80]. The main disadvantage is that the solution involves a rather complicated on-line optimization and lacks a recursive form.

A very interesting single-bit quantized innovations filter called sign-of-innovations Kalman filter (SOI-KF) has been proposed in [68, 81], where a simple recursion involving time and measurement updates as in the standard Kalman filter is provided for state estimation. Inspired by [68] and also motivated by its limitation that the very rough quantization of the SOI inevitably induces large estimation errors, a finer quantization through an introduction of a dead zone is designed in [69]. In essence, a better estimator can be obtained by ignoring an innovation of small value than quantizing it into 1 or -1 and using the quantized message to update the state estimate. When more than one bit information can be sent at each transmission, we propose a multi-level quantized innovations Kalman filter (MLQ-KF) in [69]. The distinct features of MLQ-KF lie in its simplicity and a comparable performance to the Kalman filter with a moderate number of bits.

2.4.2 Data-Driven Communication for Estimation

Sensor nodes in a WSN are usually battery driven and hence operate on an extremely frugal energy budget. Experimental studies show that communication is a major source of energy consumption in sensor nodes, and communication consumes more energy than computation in a sensor node [82]. Motivated by this observation, an effective approach for energy saving in WSNs is to minimize the number of communications for sensor nodes under a prescribed performance requirement. Toward this purpose, an estimation framework under data driven communications to reduce the number of measurement transmissions has been proposed in [79, 83–85], where a scheduler is embedded in the sensor node to decide measurement transmission, i.e. only those scheduled measurements will be transmitted to the remote estimator. Thus, the sending of the measurement from the sensor to the remote estimator is controlled by a transmission scheduler. The main purpose of the scheduler lies in the hope of selecting "important" measurements to be communicated to the estimator while discarding less important ones. This implies that only a subset of measurements will be sent to the estimator, and thus certainly reduces the number of communications of the sensor node. Besides, there are many other energy efficient algorithms in the context of WSNs in the literature, such as data quantization, sparsity of the sensed phenomena, dimension reduction, decentralized information processing and dynamical model of Markovian events [64, 68, 69, 71, 78, 80, 86–98]. For instance, the effect of sparsity of a signal on the number of samples required to perfectly recover the signal is explored in [90]. To reduce the length of the transmitted information per transmission, data quantization is adopted in the state estimation problem of a Markov process [68, 99]. Sensor selection is also shown to be a good strategy to efficiently use the energy budget in WSNs [91, 92].

Since a scheduler is usually a nonlinear function of the sensor measurement, we have to deal with a nonlinear estimation problem. In the estimation theory [100], the measurement innovation represents new information of the current measurement that is not contained in the previous measurement data, and a small innovation indicates that the corresponding measurement prediction is close to the new measurement. In this case, it is not very necessary to transmit this measurement to the estimator for saving energy in the sensor node and communication bandwidth. This essentially suggests that the measurement innovation can be a good candidate for quantifying the "importance" of the measurement. Thus, it is reasonable to devise a scheduler based on the measurement innovation for state estimation problems.

The use of measurement innovation for reducing the amount of information to be transmitted from sensor to estimator has been pursued in [68, 69, 71, 79, 83, 84, 99] under various settings. In [69], the quantized state estimation is considered for the system of scalar measurements, and if the innovation lies outside a deadzone, then the quantized innovation is sent to the estimator. Otherwise, there is no communication between the sensor and the estimator at this time. Without accounting for quantization effect, a similar transmission scheduling based on the measurement innovation is adopted in [84]. A sequential scheduling algorithm by considering the different

importance of each element of a measurement vector is proposed in [79], where an approximate MMSE estimator by using the properties of conditional expectation is given. The concept of *scheduled transmission rate*[3] is introduced to evaluate the communication cost of the sensor node, which is defined as the ratio of the number of transmitted measurements to the total number of sensor measurements in the average sense. Intuitively, the larger the scheduled transmitted rate, the larger the number of measurements tends to be transmitted in a fixed time interval, resulting in a higher communication cost.

However, the above works lack a rigorous analysis on asymptotic properties of the estimator. In [83], an innovation based scheduler has been devised for the parameter estimation problem. They also investigate asymptotic properties of the proposed estimator under scheduled measurements when the number of sensor measurements tends to infinity, and establish that under some mild persistently exciting conditions, the proposed estimate is asymptotically optimal in the sense of minimizing the mean square estimation error.

2.4.3 Estimation Over Noisy Networks

Due to random fading and congestion, observation and control packets may be lost while in transit through a network. A motivating example is given by sensor and estimator/controller communicating over a wireless channel for which the quality of the channel randomly varies over time. The unreliability of the underlying communication network is modeled stochastically by assigning probabilities to the successful transmission of packets. This requires a novel theory to generalize classical control/estimation paradigms.

It is proved in [101] that with intermittent observations, Kalman filter is still optimal in the sense of achieving MMSE. By modeling the packet loss process as an i.i.d. Bernoulli process, Sinopoli et al. [101] prove the existence of a critical packet loss rate above which the mean state estimation error covariance matrices will diverge. 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, e.g., the lower bound is tight if the observation matrix is invertible. A less restrictive condition is provided in [102] where invertibility on the observable subspace is required. Mo and Sinopoli [103] explicitly characterize the loss rate for a wider class of systems, including second-order and the so-called non-degenerate higher-order systems. A remarkable discovery in [103] is that there are counterexamples of second-order systems for which the lower bound given by [101] is not tight.

[3] Note that *transmission rate* in information theory [1] is defined as the rate at which information is processed by a transmission facility. Its unit is usually expressed as bits per second.

To capture possible temporal correlations of network conditions, a time homogeneous binary Markov process is adopted to model the packet loss process in [104]. This is usually called the Gilbert-Elliott channel model. Under the i.i.d. packet loss model, stability of the estimation error covariance matrices in the mean sense may be effectively analyzed by a modified discrete-time Riccati recursion. In contrast, this approach is no longer feasible for the Markovian packet loss model, rendering the stability analysis more challenging. Due to the temporal correlation of the Markov process, the study of Markov packet loss model is far from trivial. In [104], an interesting notion of peak covariance stability in the mean sense is introduced. They give a sufficient condition for this stability notion for vector systems, which is also necessary for systems with observation index of one. A less conservative sufficient condition for the peak covariance stability under some cases is provided by [105]. However, those works do not exploit the system structure and fail to offer necessary and sufficient conditions for the peak covariance stability, except for the special systems with observation index of one. In addition, they are unable to characterize the relationship between the peak covariance stability and the usual stability of the estimation error covariance matrices for vector systems. Actually, the problem of deriving the usual stability condition for the mean estimation error covariance matrices of vector systems with Markovian packet loss is known to be extremely challenging. In our recent work [106], necessary and sufficient conditions for mean square stability of the estimation error covariance matrices for second-order and certain classes of higher-order systems with Markovian packet loss are provided. Other related works include [52, 54, 107–114].

There are some other probabilistic descriptions to examine the behavior of the estimation error covariance matrices, which are stochastic due to random packet loss. In [115], the performance of Kalman filtering is studied by considering a different metric $\mathbb{P}(P_k \leq M)$, i.e., the probability that the one-step prediction error covariance matrix P_k is bounded by a given positive definite matrix M, which is related to finding the cumulative distribution of P_k. This probability can be exactly computed for scalar systems and only has lower and upper bounds for vector systems [115]. Another performance metric called the stochastic boundedness is introduced in [116] for the i.i.d. packet loss model. It is worth pointing out that under different metrics, the effects of random packet loss on performance would be substantially different.

On the other hand, the authors of [52–54] study the Kalman filtering with faded observations. In [52, 53], the Kalman filtering over fading channels is studied under the assumption that the receiver can decide whether to accept or to reject each noisy packet. In the presence of information on instantaneous SNR at the receiver side, they show that keeping all the packets will minimize the mean error covariance, whereas in the absence of such information, packet drop should be designed to balance information loss and communication noise in order to optimize the performance. The work [54] considers the case where single or multiple sensors transmit their measurements to a remote Kalman filter across noisy fading channels. The exact value of channel fading or exact channel state information (CSI) is assumed to be known at the filter. Following the results in [117, 118] on Kalman filtering with random coefficients

and assuming the probability of transmission failure to be zero, they prove that the mean error covariance matrix of the remote Kalman filter is bounded from above and convergent. Upper bounds for the mean error covariance matrix are provided when the fading distribution is Rayleigh or Nakagami.

2.5 Open Problems

Given the previous literature review, we have identified the following open research problems in the area of control over networks.

- Data rate theorem for stabilization over lossy channels has been enriched, which is of equal importance in NCSs as Shannon's source coding theorem in information theory. The additional bit rate to counter the effect of random packet loss on stabilizability was exactly quantified only for single input systems under the i.i.d. packet loss model, and for scalar systems under the Markovian packet loss model. However, it is not well established for the general vector systems.
- It has been shown in [7] that the minimum network requirement for stabilizability through controller design is nonconvex and has no explicit solution in general. Only the minimum capacity for state feedback stabilization of a single-input plant over a single fading channel is given in [7]. The network requirement for both state feedback and output feedback stabilization of MIMO plants over multiple fading channels remains open. In addition, is there any benefit when additional components such as pre- and post-channel processing and channel feedback are introduced to NCSs with fading channels? What is the network requirement in relation to the channel input power for stabilizability?
- For quadratic stabilization of SISO linear systems using quantized state feedback, the coarsest quantization density is related to an "expensive" control problem in [28]. However, the results in [28] are hard to be extended to the MIMO case.
- As we can see from the results in the litertaure, e.g., [119–122] to name a few, a large class of NCSs with random packet dropouts can be modeled as Markovian jump linear systems (MJLSs), for which there have been many existing results on stability, optimal control and robust control [123]. Further motivated by the results on quantized stabilization of LTI systems in [28, 29], it would be interesting to study whether logarithmic quantization is still optimal for MJLSs and whether the sector bound approach is still non-conservative in dealing with quantized stabilization of MJLSs. In addition, if the answers to the aforementioned two questions are positive, then how to design the optimal quantizer and controller jointly?
- If the exact knowledge of packet loss or channel fading is available at the filter, then the time-varying Kalman filter is proved to be optimal in the MMSE sense [54, 101]. The authors of [54] limit their attention to the signal fluctuation issue and assumed that there is no transmission failure in the channel. In wireless networks, the coexistence of transmission failure and signal fluctuation is natural and reasonable, which raises the question as to what is the stability condition for

Kalman filtering over a fading network subject to both transmission failure and signal fluctuation. In this case, the results in [117, 118] cannot be directly applied or easily extended to solve the problem.

References

1. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
2. R. Adler, A. Konheim, M. McAndrew, Topological entropy. Trans. Am. Math. Soc. **114**(2), 309–319 (1965)
3. K. Mahler, An application of Jensen's formula to polynomials. Mathematica **7**, 98–100 (1960)
4. N. Elia, When Bode meets Shannon: control-oriented feedback communication schemes. IEEE Trans. Autom. Control **49**(9), 1477–1488 (2004)
5. C. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
6. J. Hespanha, P. Naghshtabrizi, Y. Xu, A survey of recent results in networked control systems. Proc. IEEE **95**(1), 138–162 (2007)
7. N. Elia, Remote stabilization over fading channels. Syst. Control Lett. **54**(3), 237–249 (2005)
8. D. Tse, P. Viswanath, *Fundamentals of Wireless Communication* (Cambridge University Press, Cambridge, 2005)
9. A. Goldsmith, *Wireless Communications* (Cambridge University Press, Cambridge, 2005)
10. M. Simon, M. Alouini, *Digital Communication Over Fading Channels* (Wiley-IEEE Press, London, 2005)
11. A. Goldsmith, P. Varaiya, Capacity of fading channels with channel side information. IEEE Trans. Inf. Theory **43**(6), 1986–1992 (1997)
12. L. Li, A. Goldsmith, Capacity and optimal resource allocation for fading broadcast channels—I: Ergodic capacity. IEEE Trans. Inf. Theory **47**(3), 1083–1102 (2001)
13. L. Li, A. Goldsmith, Capacity and optimal resource allocation for fading broadcast channels—II: Outage capacity. IEEE Trans. Inf. Theory **47**(3), 1103–1127 (2001)
14. A. El Gamal, T. Cover, Multiple user information theory. Proc. IEEE **68**(12), 1466–1483 (1980)
15. M. Gastpar, Causal coding and feedback in Gaussian sensor networks, *Advances in Control, Communication Networks, and Transportation Systems* (Birkhauser, Boston, 2005), pp. 91–110
16. T. Linder, R. Zamir, Causal coding of stationary sources and individual sequences with high resolution. IEEE Trans. Inf. Theory **52**(2), 662–680 (2006)
17. R. Kalman, Nonlinear aspects of sampled-data control systems, in *Proceedings of the Symposium on Nonlinear Circuit Analysis*, vol. 6, pp. 273–313 (1956)
18. D. Williamson, Finite wordlength design of digital Kalman filters for state estimation. IEEE Trans. Autom. Control **30**(10), 930–939 (1985)
19. D. Delchamps, Stabilizing a linear system with quantized state feedback. IEEE Trans. Autom. Control **35**(8), 916–924 (1990)
20. W. Wong, R. Brockett, Systems with finite communication bandwidth constraints-Part I: state estimation problems. IEEE Trans. Autom. Control **42**(9), 1294–1299 (1997)
21. W. Wong, R. Brockett, Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback. IEEE Trans. Autom. Control **44**(5), 1049–1053 (1999)
22. R. Brockett, D. Liberzon, Quantized feedback stabilization of linear systems. IEEE Trans. Autom. Control **45**(7), 1279–1289 (2000)
23. J. Baillieul, Feedback coding for information-based control: operating near the data-rate limit, in *Proceedings of 41st IEEE Conference on Decision and Control* (2002)
24. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)

25. G. Nair, R. Evans, Exponential stabilisability of finite-dimensional linear systems with limited data rates. Automatica **39**(4), 585–593 (2003)
26. S. Tatikonda, S. Mitter, Control under communication constraints. IEEE Trans. Autom. Control **49**(7), 1056–1068 (2004)
27. S. Yuksel, T. Basar, Minimum rate coding for LTI systems over noiseless channels. IEEE Trans. Autom. Control **51**(12), 1878–1887 (2006)
28. N. Elia, S. Mitter, Stabilization of linear systems with limited information. IEEE Trans. Autom. Control **46**(9), 1384–1400 (2001)
29. M. Fu, L. Xie, The sector bound approach to quantized feedback control. IEEE Trans. Autom. Control **50**(11), 1698–1711 (2005)
30. M. Fu, L. Xie, Finite-level quantized feedback control for linear systems. IEEE Trans. Autom. Control **54**(5), 1165–1170 (2009)
31. K. You, W. Su, M. Fu, L. Xie, Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization. Automatica **47**(1), 170–176 (2011)
32. B. Widrow, I. Kollar, M. Liu, Statistical theory of quantization. IEEE Trans. Instrum. Meas. **45**(2), 353–361 (1996)
33. G. Nair, F. Fagnani, S. Zampieri, R. Evans, Feedback control under data rate constraints: an overview. Proc. IEEE **95**(1), 108–137 (2007)
34. V. Borkar, S. Mitter, LQG control with communication constraints, *Communications, Computation, Control and Signal Processing: A Tribute to Thomas Kailath* (Kluwer, Norwell, 1997)
35. S. Tatikonda, A. Sahai, S. Mitter, Stochastic linear control over a communication channel. IEEE Trans. Autom. Control **49**(9), 1549–1561 (2004)
36. M. Fu, Linear quadratic Gaussian control with quantized feedback, in *American Control Conference*, pp. 2172–2177 (2009)
37. K. You, L. Xie, Linear quadratic Gaussian control with quantised innovations Kalman filter over a symmetric channel. IET Control Theory Appl. **5**(3), 437–446 (2011)
38. L. Bao, M. Skoglund, K. Johansson, Iterative encoder-controller design for feedback control over noisy channels. IEEE Trans. Autom. Control **56**(2), 265–278 (2011)
39. R. Curry, *Estimation and Control with Quantized Measurements* (MIT Press, Cambridge, 1970)
40. F. Fagnani, S. Zampieri, Stability analysis and synthesis for scalar linear systems with a quantized feedback. IEEE Trans. Autom. Control **48**(9), 1569–1584 (2003)
41. F. Fagnani, S. Zampieri, Quantized stabilization of linear systems: complexity versus performance. IEEE Trans. Autom. Control **49**(9), 1534–1548 (2004)
42. S. Tatikonda, S. Mitter, Control over noisy channels. IEEE Trans. Autom. Control **49**(7), 1196–1201 (2004)
43. A. Matveev, A. Savkin, Comments on control over noisy channels and relevant negative results. IEEE Trans. Autom. Control **50**(12), 2105–2110 (2005)
44. G. Goodwin, E. Silva, D. Quevedo, Analysis and design of networked control systems using the additive noise model methodology. Asian J. Control **12**(4), 443–459 (2010)
45. J. Braslavsky, R. Middleton, J. Freudenberg, Feedback stabilization over signal-to-noise ratio constrained channels, in *Proceedings of American Control Conference*, pp. 4903–4908 (2005)
46. J. Braslavsky, R. Middleton, J. Freudenberg, Feedback stabilization over signal-to-noise ratio constrained channels. IEEE Trans. Autom. Control **52**(8), 1391–1403 (2007)
47. A. Rojas, Feedback control over signal to noise ratio constrained communication channels. Ph.D. thesis (The University of Newcastle, Callaghan, Australia, 2006)
48. R. Middleton, A. Rojas, J. Freudenberg, J. Braslavsky, Feedback stabilization over a first order moving average Gaussian noise channel. IEEE Trans. Autom. Control **54**(1), 163–167 (2009)
49. J. Freudenberg, R. Middleton, J. Braslavsky, Stabilization with disturbance attenuation over a Gaussian channel, in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 3958–3963 (2008)
50. E. Silva, A unified framework for the analysis and design of networked control systems. Ph.D. thesis (The University of Newcastle, Callaghan, Australia, 2009)

51. Y. Li, E. Tuncel, J. Chen, W. Su, Optimal tracking performance of discrete-time systems over an additive white noise channel, in *Proceedings of the 48th IEEE Conference on Decision and Control*, pp. 2070–2075 (2009)
52. Y. Mostofi, R. Murray, To drop or not to drop: design principles for Kalman filtering over wireless fading channels. IEEE Trans. Autom. Control **54**(2), 376–381 (2009)
53. Y. Mostofi, R. Murray, Kalman filtering over wireless fading channels-how to handle packet drop. Int. J. Robust Nonlinear Control **19**(18), 1993–2015 (2009)
54. S. Dey, A. Leong, J. Evans, Kalman filtering with faded measurements. Automatica **45**(10), 2223–2233 (2009)
55. N. Martins, M. Dahleh, N. Elia, Feedback stabilization of uncertain systems in the presence of a direct link. IEEE Trans. Autom. Control **51**(3), 438–447 (2006)
56. P. Minero, M. Franceschetti, S. Dey, G. Nair, Data rate theorem for stabilization over time-varying feedback channels. IEEE Trans. Autom. Control **54**(2), 243–255 (2009)
57. K. You, L. Xie, Minimum data rate for mean square stabilization of discrete LTI systems over lossy channels. IEEE Trans. Autom. Control **55**(10), 2373–2378 (2010)
58. A. Matveev, A. Savkin, Comments on control over noisy channels and relevant negative results. IEEE Trans. Autom. Control **50**(12), 2105–2110 (2005)
59. A. Matveev, A. Savkin, An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels. SIAM J. Control Optim. **46**, 1323–1367 (2007)
60. A. Sahai, S. Mitter, The necessity and sufficiency of anytime capacity for control over a noisy communication link: Part I: scalar systems. IEEE Trans. Inf. Theory **52**(8), 3369–3395 (2006)
61. K. You, L. Xie, Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses. IEEE Trans. Autom. Control **56**(4), 772–785 (2011)
62. Y. Kim, Feedback capacity of the first-order moving average Gaussian channel. IEEE Trans. Inf. Theory **52**(7), 3063–3079 (2006)
63. K. Zhou, J. Doyle, *Essentials of Robust Control* (Prentice Hall, Upper Saddle River, 1998)
64. A. Ribeiro, G. Giannakis, Bandwidth-constrained distributed estimation for wireless sensor networks-part I: Gaussian case. IEEE Trans. Signal Process. **54**(3), 1131–1143 (2006)
65. H. Papadopoulos, G. Wornell, A. Oppenheim, Sequential signal encoding from noisy measurements using quantizers with dynamic bias control. IEEE Trans. Inf. Theory **47**(3), 978–1002 (2002)
66. J. Fang, H. Li, Distributed adaptive quantization for wireless sensor networks: from delta modulation to maximum likelihood. IEEE Trans. Signal Process. **56**(10), 5246–5257 (2008)
67. D. Marelli, K. You, M. Fu, Identification of ARMA models using intermittent and quantized output observations. Automatica **49**(2), 360–369 (2013)
68. A. Ribeiro, G. Giannakis, S. Roumeliotis, SOI-KF: distributed Kalman filtering with low-cost communications using the sign of innovations. IEEE Trans. Signal Process. **54**(12), 4782–4795 (2006)
69. K. You, L. Xie, S. Sun, W. Xiao, Multiple-level quantized innovation Kalman filter, in *Proceedings of 17th IFAC World Congress*, pp. 1420–1425 (2008)
70. V. Borkar, S. Mitter, S. Tatikonda, Optimal sequential vector quantization of Markov sources. SIAM J. Control Optim. **40**(1), 135–148 (2001)
71. J. Xiao, A. Ribeiro, Z. Luo, G. Giannakis, Distributed compression-estimation using wireless sensor networks. IEEE Signal Process. Mag. **23**(4), 27–41 (2006)
72. L. Wang, G. Yin, J. Zhang, Y. Zhao, *System Identification with Quantized Observations* (Birkhäuser, Boston, 2010)
73. Z. Luo, An isotropic universal decentralized estimation scheme for a bandwith constrained ad hoc sensor network. IEEE J. Sel. Areas Commun. **23**(4), 735–744 (2005)
74. G. Pagès, H. Pham, Optimal quantization methods for nonlinear filtering with discrete-time observations. Bernoulli **11**(5), 893–932 (2005)
75. A. Ribeiro, G. Giannakis, Bandwidth-constrained distributed estimation for wireless sensor networks-part II: unknown pdf. IEEE Trans. Signal Process. **54**(7), 2784–2796 (2006)

76. M. Huang, S. Dey, Dynamic quantizer design for hidden Markov state estimation via multiple sensors with fusion center feedback. IEEE Trans. Signal Process. **54**(8), 2887–2896 (2006)
77. R. Karlsson, F. Gustafsson, Particle filtering for quantized sensor information, in *Proceedings of the 13th European Signal Processing Conference, Antalya, Turkey*, September 2005
78. Z. Luo, Universal decentralized estimation in a bandwidth constrained sensor network. IEEE Trans. Inf. Theory **51**(6), 2210–2219 (2005)
79. K. You, L. Xie, Kalman filtering with scheduled measurements. IEEE Trans. Signal Process. **61**(6), 1520–1530 (2013)
80. M. Huang, S. Dey, Dynamic quantization for multisensor estimation over bandlimited fading channels. IEEE Trans. Signal Process. **55**(9), 4696–4702 (2007)
81. E. Msechu, S. Roumeliotis, A. Ribeiro, G. Giannakis, Decentralized quantized Kalman filtering with scalable communication cost. IEEE Trans. Signal Process. **56**(8), 3727–3741 (2008)
82. I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A survey on sensor networks. IEEE Commun. Mag. **40**(8), 102–114 (2002)
83. K. You, L. Xie, S. Song, Asymptotically optimal parameter estimation with scheduled measurements. IEEE Trans. Signal Process. **61**(14), 3521–3531 (2013)
84. J. Wu, Q. Jia, K. Johansson, L. Shi, Event-based sensor data scheduling: trade-off between sensor communication rate and estimation quality. Preprint (2011)
85. G. Battistelli, A. Benavoli, L. Chisci, Data-driven communication for state estimation with sensor networks. Automatica **48**(5), 926–935 (2012)
86. C. Wikle, N. Cressie, A dimension-reduced approach to space-time Kalman filtering. Biometrika **86**(4), 815–829 (1999)
87. I. Schizas, G. Giannakis, Z. Luo, Distributed estimation using reduced-dimensionality sensor observations. IEEE Trans. Signal Process. **55**(8), 4284–4299 (2007)
88. H. Zhu, I. Schizas, G. Giannakis, Power-efficient dimensionality reduction for distributed channel-aware Kalman tracking using WSNs. IEEE Trans. Signal Process. **57**(8), 3193–3207 (2009)
89. A. Makarenko, H. Durrant-Whyte, Decentralized Bayesian algorithms for active sensor networks. Inf. Fusion **7**(4), 418–433 (2006)
90. D. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
91. Z. Quan, W. Kaiser, A. Sayed, Innovations diffusion: a spatial sampling scheme for distributed estimation and detection. IEEE Trans. Signal Process. **57**(2), 738–751 (2009)
92. Y. Mo, R. Ambrosino, B. Sinopoli, Sensor selection strategies for state estimation in energy constrained wireless sensor networks. Automatica **47**(7), 1330–1338 (2011)
93. J. Li, G. AlRegib, Rate-constrained distributed estimation in wireless sensor networks. IEEE Trans. Signal Process. **55**(5), 1634–1643 (2007)
94. C. Berger, S. Choi, S. Zhou, P. Willett, Channel energy based estimation of target trajectories using distributed sensors with low communication rate. IEEE Trans. Signal Process. **58**(4), 2339–2350 (2010)
95. J. Gubner, Distributed estimation and quantization. IEEE Trans. Inf. Theory **39**(4), 1456–1459 (1993)
96. O. Ozdemir, R. Niu, P.K. Varshney, Channel aware target localization with quantized data in wireless sensor networks. IEEE Trans. Signal Process. **57**(3), 1190–1202 (2009)
97. G. Balkan, S. Gezici, CRLB based optimal noise enhanced parameter estimation using quantized observations. IEEE Signal Process. Lett. **17**(5), 477–480 (2010)
98. Y. Zhao, L. Wang, G. Yin, J. Zhang, Identification of Wiener systems with binary-valued output observations. Automatica **43**(10), 1752–1765 (2007)
99. K. You, L. Xie, S. Sun, W. Xiao, Quantized filtering of linear stochastic system. Trans. Inst. Meas. Control **33**(6), 683–698 (2011)
100. B. Anderson, B. Moore, *Optimal Filtering* Systems Sciences Series (Prentice-Hall, Englewood Cliffs, 1979)
101. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)

102. K. Plarre, F. Bullo, On Kalman filtering for detectable systems with intermittent observations. IEEE Trans. Autom. Control **54**(2), 386–390 (2009)
103. Y. Mo, B. Sinopoli, Towards finding the critical value for Kalman filtering with intermittent observations (2010). http://arxiv.org/abs/1005.2442
104. M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses. Automatica **43**(4), 598–607 (2007)
105. L. Xie, L. Xie, Stability of a random Riccati equation with Markovian binary switching. IEEE Trans. Autom. Control **53**(7), 1759–1764 (2008)
106. K. You, M. Fu, L. Xie, Mean square stability for Kalman filtering with Markovian packet losses. Automatica **47**(12), 2647–2657 (2011)
107. M. Trivellato, N. Benvenuto, State control in networked control systems under packet drops and limited transmission bandwidth. IEEE Trans. Commun. **58**(2), 611–622 (2010)
108. V. Gupta, A. Dana, J. Hespanha, R. Murray, B. Hassibi, Data transmission over networks for estimation and control. IEEE Trans. Autom. Control **54**(8), 1807–1819 (2009)
109. S. Kluge, K. Reif, M. Brokate, Stochastic stability of the extended Kalman filter with intermittent observations. IEEE Trans. Autom. Control **55**(2), 514–518 (2010)
110. S. Hu, W. Yan, Stability robustness of networked control systems with respect to packet loss. Automatica **43**(7), 1243–1248 (2007)
111. A. Censi, Kalman filtering with intermittent observations: convergence for semi-Markov chains and an intrinsic performance measure. IEEE Trans. Autom. Control **56**(2), 376–381 (2011)
112. S. Sun, L. Xie, W. Xiao, Y. Soh, Optimal linear estimation for systems with multiple packet dropouts. Automatica **44**(5), 1333–1342 (2008)
113. M. Epstein, L. Shi, A. Tiwari, R. Murray, Probabilistic performance of state estimation across a lossy network. Automatica **44**(12), 3046–3053 (2008)
114. N. Xiao, L. Xie, M. Fu, Kalman filtering over unreliable communication networks with bounded Markovian packet dropouts. Int. J. Robust Nonlinear Control **19**(16), 1770–1786 (2009)
115. L. Shi, M. Epstein, R. Murray, Kalman filtering over a packet-dropping network: a probabilistic perspective. IEEE Trans. Autom. Control **55**(3), 594–604 (2010)
116. S. Kar, B. Sinopoli, J. Moura, Kalman filtering with intermittent observations: weak convergence to a stationary distribution. IEEE Trans. Autom. Control **57**(2), 405–420 (2012)
117. P. Bougerol, Kalman filtering with random coefficients and contractions. SIAM J. Control Optim. **31**, 942–959 (1993)
118. P. Bougerol, Almost sure stabilizability and Riccati's equation of linear systems with random parameters. SIAM J. Control Optim. **33**(3), 702–717 (1995)
119. M. Yu, L. Wang, G. Xie, T. Chu, Stabilization of networked control systems with data packet dropout via switched system approach, in *Proceedings of IEEE International Symposium on Computer Aided Control Systems Design*, pp. 362–367 (2004)
120. P. Seiler, R. Sengupta, Analysis of communication losses in vehicle control problems, in *Proceedings of American Control Conference*, pp. 1491–1496 (2001)
121. P. Seiler, R. Sengupta, An H_∞ approach to networked control. IEEE Trans. Autom. Control **50**(3), 356–364 (2005)
122. J. Xiong, J. Lam, Stabilization of linear systems over networks with bounded packet loss. Automatica **43**(1), 80–87 (2007)
123. O. Costa, M. Fragoso, R. Marques, *Discrete-Time Markov Jump Linear Systems* (Springer, London, 2005)

Chapter 3
Data Rate Theorem for Stabilization Over Noiseless Channels

In classical control theory, a common assumption is that the signals sent from sensors to controllers and from controllers to actuators take continuous values with infinite precision, which is challenged in digital and networked control systems. In such systems, outputs or control variables must be reduced into finite bit for storage, manipulation and transmission, which inevitably introduces loss of data resolution, and eventually degrades the performance or even results in instability of closed-loop systems. This process is achieved by using quantizer. If the data resolution of the plant output is very low or the quantization is very coarse, the controller receives very rough information from the plant, and it may fail to generate suitable feedback signals to stabilize the plant. To the contrary, if the data resolution is relatively high, the effect of resolution loss may be neglected. Thus, it is natural to expect that there may exist a critical data resolution (data rate) below which the controller is unable to stabilize an unstable system in an appropriate sense. This chapter studies this problem, and is concerned with the stabilization of linear systems over a noiseless channel, where the data exchanged between the plant and the controller contain only limited information.

The chapter is organized as follows. In Sect. 3.1, the problem of interest is formally described. We demonstrate in Sect. 3.2 that there is a significant limitation in the classical quantization approach by modeling the quantization error as an additive white Gaussian noise. In Sect. 3.3, a *data rate theorem* is established, which quantifies the minimum data rate required for the channel connecting the controller and plant to stabilize linear systems. In Sect. 3.4, we conclude the chapter.

3.1 Problem Statement

Consider a linear time invariant (LTI) system

$$\begin{cases} x_{k+1} = Ax_k + Bu_k, \\ y_k \quad\ = Cx_k, \end{cases} \tag{3.1}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_3

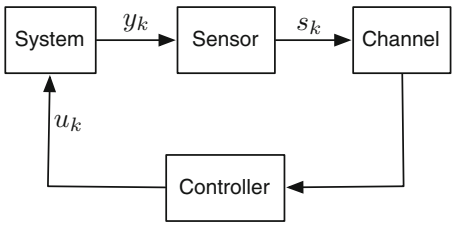

Fig. 3.1 Network configuration

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. u_k 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, and A is unstable.

Suppose that the output sensor that is equipped with an encoder communicates with the controller over a digital channel which can only support information exchange with a finite bit rate. At each time instant, the sensor sends one symbol s_k from a finite and possibly time-varying set $\mathscr{S}_k$ to the controller, see Fig. 3.1 for an illustration. In this chapter, the channel is assumed to be noiseless in the sense that s_k will be correctly received by the controller with negligible delay.

The communication *data rate* to quantify the information rate at which the channel supports is defined in the asymptotically average sense

$$R = \liminf_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} \log_2(|\mathscr{S}_j|), \tag{3.2}$$

where $|\mathscr{S}_j|$ denotes the cardinality of set $\mathscr{S}_j$.

In full generality, s_k is generated from a time-varying coder by making use of all the past and the present system outputs, and past coder outputs, i.e.,

$$s_k = \mathscr{E}_k(y_0^k, s_0^{k-1}) \in \mathscr{S}_k \tag{3.3}$$

where $\mathscr{E}_k$ is the coder mapping and $y_0^k := \{y_0, \ldots, y_k\}$.

On the channel receiver side, the controller has obtained $s_0, \ldots, s_k$ by the time k and applies a control law to generate a control feedback signal

$$u_k = \mathscr{D}_k(s_0^k), \tag{3.4}$$

where $\mathscr{D}_k$ is the controller mapping at time k.

Intuitively, if R is too small, s_k carries very limited output information, and the controller will not have accurate knowledge of the system. This may lead to that the controller fails to stabilize the system (3.1). The problem of fundamental interest is how to exactly quantify the critical data rate below which it is impossible to stabilize the system by *any* coder and control law. To this purpose, we study the networked linear system from an information-theoretic approach.

3.2 Classical Approach for Quantized Control

Quantization has been a long research topic in communications and information theory, see [1] and the references therein. Quantization refers to the process of approximating the continuous set of values with a finite (preferably small) set of values. The quantizer Q is a function whose set of output values are discrete, and usually finite, i.e.,

$$Q : \mathbb{R} \to \{q^1, \ldots, q^M\}.$$

The input to a quantizer is an analog value, and the output is one among the finite output 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, i.e.,

$$Q(x) = x + w,$$

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. While 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, i.e., (3.1) with $m = 1, A = a$ with $|a| > 1, B = C = 1$, and x_0 is unknown. 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, and w_k is uncorrelated with x_k.

The shortcoming of this approach becomes obvious. That is, the controller is impossible to asymptotically stabilize the system in the mean square sense as the noise can not be eventually eliminated.

In order to achieve the mean square stability of the closed-loop system, it requires the controller to estimate the initial state x_0 with a mean square error diminishing strictly faster than a^{-2k}. This motivates us to study the quantized control from a different perspective and with a more rigorous approach.

3.3 Data Rate Theorem for Stabilization

In this section, we present the result on the minimum data rate for the asymptotic stabilization of linear systems as follows.

Theorem 3.1 *Consider a networked control system* (3.1)*, 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*

$$R > \sum_{|\lambda_i| \geq 1} \log_2 |\lambda_i| := R_{\inf}, \tag{3.5}$$

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 [2]. 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* in Chap. 2, plays a role similar to source entropy in Shannon source coding, and can be taken 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., $R > R_{\inf}$.

A more physical insight can be gained by rewriting the inequality above 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, while 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.

3.3.1 Proof of Necessity

The necessity argument is based on volume-partitioning. We first apply a coordinate transform to decouple the unstable and stable subspaces, and the state variables in the stable subspace will automatically converge to zero for any initial state without using control inputs. This essentially implies that there is no loss of generality to assume that all the eigenvalues of A are unstable. We shall 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. After k time steps, the plant dynamics expand the initial uncertainty volume m_0 by a factor

$$\left(\prod_{|\lambda_i|\geq 1} |\lambda_i|\right)^k.$$

Under data rate R, the channel can support kR bit information transmission from the coder to the decoder. The coder can effectively divide this region into 2^{kR} disjoint and exhaustive pieces, each of which is shifted by the controller. As 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, it requires that $\lim_{k\to\infty} m_k = 0$. Thus, it follows that

$$\frac{\prod_{|\lambda_i|\geq 1} |\lambda_i|}{2^R} < 1, \tag{3.6}$$

which completes the proof of necessity.

3.3.2 Proof of Sufficiency

Similarly, there is no loss of generality to assume that all the eigenvalues of A are unstable. The unknown quantity of the system to the controller is originated from the initial state x_0, whose uncertainty growth rate due to the system dynamics is given by $\prod_{|\lambda_i|\geq 1} |\lambda_i|$. To achieve the asymptotic stability, it requires the controller to estimate x_0 with the estimation error decaying at a rate faster than $\prod_{|\lambda_i|\geq 1} |\lambda_i|$. Since R denotes the communication rate of the channel to reduce the uncertainty volume, and satisfies (3.5), the controller can thus achieve this goal by appropriately designing a causal coder.

Note that (C, A) is detectable, and all the eigenvalues of A are unstable, it follows that (C, A) is observable. Then, there exists a dead-beat observer by using y_k to estimate the system state with zero estimation error after n time steps. This implies that it does not lose generality to consider the fully observed system. In particular, $C = I$ in system (3.1), the purpose of which is again to simplify the presentation. In addition, we assume that A is in a real Jordan form.

Let $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$ be the distinct unstable eigenvalues of A (if λ_i is not a real number, we exclude its conjugate λ_i^* from the list) and let m_i be the corresponding algebraic multiplicity of λ_i. Then, there exists a real transformation matrix $T \in \mathbb{R}^{n\times n}$ such that $J = TAT^{-1}$. The real Jordan canonical form [3] has the block diagonal structure

$$J = \text{diag}(J_1, \ldots, J_d) \in \mathbb{R}^{n\times n}$$

with $J_i \in \mathbb{R}^{\mu_i \times \mu_i}$ and $|\det(J_i)| = |\lambda_i|^{\mu_i}$, where

$$\mu_i = \begin{cases} m_i, & \text{if } \lambda_i \in \mathbb{R}; \\ 2m_i, & \text{otherwise.} \end{cases}$$

In summary, consider the vector system as follows:

$$x_{k+1} = Jx_k + Bu_k, \tag{3.7}$$

where the state vector

$$x_k = [x_k^{(1)}, \ldots, x_k^{(d)T}]^T \in \mathbb{R}^n$$

is partitioned in conformity with the block diagonal structure of J and the pair (J, B) is controllable. Moreover, the initial state x_0 has a known bounded support, i.e., $\|x_0\|_\infty \leq l_0$ for some $l_0 > 0$.

Lemma 3.1 *There is a positive ζ such that for any $m \in \mathbb{N}$,*

$$\|J_i^g\|_\infty \leq \zeta \sqrt{\mu_i} g^{\mu_i - 1} |\lambda_i|^g, \textit{for any } g \in \mathbb{N}. \tag{3.8}$$

Proof It is known that there exists a $\zeta > 0$, independent of J_i, μ_i, and g, such that

$$\|J_i\| \leq \zeta g^{\mu_i - 1} |\lambda_i|^g,$$

where $\|\cdot\|$ is the spectral norm induced from the Euclidean norm [4]. Together with the fact that $\|J_i\|_\infty \leq \sqrt{\mu_i}\|J_i\|$, (3.8) is immediately inferred. □

In control strategies to be developed in the sequel, we will utilize a *uniform* quantizer. Precisely, if x is a real-valued number between -1 and 1, i.e., $x \in [-1, 1]$, a mid-rise uniform quantization operator that uses N bits of precision to represent each quantizer output is expressed as

$$q_N(x) = \begin{cases} \frac{\lfloor 2^{N-1}x \rfloor + 0.5}{2^{N-1}}, & \text{if } -1 \leq x < 1; \\ 1 - \frac{0.5}{2^{N-1}}, & \text{if } x = 1, \end{cases} \tag{3.9}$$

where $\lfloor \cdot \rfloor$ is the standard floor function. The quantization interval are labeled from left to right by $I_{2^N}(0), \ldots, I_{2^N}(2^N - 1)$. Thus, if $|x| \leq M$ with M known, the quantization error induced by the above N-bit uniform quantizer is bounded as

$$|x - Mq_N(\frac{x}{M})| \leq \frac{M}{2^N}.$$

Since $R > \sum_{|\lambda_i| \geq 1} \log_2 |\lambda_i|$, we can find R_i such that $R_i > \log_2 |\lambda_i|$ for all $i \in \{1, \ldots, d\}$ and $\sum_{i=1}^d \mu_i R_i \leq R$. Then, there exist integers α_i and β such that

$$R_i \geq \frac{\alpha_i}{\beta} > \log_2 |\lambda_i|,$$

which implies that we can select an integer g satisfying that

$$\zeta \sqrt{\mu_i} g^{\mu_i - 1} |\lambda_i|^g < 2^{g\alpha_i/\beta}. \tag{3.10}$$

By using the quantized signals, both the coder and the controller can keep an identical state estimate $\widehat{x}_k$. The control law is given by $u_k = K\widehat{x}_k$, where the control gain is designed to satisfy that $J + BK$ is stable. Let the state estimation error be $\widetilde{x}_k := x_k - \widehat{x}_k$.

Divide the time into cycles with length $\tau = g\beta$, i.e., each cycle is given by $\{k\tau, k\tau + 1, \ldots, (k+1)\tau - 1\}$. Within each cycle, we use d scalar quantizers to quantize each variable of $\widetilde{x}_{k\tau}$. In particular, an $g\alpha_i$-bit uniform quantizer is used to quantize each variable of $\widetilde{x}^{(i)}_{k\tau}$. Under the above communication protocol, the average data rate is computed as

$$\frac{g}{\tau} \sum_{i=1}^{d} \mu_i \alpha_i \le \sum_{i=1}^{d} \mu_i R_i \le R. \tag{3.11}$$

Let the quantized signal be $s_k \in \mathbb{R}^n$ and $L_k = \mathrm{diag}(l^{(i)}_k I_{\mu_i})$, the estimate $\widehat{x}_k$ and $l^{(i)}_k$ are designed as follows

$$\widehat{x}_0 = 0, l^{(i)}_0 = l_0, \forall i \in \{1, \ldots, d\}, \tag{3.12}$$

$$\widehat{x}_{k\tau+j} = J\widehat{x}_{k\tau+(j-1)} + Bu_{k\tau+(j-1)}, \quad \forall j \in \{1, \ldots, \tau - 1\}, \tag{3.13}$$

$$\widehat{x}_{(k+1)\tau} = J^\tau(\widehat{x}_{k\tau} + L_k s_k) + \sum_{j=k\tau}^{(k+1)\tau-1} J^{(k+1)\tau-j-1} Bu_j, \tag{3.14}$$

$$l^{(i)}_{k+1} = \frac{l^{(i)}_k}{2^{g\alpha_i}} \left(\zeta \sqrt{\mu_i} g^{\mu_i - 1} |\lambda_i|^g\right)^\beta. \tag{3.15}$$

Assume that

$$\widetilde{x}^{(i)}_{k\tau} \in [-l^{(i)}_k, l^{(i)}_k] \times \cdots \times [-l^{(i)}_k, l^{(i)}_k].$$

Each variable of $\widetilde{x}^{(i)}_{k\tau}$ is scaled by $l^{(i)}_k$ and the scaled version is quantized by an $m\alpha_i$-bit uniform quantizer. Then, we obtain that

$$\|x^{(i)}_{(k+1)\tau} - \widehat{x}^{(i)}_{(k+1)\tau}\|_\infty \le \frac{l^{(i)}_k}{2^{g\alpha_i}} \|J_i\|^\tau_\infty \le l^{(i)}_{k+1}. \tag{3.16}$$

Let

$$\eta = \zeta \sqrt{\mu_i} g^{\mu_i - 1} |\lambda_i|^g 2^{-g\alpha_i/\beta},$$

which is strictly less than one by (3.10). Hence, it follows that

$$l^{(i)}_{k+1} \le \eta^\beta l^{(i)}_k.$$

This implies that

$$\lim_{k\to\infty} \|\widetilde{x}_{k\tau}\|_\infty \le \lim_{k\to\infty} \max_i \{l_k^{(i)}\} = 0.$$

Since τ is finite, we conclude that $\lim_{k\to\infty} \|\widetilde{x}_k\|_\infty = 0$. Moreover, it is not difficult to verify that

$$x_{k+1} = (A + BK)x_k - BK\widetilde{x}_k. \tag{3.17}$$

Note that $A + BK$ is stable, and $\lim_{k\to\infty} \|\widetilde{x}_k\|_\infty = 0$, we finally obtain that

$$\lim_{k\to\infty} \|x_k\|_\infty = 0.$$

which completes the proof. □

Remark 3.1 To attain the lower bound of (3.5), several scalar quantizers have been designed to separately quantize each variable of the error vector. The main motivation lies in that the error vector is independent of any control input, which shares the same philosophy of the celebrated separation principle. Note that the main focus is on stability, the transient performance of the closed-loop system might be typically poor as we have lifted the system for coding.

3.4 Summary

In this chapter, we briefly demonstrated the main limitation of the classical quantization approach, which fails to address the stability of the closed-loop systems under any arbitrarily large number of quantization levels. To overcome this limitation, a more rigorous information-theoretic approach was adopted. It turns out that the system dynamics poses a fundamental limitation on the communication data rate required for stabilizing noise free linear systems.

It is worthy noting that the derivation of data rate theorem bears a long history in the literature under various settings [4–9]. The pioneering work [5] interprets the quantized data as the information coder, and established the data rate theorem for a scalar system, which was further studied in [6] from an information flow point of view. For general linear vector systems, the data rate theorem was independently reported in [4, 8]. In [9], this problem was studied by using an information-theoretic approach. It is fair to say that the data rate problem over a noiseless digital channel is relatively mature.

References

1. A. Gersho, R. Gray, *Vector Quantization and Signal Compression* (Kluwer Academic Publishers, Norwell, 1991)
2. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
3. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)

4. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)
5. W. Wong, R. Brockett, Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback. IEEE Trans. Autom. Control **44**(5), 1049–1053 (1999)
6. J. Baillieul, Feedback coding for information-based control: operating near the data-rate limit, in *Proceedings of 41st IEEE Conference on Decision and Control* (2002)
7. G. Nair, R. Evans, Exponential stabilisability of finite-dimensional linear systems with limited data rates. Automatica **39**(4), 585–593 (2003)
8. S. Tatikonda, S. Mitter, Control under communication constraints. IEEE Trans. Autom. Control **49**(7), 1056–1068 (2004)
9. S. Yuksel, T. Basar, Minimum rate coding for LTI systems over noiseless channels. IEEE Trans. Autom. Control **51**(12), 1878–1887 (2006)

Chapter 4
Data Rate Theorem for Stabilization Over Erasure Channels

The quantization process induces information loss in the feedback loop which may significantly affect the operation of the closed-loop system. To achieve stabilization, the data rate must be large enough. In this chapter, we focus on a lossy communication channel where the quantized data might be randomly lost while in transit through the network. Clearly, the loss of packet also induces information loss. This intuitively implies that it will require a higher data rate for stabilization of the networked system. The fundamental question is how the finite communication data rate and packet loss jointly affect the stabilization of a linear system? This chapter is an attempt to answer this question. We are able to exactly quantify the number of bits required to compensate for the packet losses for single input linear systems by modeling the packet loss process of the channel as an i.i.d. process.

For general single input systems, the minimum data rate is explicitly given in terms of unstable eigenvalues of the open loop matrix and the packet loss rate, which explicitly reveals the amount of the additional bit rate required to counter the effect of packet losses on stabilization. In addition, sufficient data rate conditions for the mean square stabilization of multiple input systems are derived as well.

The chapter is organized as follows. The problem of interest is formulated in Sect. 4.1. The main results are presented in Sect. 4.2, where we start at the investigation of the minimum data rate for single input systems followed by multiple input systems in Sect. 4.3. Concluding remarks are made in Sect. 4.4.

4.1 Problem Formulation

Consider a discrete LTI system

$$x_{k+1} = Ax_k + Bu_k, \tag{4.1}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_4

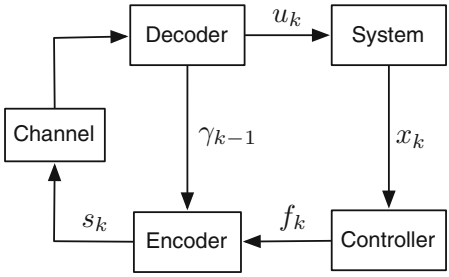

Fig. 4.1 Network configuration

where $x_k \in \mathbb{R}^n$ is the measurable state for feedback and $u_k \in \mathbb{R}^m$ is the control input. To make the problem interesting, $A \in \mathbb{R}^{n\times n}$ is assumed to be unstable.

The network configuration under consideration is described in Fig. 4.1, where the controller is collocated with the system and can access the state x_k at time k. Under a data rate R, the control signal f_k generated by the controller needs to be quantized with a packet size 2^R for each transmission. Denote the quantized signal of f_k by s_k which will be transmitted via a lossy communication channel. The packet reception/loss is represented by a binary random variable γ_k with $\gamma_k = 1$ indicating that the packet has been successfully delivered to the decoder and $\gamma_k = 0$ the loss of the packet. The packet loss process $\{\gamma_k\}_{k\geq 0}$ is assumed to be an i.i.d. process with probability distribution

$$\mathbb{P}\{\gamma_k = 1\} = 1 - p = 1 - \mathbb{P}\{\gamma_k = 0\},$$

where $p \in (0, 1)$ is named as the packet loss rate. In addition, suppose that there exists a perfect (errorless and no packet loss) channel connecting the decoder to the encoder to acknowledge the packet reception, which is used to acquire the packet delivery status (packet received or dropped) for the encoder. In comparison with the preceding chapter, the quantized data f_k might be lost during the transmission.

It should be noted that our objective is to address the issue of limited communication rather than that of limited computation and storage. Hence, the quantized signal s_k at time k is generated by allowing the encoder to access all the past and present un-quantized control input $f_0, \ldots, f_k$, the past quantized symbols $s_0, \ldots, s_{k-1}$ and packet reception status $\gamma_0, \ldots, \gamma_{k-1}$, i.e., by defining $f_0^k = \{f_0, \ldots, f_k\}$ and similarly s_0^k and γ_0^k, $s_k = \mathscr{E}_k(f_0^k, s_0^{k-1}, \gamma_0^{k-1})$, where $\mathscr{E}_k(\cdot)$ is the coder mapping at time k. Likewise, at time k, the decoder generates the control input u_k by $u_k = \mathscr{D}_k((s\gamma)_0^k, \gamma_0^k)$, where $(s\gamma)_0^k = \{s_0\gamma_0, \ldots, s_k\gamma_k\}$ and $\mathscr{D}_k(\cdot)$ is the decoder mapping at time k. Actually, the encoder and decoder to be designed later require only a finite memory.

Definition 4.1 The system (4.1) is said to be asymptotically stabilizable in the mean square sense via quantized feedback if for any finite initial state x_0, there is a control policy relying on the quantized information such that $\lim_{k\to\infty} \mathbb{E}[\|x_k\|^2] = 0$, where the mathematical expectation operator $\mathbb{E}[\cdot]$ is taken w.r.t. the random packet loss process $\{\gamma_k\}_{k\geq 0}$.

Our purpose is to derive the minimum data rate $R_{\inf}$ that requires to be communicated between the encoder and the decoder to asymptotically stabilize the system in the mean square sense. For ease of exposition, we make the following assumption:

Assumption 4.1 All the eigenvalues of A lie outside or on the unit circle and (A, B) is a controllable pair.

Remark 4.1 In general, one may like to consider systems with output feedback and A containing stable eigenvalues. As in the previous chapter, we can easily extend the results of this chapter to the general case. In addition, the following results continue to hold for single input/vector state systems driven by bounded disturbance.

4.2 Single Input Case

In this case, $u_k \in \mathbb{R}$ in (4.1). We have the following main result.

Theorem 4.1 *Consider the single input system* (4.1) *satisfying Assumption* 4.1 *and the network configuration in Fig.* 4.1 *under the i.i.d. packet loss model with packet loss rate* $p \in (0, 1)$. *Then, the system is asymptotically stabilizable in the mean square sense via quantized feedback if and only if*

(a) *The packet loss rate is less than the threshold given in* [1]*, namely,*

$$p < \frac{1}{M(A)^2}, \tag{4.2}$$

where M(A) is the Mahler measure of the LTI system.

(b) *The data rate R satisfies that*

$$R > R_{\inf} = H_T(A) + \frac{1}{2}\log_2 \frac{1-p}{1-pM(A)^2}, \tag{4.3}$$

where $H_T(A)$ *is the topological entropy of the LTI system.*

Remark 4.2 Due to the existence of packet loss, additional bits are required to counter the packet loss effect on the stabilizability of the system. They are explicitly quantified by the second term of the right hand side of (4.3) which is a function of the packet loss rate and the intrinsic entropy of the system. One can easily verify that this term is a monotonically increasing function of packet loss rate p satisfying that

$$0 < p < M(A)^{-2}.$$

For the perfect channel, i.e., $p = 0$, (4.2) is obviously satisfied and (4.3) recovers the well-known minimum data rate in (3.1). On the other hand, when the channel bandwidth is infinite, i.e., $R = \infty$, the data rate inequality of (4.3) is automatically satisfied and a necessary and sufficient condition for asymptotic stabilization in the

mean square sense is fully characterized in (4.2). It is interesting to note that the same conclusion can be found in [2, 3] where their focus is exclusively on the packet loss without the consideration of bandwidth limitation.

To sum up, Theorem 4.1 characterizes the minimum data rate for the mean square stabilization over a lossy communication channel. It contains the existing well known minimum data rate for stabilization over a communication channel without packet loss and the critical packet loss rate for stabilization over a communication channel of an infinite bandwidth as special cases.

In the rest of this section, we shall prove the above result.

4.2.1 Proof of Necessity

For brevity, we let the dimension of the state be $n = 2$. The extension to a higher order system is trivial by repeating the same arguments.

Since (A, B) is controllable, we adopt the controllable canonical form for the system (4.1), i.e.,

$$x_{k+1} = \begin{bmatrix} 0 & 1 \\ -a_2 & -a_1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k, \tag{4.4}$$

where $\det(\lambda I - A) = \lambda^2 + a_1\lambda + a_2$. Clearly, $\det(A) = a_2$. Denote the ith element of x_k by $x_k^{(i)}$, it follows from (4.4) that if the system is asymptotically stabilized in the mean square sense, then

$$\lim_{k\to\infty} \mathbb{E}[\|x_k\|^2] = \lim_{k\to\infty} \mathbb{E}[|x_k^{(1)}|^2 + |x_k^{(2)}|^2] = 2 \lim_{k\to\infty} \mathbb{E}[|x_k^{(2)}|^2].$$

Thus, the asymptotic stabilization of the system (4.4) in the mean square sense is equivalent to that of the sub-system:

$$x_{k+1}^{(2)} = -a_1 x_k^{(2)} - a_2 x_{k-1}^{(2)} + u_k, \tag{4.5}$$

where $x_{-1}^{(2)} = x_0^{(1)}$.

Observe that after the kth transmission of the quantized control signal s_{k-1}, the decoder is only able to determine that $x_k^{(2)}$ belongs to some set $\chi_k \subset \mathbb{R}$. Following (4.5), the decoder knows that $x_{k+1}^{(2)}$ is in the set

$$\bar{\chi}_{k+1} = -a_1\chi_k - a_2\chi_{k-1}$$

before the reception of the quantized control signal s_k, where for any two Lebesgue measurable sets $\mathscr{X}, \mathscr{Y} \subset \mathbb{R}^n$, we define the sum of sets by

$$a\mathscr{X} + b\mathscr{Y} \triangleq \{ax + by | x \in \mathscr{X}, y \in \mathscr{Y}\}.$$

In addition, denote the Lebesgue measure of a measurable set $\mathscr{X} \subset \mathbb{R}^n$ by $\mu(\mathscr{X})$. Applying the Brunn-Minkowski inequality [4], we obtain that

$$\begin{aligned}\mu(\chi_{k+1}^{-}) &\geq \mu(-a_1\chi_k) + \mu(-a_2\chi_{k-1}) \\ &= |a_1|\mu(\chi_k) + |a_2|\mu(\chi_{k-1}).\end{aligned} \tag{4.6}$$

Suppose now the coding is optimal in the sense that it would minimize the uncertainty of the set χ_{k+1} which $x_{k+1}^{(2)}$ will belong to. To maximize the uncertainty reduction by the quantized control signal s_k with a packet size of $R\gamma_{k+1}$ bits, the decoder will locate $x_{k+1}^{(2)}$ in one of the $2^{R\gamma_{k+1}}$ subsets of χ_{k+1}^{-}. Thus, in view of (4.5) and (4.6), it holds that

$$\mu(\chi_{k+1}) \geq \frac{1}{2^{R\gamma_{k+1}}}[|a_1|\mu(\chi_k) + |a_2|\mu(\chi_{k-1})] \tag{4.7}$$

with the equality achievable through the optimal uncertainty reduction coding.

Denote the Lebesgue measure of χ_k by $L_k = \mu(\chi_k) \geq 0$, which is stochastic due to the randomness of γ_k. Noting that $L_j \geq 0, \forall j \in \mathbb{N}$ and taking square on both sides of (4.7), the following inequalities are in force.

$$\begin{aligned}\mathbb{E}[L_{k+1}^2] &\geq \mathbb{E}[\frac{|a_1|^2}{2^{2R\gamma_{k+1}}}]\mathbb{E}[L_k^2] + \mathbb{E}[\frac{|a_2|^2}{2^{2R\gamma_{k+1}}}]\mathbb{E}[L_{k-1}^2] + 2\mathbb{E}[\frac{|a_2||a_1|}{2^{2R\gamma_{k+1}}}]\mathbb{E}[L_kL_{k-1}] \\ &\geq |a_2|^2\mathbb{E}[\frac{1}{2^{2R\gamma_{k+1}}}]\mathbb{E}[L_{k-1}^2] \\ &= |a_2|^2[p + (1-p)2^{-2R}]\mathbb{E}[L_{k-1}^2],\end{aligned} \tag{4.8}$$

where we have used the fact that the binary valued random variable γ_{k+1} is independent of L_j, $j \leq k$ since L_j is adapted to $\sigma(\{\gamma_j, j \leq k\})$, which is the σ-algebra generated by the random variables γ_j, $j \leq k$ [5]. Also it clearly holds that

$$2\{\sup|x|, x \in \mathscr{X}\} \geq \mu(\mathscr{X})$$

for any Lebesgue measurable subset $\mathscr{X} \subset \mathbb{R}$, the asymptotic stabilization in the mean square sense of $x_k^{(2)}$ implies that

$$\liminf_{k\to\infty} \mathbb{E}[L_k^2] = 0, \forall L_0 \geq 0.$$

Together with the recursion of (4.8), it infers that

$$1 > |a_2|^2[p + (1-p)2^{-2R}] \Leftrightarrow R > \log_2|a_2| + \frac{1}{2}\log_2\frac{1-p}{1-p|a_2|^2}.$$

Hence, the necessity is established.

4.2.2 Proof of Sufficiency

Define the subset $\mathscr{L}(A) \subseteq \mathbb{N}$ by

$$\mathscr{L}(A) = \begin{cases} \{l \in \mathbb{N} | \lambda_1^l \neq \lambda_2^l\}, & \text{if } \lambda_1 \neq \lambda_2; \\ \mathbb{N}, & \text{otherwise,} \end{cases}$$

where $\lambda_i, i \in \{1, 2\}$ are the unstable eigenvalues of the open loop matrix A. It is clear that $\mathscr{L}(A)$ has infinitely many elements. Since (A, B) is a controllable pair, it can be readily verified that (A^l, B) is a controllable pair if $l \in \mathscr{L}(A)$.

Divide the times $j \in \mathbb{N}$ into blocks $\{kl, \ldots, (k+1)l - 1\}$, $k \in \mathbb{N}$, of uniform duration, where $l \in \mathscr{L}(A)$ is an integer to be determined later. We shall design the control law f_j and the corresponding encoder/decoder to achieve the mean square stabilization of x_{kl}. To this end, at the start of the time block $\{kl, \ldots, (k+1)l - 1\}$, i.e., at time kl, the controller generates a control law f_{kl} and set $f_{kl+t} = 0, t \in \{1, \ldots, l-1\}$. The idea is to use the time period from kl to $(k+1)l-1$ to sequentially transmit the quantized f_{kl}. Accordingly, the decoder applies the decoded control input $u_{kl+t} = 0,\ t \in \{0, 1, \ldots, l-2\}$ and $u_{(k+1)l-1}$, which is the decoder's estimate of f_{kl}. In this case, the down-sampled system of (4.1) with the down-sampling factor l is expressed as

$$x_{(k+1)l} = A^l x_{kl} + B u_{(k+1)l-1}. \tag{4.9}$$

Due to the controllability of (A^l, B), (4.9) can be transformed into a controllable canonical form. In particular, there exists a nonsingular matrix Q such that the controllable canonical form of (4.9) is expressed by

$$x_{(k+1)l} = \begin{bmatrix} 0 & 1 \\ -\alpha_2 & -\alpha_1 \end{bmatrix} x_{kl} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{(k+1)l-1} \tag{4.10}$$

where we abuse notation by defining $x_{kl} \triangleq Q x_{kl}$ and $|\alpha_2| = |\det(A)|^l$. Let

$$\alpha_0 = \max_{k \in \{1,2\}} \binom{2}{k}.$$

Since all of the eigenvalues of A are assumed to be unstable, it follows that

$$|\alpha_k| \leq \alpha_0 |\alpha_2|, \forall k \in \{1, 2\}. \tag{4.11}$$

As in the proof of the necessity part, if we can stabilize the last element of the state vector, denoted by $x_{kl}^{(2)}$, in the mean square sense, x_{kl} will be stabilized as well, which further implies the stabilization of the original system (4.1) due to $l < \infty$. Given any data rate R that is strictly greater than the minimum data rate in (4.3), i.e., $R > R_{\inf}$ and a packet loss rate p satisfying (4.2), the control signal f_{kl} and the correspondingly decoded signal $u_{(k+1)l-1}$ are proposed as follows:

$$\begin{cases} u_{(k+1)l-1} = \tilde{f}_{kl} q_{N_k}(f_{kl}/\tilde{f}_{kl}); \\ f_{kl} = \begin{cases} 0, & \text{if } k < 2; \\ \alpha_1 x_{kl}^{(2)} + \alpha_2 x_{(k-1)l}^{(2)}, & \text{if } k \geq 2, \end{cases} \end{cases} \tag{4.12}$$

where the random variable

$$N_k = R \sum_{j=kl}^{(k+1)l-1} \gamma_j \tag{4.13}$$

represents the cumulative successfully transmitted bit number during the time block $\{kl, \ldots, (k+1)l-1\}$ and $\tilde{f}_{kl}$ is the scaling factor to capture f_{kl}, i.e., $|f_{kl}| \leq \tilde{f}_{kl}$. Since f_{kl} and the scaling factor $\tilde{f}_{kl}$ are produced at time kl, we denote $f_k \triangleq f_{kl}$ and $\tilde{f}_k \triangleq \tilde{f}_{kl}$ in the rest of the chapter for notational simplicity.

Before proceeding to the proof of stabilizability, a recursive implementation of the quantizer is first described. At the start of each time block $\{kl, \ldots, (k+1)l-1\}$, the scaled $f_k/\tilde{f}_k$ will first be quantized by the uniform quantizer in (3.9) with bit number R [6]. The quantized message s_{kl} will then be sent to the decoder. If the packet is received by the decoder, i.e., $\gamma_{kl} = 1$, the encoder and decoder determine that $f_k/\tilde{f}_k$ belongs to one of the 2^R subintervals $I_{2^R}(\cdot)$ by using the quantized signal s_{kl}. Otherwise, the packet is dropped ($\gamma_{kl} = 0$), the encoder and decoder agree that $f_k/\tilde{f}_k \in [-1, 1] \triangleq I_1$. Thus, after the first transmission at time kl, the encoder and decoder agree on the fact that $f_k/\tilde{f}_k \in I_{2^{R\gamma_{kl}}}(\cdot)$.

The remaining $l-1$ transmissions in the time block $\{kl, \ldots, (k+1)l-1\}$ are devoted to reducing the size of the subinterval $I_{2^{R\gamma_{kl}}}(\cdot)$. Specifically, at time $kl+1$, the encoder and decoder equally divide $I_{2^{R\gamma_{kl}}}(\cdot)$ into 2^R subintervals, sequentially generating the partitions of $I_{2^{R\gamma_{kl}}}(\cdot)$ of the quantizer $q_{2^{R\gamma_{kl}+2^R}}(\cdot)$. Similarly, after the second transmission of s_{kl+1}, the encoder and decoder agree on the fact that $f_k/\tilde{f}_k \in I_{2^{R\gamma_{kl}+2^{R\gamma_{kl+1}}}}(\cdot)$. Continuing the above process until the end of the time block $\{kl, \ldots, (k+1)l-1\}$, the encoder and decoder agree on the final uncertainty interval $I_{N_k}(\cdot)$ of the quantizer $q_{N_k}(\cdot)$, where N_k is given in (4.13). Using the above protocol, we may get an accurate estimate of f_{kl} and thus a better control input $u_{(k+1)l-1}$.

Similarly, we also respectively rewrite $x_{kl}^{(2)}, u_{(k+1)l-1}$ as $x_k^{(2)}$, u_k in the sequel of this chapter. Furthermore, for ease of presentation, assume that $|x_0^{(2)}| \leq \triangle_0$ and $|x_1^{(2)}| \leq \triangle_1$ and the upper bounds $\triangle_i, i \in \{0, 1\}$ are known by the encoder and decoder. This assumption can be removed by using the approach in [7]. In light of (4.10), it immediately holds that

$$x_{k+1}^{(2)} = -\alpha_1 x_k^{(2)} - \alpha_2 x_{k-1}^{(2)} + u_k, \quad k \geq 2. \tag{4.14}$$

Define the upper bound of $x_k^{(2)}$ by $\triangle_k$, i.e., $|x_k^{(2)}| \leq \triangle_k$ and $\triangle_k' = \max\{\triangle_k, \triangle_{k-1}\}$ with $\triangle_k = 0$ if $k < 0$. The synchronization of the time-varying factors $\triangle_k$ and $\triangle_k'$ is ensured at the encoder and decoder by the quantized message s_k, packet reception

acknowledgement γ_k and the fact that the initial uncertainty $\triangle_i, i \in \{0, 1\}$ is transparent to the decoder and encoder by assumption.

By (4.11), it follows that

$$|\sum_{j=0}^{1} \alpha_{j+1} x_{k-j}^{(2)}| \leq \sum_{j=0}^{1} |\alpha_{j+1}| \cdot |x_{k-j}^{(2)}| \leq 2\alpha_0|\alpha_2|\triangle_k' \triangleq \tilde{f}_k.$$

Then, inserting the control law in (4.12) to the system (4.14) obtains that

$$|x_{k+1}^{(2)}| = \tilde{f}_k|\frac{f_k}{\tilde{f}_k} - q_{N_k}(\frac{f_k}{\tilde{f}_k})| \leq \frac{\tilde{f}_k}{N_k} = \frac{2\alpha_0|\alpha_2|}{N_k}\triangle_k' \triangleq \triangle_{k+1}. \tag{4.15}$$

Note that $\tilde{f}_k$ is available for the encoder and decoder. To investigate the expansion of $\triangle_k'$, four cases are separately discussed.

Case1: If $\triangle_{k+1} > \triangle_k$, then $\triangle_{k+1}' = \triangle_{k+1} = \frac{2\alpha_0|\alpha_2|}{N_k}\triangle_k'$.

Case2: Otherwise, if $\triangle_{k+1} \leq \triangle_k$, it is obvious that $\triangle_{k+1}' = \triangle_k \leq \triangle_k'$. Under this situation, repeating the derivation of (4.15) yields that

$$|x_{k+2}^{(2)}| \leq \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\triangle_{k+1}' \leq \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\triangle_k' \triangleq \triangle_{k+2}.$$

Furthermore, if $\triangle_{k+2} > \triangle_{k+1}$, then

$$\triangle_{k+2}' = \triangle_{k+2} = \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\triangle_k'.$$

Otherwise, $\triangle_{k+2} \leq \triangle_{k+1}$, we conclude that

$$\triangle_{k+2}' = \triangle_{k+1} = \frac{2\alpha_0|\alpha_2|}{N_k}\triangle_k'.$$

Thus, at any circumstance, there must exist $i_k \in \{1, 2\}, l_k \in \{0, 1\}$ such that

$$\triangle_{k+i_k}' = \frac{2\alpha_0|\alpha_2|}{N_{k+l_k}}\triangle_k'.$$

Next, select a subsequence from $\{\triangle_k'\}$ as follows:

$$k_1 = 1, k_{j+1} = k_j + i_{k_j}.$$

Consequently, we have the recursion

$$\triangle_{k_{j+1}}' = \frac{2\alpha_0|\alpha_2|}{N_{k_j+l_{k_j}}}\triangle_{k_j}', \quad \forall j \in \mathbb{N}.$$

Taking square and then expectation on both sides of the equality yields that

$$\begin{aligned}\mathbb{E}[(\triangle'_{k_{j+1}})^2] &= (2\alpha_0)^2(|\alpha_2|)^2\mathbb{E}[\frac{1}{N^2_{k_j+l_{k_j}}}]\mathbb{E}[(\triangle'_{k_j})^2] \\ &= (2\alpha_0)^2[|\det A|^2(p+(1-p)2^{-2R})]^m\mathbb{E}[(\triangle'_{k_j})^2] \\ &\triangleq \eta\mathbb{E}[(\triangle'_{k_j})^2]. \end{aligned} \tag{4.16}$$

Here the first equality is due to that $N_{k_j+l_{k_j}}$ is independent of $\triangle'_{k_j}$ since the packet loss process is assumed to be an i.i.d. process while the second equality uses the fact that

$$\mathbb{E}[2^{-2R\sum_{j=kl}^{(k+1)l-1}\gamma_j}] = (\mathbb{E}[2^{-2R\gamma_1}])^l$$

and $|\alpha_2| = |\det(A)|^l$.

The remaining problem is to select l from $\mathscr{L}(A)$ to make

$$\eta = (2\alpha_0)^2[|\det A|^2(p+(1-p)2^{-2R})]^l < 1.$$

In particular, in light of (4.3), one can test that

$$|\det A|^2(p+(1-p)2^{-2R}) < 1.$$

Thus, it is possible to choose a sufficiently large $l \in \mathscr{L}(A)$, i.e.,

$$l > -\frac{2\log_2(2\alpha_0)}{\log_2[|\det A|^2(p+(1-p)2^{-2R})]}, \tag{4.17}$$

such that $\eta < 1$, which immediately infers that $\lim_{j\to\infty}\mathbb{E}[(\triangle'_{k_j})^2] = 0$ by (4.16).

Observe that

$$\triangle_v \le \min\{\triangle'_{v-1}, \triangle'_v, \triangle'_{v+1}\}, \quad \forall v \in \mathbb{N}$$

and $k_{j+1} - k_j = i_{k_j} \le 2$, there must exist a positive $k_{j_v} \in \{v-1, v, v+1\}$ such that $\triangle_v \le \triangle'_{k_{j_v}}$. Together with the fact that $v \to \infty$ implies $k_{j_v} \to \infty$, we obtain

$$\lim_{v\to\infty}\mathbb{E}[\triangle_v^2] \le \lim_{v\to\infty}\mathbb{E}[(\triangle'_{k_{j_v}})^2] = \lim_{k_{j_v}\to\infty}\mathbb{E}[(\triangle'_{k_{j_v}})^2] = 0.$$

Then, it follows that

$$\lim_{k\to\infty}\mathbb{E}[(x_k^{(2)})^2] \le \lim_{k\to\infty}\mathbb{E}[\triangle_k^2] = 0,$$

which further implies that $\lim_{k\to\infty}\mathbb{E}[\|x_k\|^2] = 0$.

Remark 4.3 Similar conditions as in Theorem 4.1 have been obtained for scalar systems in [8, 9]. The above result establishes the minimum data rate for stabilization in the mean square sense for general single input vector systems. It is worthy mentioning that the result can be easily generalized to m-moment stabilization, i.e.,

$$\lim_{k\to\infty} \mathbb{E}[\|x_k\|^l] = 0.$$

The idea is essentially the same as the case of mean square stabilization ($m = 2$).

4.3 Multiple Input Case

Consider system (4.1) with multiple inputs, i.e., $u_k \in \mathbb{R}^m$ with $m > 1$, and satisfying Assumption 4.1. Without loss of generality, assume that the control input matrix $B \in \mathbb{R}^{n\times m}$ has full column rank, namely, $\text{rank}(B) = m$.

Compared to the case of single input, the main difficulty in deriving the minimum data rate for stabilizing a multiple input system over a lossy channel consists of optimally allocating bits to each input. In this section, a sub-optimal bit-allocation scheme is provided to quantize the vector control input. Specifically, at any time instant k, each control input $u_k^{(j)}$ is separately quantized by a R_j bits scalar quantizer. Thus, the vector control input is to be quantized by a product quantizer. The data rate R is defined as the summation of the rates of all scalar quantizers, i.e.,

$$R = \sum_{j=1}^{m} R_j.$$

The following lemma is needed.

Lemma 4.1 *Suppose that the sequence* $\{z_k\} \subset \mathbb{R}$ *is recursively computed by the formula*

$$z_{k+1} = (1 - a_k)z_k + b_k, \quad \forall k \in \mathbb{N}$$

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 that*

$$\lim_{k\to\infty} z_k = \lim_{k\to\infty} \frac{b_k}{a_k}.$$

Proof Define

$$\phi(k + 1, j) = (1 - a_k)\phi(k, j), \phi(j, j) = 1, \quad \forall k \geq j,$$

it is easy to verify that

$$\sum_{j=0}^{k} \phi(k + 1, j + 1)a_j = 1 - \prod_{j=0}^{k}(1 - a_j) \to 1 \text{ as } k \to \infty.$$

By simple manipulation, it immediately follows that

$$z_{k+1} = \phi(k+1, 0)z_0 + \sum_{j=0}^{k} \phi(k+1, j+1)a_j \times \frac{b_j}{a_j}.$$

Noticing that $\phi(k, j) \to 0$ as $k \to \infty, \quad \forall j \in \mathbb{N}$, the result is established by the Toeplitz Lemma [5, pp.235–236].

We now apply the Wonham decomposition [10] to (A, B). Precisely, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that $\bar{A} = T^{-1}AT$ and $\bar{B} = T^{-1}B$ and take the forms

$$\bar{A} = \begin{bmatrix} A_1 & A_{12} & \cdots & A_{1m} \\ 0 & A_2 & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} b_1 & b_{12} & \cdots & b_{1m} \\ 0 & b_2 & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_m \end{bmatrix}, \tag{4.18}$$

where the matrices $A_i \in \mathbb{R}^{n_i\times n_i}$ and $b_i \in \mathbb{R}^{n_i}$, $i \in \{1, \ldots, m\}$ satisfy $\sum_{i=1}^{m} n_i = n$. Using the proposed quantizer scheme, we have the following sufficient data rate condition for stabilization in the mean square sense.

Theorem 4.2 *Consider the multiple input system* (4.1) *satisfying Assumption* 4.1 *and the network configuration in Fig.* 4.1. *Assume that the packet loss process of the lossy digital link is an i.i.d. process with packet loss rate* $p \in (0, 1)$. *Then, the system is asymptotically stabilizable in the mean square sense via a quantized feedback if*

(a) *The packet loss rate is small enough, i.e.,*

$$p < \frac{1}{\max_{i\in\{1,\ldots,m\}}\{H_T(A_i)\}}. \tag{4.19}$$

(b) *The data rate* R *satisfies that*

$$R > H_T(A) + \frac{1}{2}\sum_{i=1}^{m} \log_2 \frac{1-p}{1-pH_T(A_i)}. \tag{4.20}$$

Proof (Sketch of Proof) For brevity, assume $m = 2$ in (4.18) and the system is already given by

$$x_{k+1} = \bar{A}x_k + \bar{B}u_k.$$

As in the sufficiency proof of Theorem 4.1, by selecting a sufficiently large but finite down-sampling factor l, which is determined by the given data rate and its ability to preserve the controllability of $(\bar{A}^l, \bar{B})$, to down-sample the system, we get that

$$x_{(k+1)l} = \bar{A}^l x_{kl} + \bar{B}u_{(k+1)l-1}.$$

By denoting

$$\bar{A}^l = \begin{bmatrix} A_1^l & \bar{A}_{12} \\ 0 & A_2^l \end{bmatrix}$$

and partitioning the state vector

$$x_{kl} \triangleq [(x_k^{(1)})^T, (x_k^{(2)})^T]^T$$

in conformity with $\bar{A}^l$, the subsystems are described by

$$x_{k+1}^{(1)} = A_1^l x_k^{(1)} + b_1 u_k^{(1)} + \bar{A}_{12} x_{kl}^{(2)} + b_{12} u_k^{(2)}, \tag{4.21}$$

$$x_{k+1}^{(2)} = A_2^l x_k^{(2)} + b_2 u_k^{(2)}, \tag{4.22}$$

where the control vector is written as

$$u_{(k+1)l-1} \triangleq [u_k^{(1)}, u_k^{(2)}]^T.$$

Given any data rate R and the packet loss rate $p \in (0, 1)$ respectively satisfying (4.19) and (4.20), we separate R into $R = R_1 + R_2$ such that

$$R_i > \log_2 |\det(A_i)| + \frac{1}{2} \log_2 \frac{1-p}{1-p|\det A_i|^2}, \quad i \in \{1, 2\}.$$

Repeating the process in the sufficiency proof of Theorem 4.1 on the subsystem (4.22), we can stabilize $x_k^{(2)}$ in the mean square sense using the data rate R_2, i.e.,

$$\lim_{k\to\infty} \mathbb{E}[\|x_k^{(2)}\|^2] = 0.$$

As for the subsystem (4.21), in view of the sufficiency of the proof of Theorem 4.1, with the remaining data rate R_1, one can derive the inequality similar to (4.16):

$$\mathbb{E}[(\triangle'_{k_{j+1}})^2] \le \eta \mathbb{E}[(\triangle'_{k_j})^2] + c\mathbb{E}[\|x_{k_j}^{(2)}\|^2],$$

where $c \in \mathbb{R}$ is a constant, $\eta < 1$ and $\triangle'_k$ has the same meaning as in the proof of the sufficiency of Theorem 4.1. Together with Lemma 4.1, it gives that

$$\lim_{j\to\infty} \mathbb{E}[(\triangle'_{k_j})^2] = 0,$$

which further implies

$$\lim_{k\to\infty} \mathbb{E}[\|x_k^{(1)}\|^2] = 0.$$

Due to $m < \infty$, it finally holds that

$$\lim_{k\to\infty} \mathbb{E}[\|x_k\|^2] = 0.$$

4.4 Summary

Motivated by the fact that practical communication channels are often lossy, this chapter has studied the stabilization problem over lossy channels. We have derived a necessary and sufficient condition for the asymptotic stabilization of single input discrete LTI systems in the mean square sense over an erasure channel. The packet loss process is characterized by an i.i.d. packet loss process. The condition is explicitly given by the unstable eigenvalues of the open loop matrix and the packet loss rate and can recover the existing results on minimum data rate and critical packet loss rate in the literature. Finally, a sufficient data rate condition for the stabilization of multiple input systems in the mean square sense has also been derived.

Recently, much effort has been devoted to examining how the limited data rate of the communication channel and the random channel variation affect the stabilization of an LTI system in an appropriate sense. In [6], the authors try to use the Shannon capacity of the channel to quantify how the erasure channel affects the almost sure stabilization of the networked linear systems. However, their results are proved to be incorrect for the system with non-vanishing disturbance [11]. It is demonstrated clearly in [8, 9] that the Shannon capacity is not a good quantity to consider the mean square stabilization problem of networked control systems. In [12], the random variation of the channel follows a Markov process, which is to be studied in the next chapter.

References

1. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
2. N. Elia, Remote stabilization over fading channels. Syst. Control Lett. **54**(3), 237–249 (2005)
3. L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks. Proc. IEEE **95**(1), 163–187 (2007)
4. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
5. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)
6. S. Tatikonda, S. Mitter, Control over noisy channels. IEEE Trans. Autom. Control **49**(7), 1196–1201 (2004)
7. K. You, W. Su, M. Fu, L. Xie, Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization. Automatica **47**(1), 170–176 (2011)
8. P. Minero, M. Franceschetti, S. Dey, G. Nair, Data rate theorem for stabilization over time-varying feedback channels. IEEE Trans. Autom. Control **54**(2), 243–255 (2009)
9. N. Martins, M. Dahleh, N. Elia, Feedback stabilization of uncertain systems in the presence of a direct link. IEEE Trans. Autom. Control **51**(3), 438–447 (2006)
10. C. Chen, *Linear System: Theory and Design* (Saunders College Publishing, Philadelphia, 1984)
11. A. Matveev, A. Savkin, Comments on "Control over noisy channel" and relevant negative results. IEEE Trans. Autom. Control **50**(12), 2105–2110 (2005)
12. K. You, L. Xie, Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses. IEEE Trans. Autom. Control **56**(4), 772–785 (2011)

Chapter 5
Data Rate Theorem for Stabilization Over Gilbert-Elliott Channels

This chapter continues to investigate the minimum data rate for mean square stabilization of linear systems over a lossy digital channel. However, the packet loss process of the channel is modeled as a time-homogeneous Markov process, which is more general and realistic than the i.i.d. packet loss model. To overcome the difficulties induced by the temporal correlations of the packet loss process and stochastically time-varying data rate due to packet loss, a randomly sampled system approach is developed to study the minimum data rate for mean square stabilization. It is shown that the minimum data rate for scalar systems can be explicitly given in terms of the magnitude of the unstable mode and the transition probabilities of the Markov chain. The number of additional bits required to counter the effect of Markovian packet loss on stabilization is exactly quantified. Our result provides a means for better bandwidth utilization by jointly considering bits per sample and an effective sampling. Necessary condition and sufficient condition on the minimum data rate problem for mean square stabilization of vector systems are provided respectively and shown to be optimal under some special cases.

The chapter is organized as follows. The problem formulation is described in Sect. 5.1. Some important preliminary results are provided in Sect. 5.2. The study of the minimum data rate for scalar systems is carried out in Sect. 5.3, where we start from an unstable noise free system with bounded initial state and then proceed to the more general case with unbounded initial state and unbounded process noise. Vector systems are studied in Sect. 5.5. Conclusion remarks are drawn in Sect. 5.6.

5.1 Problem Formulation

Consider the following stochastic linear time-invariant system

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{5.1}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_5

where $x_k \in \mathbb{R}^n$ is the measurable state, $u_k \in \mathbb{R}^m$ is the input, and $w_k \in \mathbb{R}^n$ is the stochastic disturbance of zero mean. We make the following assumptions.

- The random variables x_0 and w_k have uniformly bounded $(2+\delta)$th absolute moment, i.e., $\exists \delta > 0$ and $\varpi > 0$, such that $\mathbb{E}[\|x_0\|^{2+\delta}] < \infty$ and

$$\sup_{k\in\mathbb{N}} \mathbb{E}[\|w_k\|^{2+\delta}] < \varpi.$$

- x_0 and w_k are mutually independent and have probability densities.
- The differential entropy [1] of x_0, denoted by $h(x_0)$, exists and $\exists \Delta > 0$ such that $\inf_{k\in\mathbb{N}} e^{2h(w_k)} \geq \Delta$.
- To make the problem interesting, we focus on unstable systems and that (A, B) is a stabilizable pair.

It is worth mentioning that the above assumptions have been adopted in [2] which contain the general additive white Gaussian noise assumption as a special case. Suppose that the state measurement and the controller are connected by a lossy *forward* digital channel. See Fig. 5.1 for the networked control configuration. At each time slot, the encoder measures the state, quantizes it with a packet size of R bits and transmits the quantizer output to the decoder via the *forward* channel. Due to random fading of the channel, the packet may be lost while in transit through the network. It is assumed that there is an additional perfect (without packet loss and transmission errors) *feedback* channel to send a reception/loss acknowledgement to the encoder. Neglecting transmission errors for both *forward* and *feedback* channels, we further assume that the transmission of the quantized symbol and acknowledgement can be completed within one sampling interval.

As in [3, 4], the packet loss process in the *forward* channel is modeled as a time-homogenous Markov process $\{\gamma_k\}_{k\geq 0}$, which is more general and realistic than the i.i.d. case studied in the last chapter due to the existence of temporal correlations of channel conditions. Furthermore, $\{\gamma_k\}_{k\geq 0}$ does not contain any information of the system state, suggesting that it is independent of the channel input. Similar to the previous chapter, let $\gamma_k = 1$ indicate that the packet has been successfully delivered to the decoder while $\gamma_k = 0$ corresponds to the loss of the packet. Moreover, the Markov chain has a transition probability matrix defined by

$$(\mathbb{P}\{\gamma_{k+1} = j | \gamma_k = i\})_{i,j\in\mathbb{S}} = \begin{bmatrix} 1-q & q \\ p & 1-p \end{bmatrix}, \tag{5.2}$$

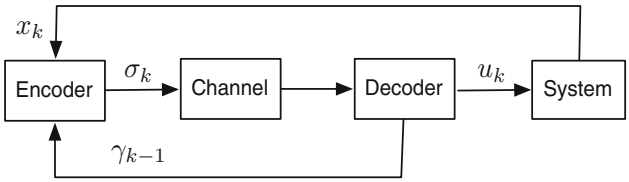

Fig. 5.1 Network configuration

where $\mathbb{S} \triangleq \{0, 1\}$ is the state space of the Markov chain. Besides, the failure rate p and recovery rate q of the channel are assumed to be strictly positive and less than one, i.e., $0 < p, q < 1$, so that the Markov chain $\{\gamma_k\}_{k\geq 0}$ is ergodic. Obviously, a smaller value of p and a larger value of q indicate a more reliable channel.

Definition 5.1 The system (5.1) with network configuration of Fig. 5.1 is said to be mean square stabilizable (MS-Stabilizable) if for any initial state x_0 and γ_0, there is a control policy relying on the quantized data such that the state of the closed-loop system is uniformly bounded in the mean square sense, i.e.,

$$\sup_{k\in\mathbb{N}} \mathbb{E}[\|x_k\|^2] < \infty,$$

where the mathematical expectation operator $\mathbb{E}[\cdot]$ is taken w.r.t. the packet loss process $\{\gamma_k\}$, the noise sequence $\{w_k\}$ and the initial state x_0.

The objective of this chapter is to find necessary and sufficient conditions on the data rate R in relation to the failure rate p and recovery rate q such that there exists a control strategy and a coding-decoding scheme to achieve MS-Stabilization of system (5.1).

5.2 Preliminaries

5.2.1 Random Down Sampling

Denote $(\Omega, \mathscr{F}, \mathbb{P})$ the common probability space for all random variables in the chapter and let $\mathscr{F}_k \triangleq \sigma(\gamma_0^k) \subset \mathscr{F}$ be an increasing sequence of σ-fields (filtration) generated by random variables $\{\gamma_0, \ldots, \gamma_k\}$. In the sequel, the terminology of almost everywhere (abbreviated as *a.e.*) is always with respect to the probability measure $\mathbb{P}$. Associated with the Markov chain $\{\gamma_k\}_{k\geq 0}$, the stochastic time sequence $\{t_k\}_{k\geq 0}$ is introduced to denote the time at which the encoder receives a packet reception acknowledgement from the decoder. Without loss of generality, let $\gamma_0 = 1$ [4]. Then, $t_0 = 1$ and t_k, $k \geq 1$ is precisely defined by

$$\begin{aligned} t_1 &= \inf\{k : k \geq 1, \gamma_k = 1\} + 1, \\ t_2 &= \inf\{k : k \geq t_1, \gamma_k = 1\} + 1, \\ &\vdots \quad \vdots \\ t_j &= \inf\{k : k \geq t_{j-1}, \gamma_k = 1\} + 1. \end{aligned} \tag{5.3}$$

In what follows, we firstly study the MS-stability of the down sampled system $\{x_{t_k}\}$, which is then used to analyze the MS-Stabilization of the origin discrete-time system.

5.2.2 Statistical Properties of Sojourn Times

By the ergodic property of the Markov chain $\{\gamma_k\}_{k\geq 0}$, t_k is finite *a.e.* for any $k \in \mathbb{N}$ [3]. Thus, the integer valued sojourn times $\{t_k^*\}_{k>0}$ to denote the time duration between

two successive packet received times are well-defined *a.e.*, where

$$t_k^* \triangleq t_k - t_{k-1} > 0. \tag{5.4}$$

With regard to the probability distribution of sojourn times $\{t_k^*\}$, we recall the following interesting result.

Lemma 5.1 ([4]) *The sojourn times $\{t_k^*\}_{k>0}$ are independent and identically distributed. Furthermore, the distribution of t_1^* is explicitly expressed as*

$$\mathbb{P}(t_1^* = i) = \begin{cases} 1-p, & i = 1; \\ pq(1-q)^{i-2}, & i > 1. \end{cases} \tag{5.5}$$

Proof The i.i.d. property directly follows from the fact that the excursion time for a Markov process to visit a state is i.i.d. [5]. Note that we have set $\gamma_0 = 1$. By the transition probability matrix in (5.2), the distribution can be easily computed. □

We shall exploit this fact to establish our results by developing a new framework by down sampling the system of (5.1) with the sampling interval equal to t_k^*. Here $\{t_k^*\}$ can be also interpreted as a "communication logic" to trigger the transmission of the packet over a channel without packet loss.

5.3 Scalar Systems

To better convey our idea, we first focus on a scalar system of the form

$$x_{k+1} = \lambda x_k + bu_k + w_k, \tag{5.6}$$

where $|\lambda| \geq 1$ and $b \neq 0$.

5.3.1 Noise Free Systems with Bounded Initial Support

Consider a noise free system as follows:

$$x_{k+1} = \lambda x_k + bu_k, \tag{5.7}$$

where the initial state $x_0 \in \mathbb{R}$ is a random variable with a known bounded support, i.e., there exists an $l_0 > 0$ such that $|x_0| \leq l_0$, and a probability density $P_{x_0}(\cdot)$.

Definition 5.2 The system (5.7) is said to be *asymptotically* MS-Stabilizable via quantized feedback if for any initial state x_0 and γ_0, there is a control policy relying on the quantized data such that the state of the closed-loop system is asymptotically driven to zero in the mean square sense, namely,

$$\lim_{k\to\infty} \mathbb{E}[\|x_k\|^2] = 0,$$

where the mathematical expectation operator $\mathbb{E}[\cdot]$ is taken w.r.t. the packet loss process $\{\gamma_k\}_{k\geq 0}$ and the initial random variable x_0.

We are now in the position to present the first main result of this chapter.

Theorem 5.1 *Consider the system* (5.7) *and the network configuration in Fig.* 5.1 *where the packet loss process of the* forward *channel is a time-homogeneous Markov process with the transition probability matrix* (5.2). *The networked system is* asymptotically *MS-Stabilizable* if and only if *the following conditions hold:*

(a) *The probability of the channel recovering from packet loss is large enough,*

$$q > 1 - 1/|\lambda|^2; \tag{5.8}$$

(b) *The data rate R satisfies the following strict inequality*

$$R > \frac{1}{2}\log_2 \mathbb{E}[|\lambda|^{2t_1^*}] \tag{5.9}$$

$$= \log_2 |\lambda| + \frac{1}{2}\log_2[1 + \frac{p(|\lambda|^2 - 1)}{1 - (1-q)|\lambda|^2}]. \tag{5.10}$$

Remark 5.1

(a) The data rate condition (5.9) has an intuitive interpretation. At the time interval $[t_k, t_{k+1})$, the square of state estimation error at the decoder grows by $|\lambda|^{2t_{k+1}^*}$. By the definition of t_k, only one packet is successfully sent to the decoder during this time interval, which can reduce the square of the estimation error at most by 2^{2R}. Thus, if the growth of the mean square estimation error

$$\mathbb{E}[|\lambda|^{2t_{k+1}^*}] = \mathbb{E}[|\lambda|^{2t_1^*}]$$

equals or exceeds 2^{2R}, i.e.,

$$\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}] \geq 1,$$

it is impossible to *asymptotically* stabilize the system in the mean square sense.

(b) Neglecting quantization effects, i.e., $R = \infty$, the inequality of (5.10) is automatically satisfied. It is interesting to note that this condition recovers the result in [3, 6].

(c) Due to stochastic packet loss, additional bits are required to asymptotically stabilize the system (5.7). When the packet loss process is specialized to an i.i.d. process, corresponding to $p = 1 - q$ in the transition probability matrix, the necessary and sufficient condition reduces to that of [7–9], see Theorem 4.1.

(d) In light of (5.10), the larger the magnitude of the unstable mode, the more bits are needed to compensate the effect of packet losses. As mentioned in Sect. 5.1, a smaller value of p and a larger value of q correspond to a more reliable channel. Thus, fewer bits are required to counter the loss effect on MS-stabilization,

which is confirmed in (5.10). For the special case that there is no packet loss, corresponding to the limiting case $p \to 0$ and $q \to 1$, the minimum data rate in (5.7) converges to the well-known data rate theorem for the stabilization of a linear system [10–12].

The following technical lemma is used to establish the result of Theorem 5.1.

Lemma 5.2 ([11]) *Let the distribution of a real valued random variable $x \in \mathbb{R}$ be absolutely continuous w.r.t. Lebesgue measure with density $P_x(\cdot)$ and has the second absolute moment, i.e., $\mathbb{E}[\|x\|^2] < \infty$. Denote the Borel measurable quantizer $c_\omega(\cdot) : \mathbb{R} \to \mathbb{R}$ with the number of quantization levels not greater than $\omega \in \mathbb{N}$.*

Then, $\forall \theta \in (\frac{1}{3}, 1)$, $\forall \omega \in \mathbb{N}$, it follows that

$$\mathbb{E}[\|x - c_\omega(x)\|^2] \geq \beta \omega^{-2} \|P_x\|_\theta^{2\theta/(1-\theta)},$$

where $\beta > 0$ is a parameter determined by θ and $\|P_x\|_\theta^{2\theta/(1-\theta)}$ is the Rényi differential entropy power of order θ [1]*, i.e.,*

$$\|P_x\|_\theta^{2\theta/(1-\theta)} = \Big(\int_{\mathbb{R}} \|P_x(x)\|^\theta dx\Big)^{\frac{2}{1-\theta}}.$$

5.3.2 Proof of Necessity

Since an acknowledgement from the decoder to indicate the packet reception/loss status will be sent to the encoder through a perfect *feedback* channel, the encoder can access the full knowledge of the decoder and recover the control produced by the decoder. By [11], the *asymptotic* MS-Stabilization of the system (5.7) is equivalent to finding a sequence of admissible quantizers $\{c^*_{\omega_k}(\cdot)\}_{k>1}$ to satisfy that

$$\lim_{k\to\infty} \mathbb{E}[\|\lambda^k x_0 - c^*_{\omega_k}(x_0)\|^2] = 0. \tag{5.11}$$

Here the admissible quantizer means that the number of quantization level $\omega_k \in \mathbb{N}$ of the quantizer $c^*_{\omega_k}(\cdot)$ is adapted to the information available at time k for the decoder.

Given any data rate R bits/transmission for the *forward* channel, denote the random variable

$$\Gamma_k = 2^{R\sum_{j=0}^k \gamma_j}$$

as the accumulative number of quantization levels that has been received by the decoder at time k. For any $\theta \in (\frac{1}{3}, 1)$, it follows that

$$\begin{aligned}\mathbb{E}[\|\lambda^k x_0 - c^*_{\Gamma_k}(x_0)\|^2] &= \mathbb{E}[\mathbb{E}[\|\lambda^k x_0 - c^*_{\Gamma_k}(x_0)\|^2 | \Gamma_k]] \\ &\geq \mathbb{E}[\beta \Gamma_k^{-2} \|P_{\lambda^k x_0}\|_\theta^{2\theta/(1-\theta)}]\end{aligned}$$

$$= \beta\mathbb{E}[\Gamma_k^{-2}](\int_{\mathbb{R}} \|P_{\lambda^k x_0}(x)\|^\theta dx)^{\frac{2}{1-\theta}}$$

$$= \beta\mathbb{E}[\Gamma_k^{-2}](\int_{\mathbb{R}} \|P_{x_0}(\lambda^{-k}x)\lambda^{-k}\|^\theta dx)^{\frac{2}{1-\theta}}$$

$$= \beta\mathbb{E}[\Gamma_k^{-2}](\lambda^k \int_{\mathbb{R}} \|P_{x_0}(y)\lambda^{-k}\|^\theta dy)^{\frac{2}{1-\theta}}$$

$$= \beta\|P_{x_0}\|_\theta^{2\theta/(1-\theta)}\mathbb{E}[\frac{\lambda^{2k}}{\Gamma_k^2}], \tag{5.12}$$

where the inequality is due to Lemma 5.2 and the change of integration variable $y = \lambda^{-k}x$ was performed in the second last equality.

Let $\xi_k = \lambda^k/\Gamma_k$, $\forall k \in \mathbb{N}$, it leads to that

$$\xi_{k+1} = \frac{\lambda}{2^{R\gamma_{k+1}}}\xi_k. \tag{5.13}$$

By (5.11), it follows from (5.12) that

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{\lambda^{2k}}{\Gamma_k^2}\right] = 0$$

since $\beta\|P_{x_0}\|_\theta > 0$.

Consider the randomly sampled system of (5.13) at stopping times $\{t_k - 1\}$ and the definition of t_k^* in (5.4), we obtain that

$$\xi_{t_{k+1}-1} = \frac{\lambda^{t_{k+1}^*}}{2^R}\xi_{t_k-1}. \tag{5.14}$$

By Theorem 4 of [4], the *asymptotic* MS-Stability of the system (5.13) in discrete time is equivalent to the *asymptotic* MS-Stability of the system (5.14) in the random times $\{t_k - 1\}$, i.e.,

$$\lim_{k\to\infty} \mathbb{E}[\|\xi_k\|^2] = 0 \Leftrightarrow \lim_{k\to\infty} \mathbb{E}[\|\xi_{t_k-1}\|^2] = 0.$$

Thus, it is sufficient to focus on the randomly sampled system of (5.14). Since in view of Lemma 5.1, the sojourn times $\{t_k^*\}$ are i.i.d., it is easy to derive that

$$\mathbb{E}[\|\xi_{t_k-1}\|^2] = \mathbb{E}\left[\frac{|\lambda|^{2\sum_{j=0}^{k-1} t_{j+1}^*}}{2^{2Rk}}|\xi_0|^2\right] = \frac{1}{2^{2R}}\left(\mathbb{E}\left[\frac{|\lambda|^{2t_1^*}}{2^{2R}}\right]\right)^k.$$

Consequently, a necessary condition to make $\lim_{k\to\infty}\mathbb{E}[\|\xi_k\|^2]=0$ is that

$$\mathbb{E}\left[\frac{|\lambda|^{2t_1^*}}{2^{2R}}\right]<1$$

by the equivalence property. By Lemma 5.1, the proof of the necessity is completed by the following arguments:

$$\mathbb{E}[|\lambda|^{2t_1^*}]=\begin{cases}\infty, & \text{if } (1-q)|\lambda|^2\geq 1;\\ |\lambda|^2[1+\frac{p(|\lambda|^2-1)}{1-(1-q)|\lambda|^2}], & \text{if } (1-q)|\lambda|^2<1.\end{cases}$$

5.3.3 *Proof of Sufficiency*

In control strategies to be developed in this section, a uniform quantizer in (3.9) will be utilized. Given any data rate R satisfying (5.10), a sequence of R-bit uniform quantizers to recursively acquire initial state information are to be designed. At time $j\in\mathbb{N}$, the encoder and decoder share a state estimator $\hat{x}_j$ based on the symbols sent via the *forward* channel and packet acknowledgement and update the estimator as follows:

$$\hat{x}_0=0,\ \hat{x}_1=\lambda l_0 q_R(\frac{x_0}{l_0}); \tag{5.15}$$

$$\hat{x}_{j+1}=(\lambda+b\mu)\hat{x}_j, j\in\{t_k,t_k+1,\ldots,t_{k+1}-2\}; \tag{5.16}$$

$$\hat{x}_{t_{k+1}}=\lambda^{t_{k+1}^*}(\hat{x}_{t_k}+l_{t_k}q_R(\frac{x_{t_k}-\hat{x}_{t_k}}{l_{t_k}}))+\sum_{j=t_k}^{t_{k+1}-1}\lambda^{t_{k+1}-1-j}b\mu\hat{x}_j, \tag{5.17}$$

where the stabilizing control gain $\mu\in\mathbb{R}$ is chosen to satisfy that $|\lambda+b\mu|<1$. The input signal is produced by $u_j=\mu\hat{x}_j,\forall j\in\mathbb{N}$. The scaling factor l_j is simultaneously updated on both sides of the channel via

$$\begin{aligned} l_1&=\frac{|\lambda|}{2^R}l_0;\\ l_{j+1}&=|\lambda|l_j, j\in\{t_k,t_k+1,\ldots,t_{k+1}-2\}; \qquad (5.18)\\ l_{t_{k+1}}&=\frac{|\lambda|^{t_{k+1}^*}}{2^R}l_{t_k}. \qquad (5.19)\end{aligned}$$

The above algorithm in the encoder and decoder is executed as follows. At the random time t_k, both the encoder and decoder have a state estimator $\hat{x}_{t_k}$ and the corresponding scaling l_{t_k}. The encoder quantizes the "normalized innovation", denoted as $\frac{x_{t_k}-\hat{x}_{t_k}}{l_{t_k}}$,

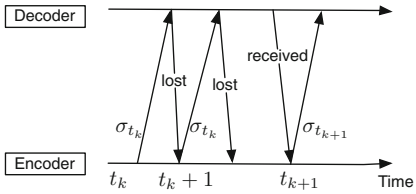

Fig. 5.2 Communication protocol

by a R-bit uniform quantizer. The quantizer output

$$\sigma_{t_k} \triangleq q_R(\frac{x_{t_k} - \hat{x}_{t_k}}{l_{t_k}})$$

is transmitted to the decoder via the *forward* channel. If σ_{t_k} is lost at time t_k's transmission, the decoder sends a packet loss acknowledgment ($\gamma_{t_k} = 0$) to the encoder and updates its estimator and scaling respectively according to (5.16) and (5.18) in the next time instant $t_k + 1$. Since we assume there is a perfect *feedback* channel to send the packet acknowledgement, the encoder can update its estimator and scaling in the same manner as the decoder. The encoder retransmits the same signal σ_{t_k} at time $t_k + 1$ until it receives the packet reception acknowledgment from the decoder ($\gamma_{t_{k+1}-1} = 1$) at time t_{k+1}. See Fig. 5.2 for illustrations. By the definition of random time t_k, the packet σ_{t_k} will be successfully delivered at time $t_{k+1} - 1$'s transmission. Then, both the encoder and decoder update their estimator and scaling according to (5.17) and (5.19) at time t_{k+1}. Thus, the synchronization between the encoder and decoder is guaranteed and this process can be continued.

Denote the estimation error by $\widetilde{x}_j = x_j - \hat{x}_j, \forall j \in \mathbb{N}$, the dynamical equation governing the error evolution is given by

$$\begin{cases} \widetilde{x}_0 = x_0,\ \widetilde{x}_1 = \lambda(x_0 - l_0 q_R(\frac{x_0}{l_0})); \\ \widetilde{x}_{j+1} = \lambda \widetilde{x}_j,\ j \in \{t_k, t_k + 1, \ldots, t_{k+1} - 2\}; \\ \widetilde{x}_{t_{k+1}} = \lambda^{t^*_{k+1}} l_{t_k}(\frac{\widetilde{x}_{t_k}}{l_{t_k}} - q_R(\frac{\widetilde{x}_{t_k}}{l_{t_k}})). \end{cases} \tag{5.20}$$

It can be shown that the quantizer does not overflow at all times, i.e., $|\widetilde{x}_j| \leq l_j, \forall j \in \mathbb{N}$. In fact, it obviously holds for $j = 0, 1$ since $|x_0| \leq l_0$ and

$$|\widetilde{x}_1| \leq \frac{|\lambda|}{2^R} l_0 = l_1.$$

Assume that $\exists k \geq 0$, such that $|\widetilde{x}_{t_k}| \leq l_{t_k}$, then

$$|\widetilde{x}_{t_{k+1}}| = |\lambda|^{t^*_{k+1}} l_{t_k} |\frac{\widetilde{x}_{t_k}}{l_{t_k}} - q(\frac{\widetilde{x}_{t_k}}{l_{t_k}})| \leq \frac{|\lambda|^{t^*_{k+1}}}{2^R} l_{t_k} = l_{t_{k+1}}.$$

Further, for any $j \in \{t_k, t_k + 1, \ldots, t_{k+1} - 2\}$, it holds that

$$|\widetilde{x}_j| = |\lambda|^{j - t_k} |\widetilde{x}_{t_k}| \leq |\lambda|^{j - t_k} l_{t_k} = l_j.$$

Thus, by induction, we have that $|\widetilde{x}_j| \leq l_j, \forall j \in \mathbb{N}$. On the other hand, the mean square of the scaling at random times $\{t_k\}$ will exponentially converge to zero.

Let $\eta = \mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}] < 1$ by (5.10), we have that

$$\lim_{k\to\infty} \mathbb{E}[l_{t_k}^2] = \lim_{k\to\infty} \mathbb{E}[\prod_{j=0}^{k-1} \frac{|\lambda|^{2t_{j+1}^*}}{2^{2R}}] l_1^2 = \lim_{k\to\infty} \left(\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}]\right)^k l_1^2 = l_1^2 \lim_{k\to\infty} \eta^k = 0,$$

where the second equality is due to that sojourn times $\{t_k^*\}$ are a sequence of i.i.d random variables by Lemma 5.1. Next, one can further derive that

$$\begin{aligned}
\mathbb{E}[\sum_{k=1}^{\infty} l_k^2] &= \sum_{k=0}^{\infty} \mathbb{E}[\sum_{j=t_k}^{t_{k+1}-1} |\lambda|^{2(j-t_k)} l_{t_k}^2] \\
&\leq \sum_{k=0}^{\infty} \mathbb{E}[t_{k+1}^* |\lambda|^{2t_{k+1}^*} l_{t_k}^2] = \mathbb{E}[t_1^* |\lambda|^{2t_1^*}] \sum_{k=0}^{\infty} \mathbb{E}[l_{t_k}^2] \qquad (5.21) \\
&= \mathbb{E}[t_1^* |\lambda|^{2t_1^*}] l_1^2 \sum_{k=0}^{\infty} \eta^k = \frac{\mathbb{E}[t_1^* |\lambda|^{2t_1^*}] l_1^2}{1-\eta} < \infty, \qquad (5.22)
\end{aligned}$$

since for $|\lambda|^2(1-q) < 1$ in (5.8), it is easy to see that $\mathbb{E}[t_1^*|\lambda|^{2t_1^*}] < \infty$. Here we have utilized the fact that the identically distributed random variable t_{k+1}^* is independent of $x_0, t_1^*, \ldots, t_k^*$ and l_{t_k} is adapted to $\sigma(x_0, t_1^*, \ldots, t_k^*)$ in (5.21).

Together with that $|\widetilde{x}_k| \leq l_k, \forall k \in \mathbb{N}$, it immediately follows from (5.22) that

$$\lim_{k\to\infty} \mathbb{E}[|\widetilde{x}_k|^2] \leq \lim_{k\to\infty} \mathbb{E}[l_k^2] = 0.$$

By iteration of (5.7) and substituting the control $u_k = \mu\hat{x}_k$ into (5.7), it yields that

$$\begin{aligned}
x_{k+1} &= \lambda x_k + b\mu\hat{x}_k \\
&= (\lambda + b\mu)x_k + b\mu\widetilde{x}_k \\
&= (\lambda + b\mu)^{k+1} x_0 + \sum_{j=0}^{k} (\lambda + b\mu)^{k-j} b\mu\widetilde{x}_j.
\end{aligned}$$

Jointly with the inequality

$$(\mathbb{E}[|x+y|^2])^{1/2} \leq (\mathbb{E}[|x|^2])^{1/2} + (\mathbb{E}[|y|^2])^{1/2},$$

it implies that

$$\lim_{k\to\infty} \mathbb{E}[|x|_{k+1}^2] \leq [\lim_{k\to\infty} \sum_{j=0}^{k} |\lambda + b\mu|^{k-j} |b\mu| (\mathbb{E}[|\widetilde{x}_j|^2])^{1/2}]^2 = 0$$

by the Toeplitz lemma [13, Sect. 6.1.2] and the result that $\lim_{k\to\infty}(\mathbb{E}[|\widetilde{x}_k|^2])^{1/2} = 0$.

5.4 General Stochastic Scalar Systems

Theorem 5.2 *Consider the system* (5.6) *and the network configuration in Fig.* 5.1 *where the packet loss process of the* forward *channel is a time-homogeneous Markov process with the transition probability matrix* (5.2)*. The networked system is MS-Stabilizable* if and only if *the following conditions hold:*

(a) *The probability of the channel recovering from losing packet is large enough,*

$$q > 1 - 1/|\lambda|^2; \tag{5.23}$$

(b) *The data rate satisfies the following strict inequality*

$$R > \log_2 |\lambda| + \frac{1}{2}\log_2\left(1 + \frac{p(|\lambda|^2 - 1)}{1 - (1-q)|\lambda|^2}\right). \tag{5.24}$$

Remark 5.2 Although the necessary and sufficient condition remains the same as that for the case of noise free systems with a finite initial state support, the proof is much more challenging. Due to the unbounded noise, uncertainties about the system state at the decoder arise from both the initial state and the noise. Thus, a completely different method is developed to establish the result.

As in [8], we will find a lower bound for the second moment of the state to establish the necessity. To this end, let

$$\Phi_k = \frac{1}{2\pi e}\mathbb{E}_{S_{k-1}}[e^{2h(x_k|S_{k-1}=s_{k-1})}]$$

be the conditional entropy power of x_k conditioned on the event $\{S_{k-1} = s_{k-1}\}$, averaged over all possible s_{k-1}. Here s_{k-1} is a particular realization of the random vector S_{k-1}. It is clear [1] that Φ_k is a lower bound of the second moment of the state x_k, i.e.,

$$\mathbb{E}[|x_k|^2] \geq \Phi_k, \quad \forall k \in \mathbb{N}.$$

The following lemma is essential to the proof of the necessity.

Lemma 5.3 *Given any data rate R bits/transmission, the following inequality holds:*

$$\mathbb{E}_{S_k|S_{k-1},\gamma_k}[e^{2h(x_k|S_k=s_k)}] \geq \frac{1}{2^{2R\gamma_k}} e^{2h(x_k|S_{k-1}=s_{k-1})}$$

Proof In view of Lemma 4.2 in [8], the proof is straightforward. □

5.4.1 Proof of Necessity

Let the data rate be R bits/transmission, it follows that

$$\begin{aligned}
\mathbb{E}_{S_k}[e^{2h(x_{k+1}|S_k=s_k)}] &= \mathbb{E}_{S_k}[e^{2h(\lambda x_k+w_k|S_k=s_k)}] \\
&\geq \mathbb{E}_{S_k}[e^{2h(\lambda x_k|S_k=s_k)}+e^{2h(w_k)}] \geq |\lambda|^2\mathbb{E}_{S_k}[e^{2h(x_k|S_k=s_k)}]+\Delta \\
&= |\lambda|^2\mathbb{E}_{S_{k-1},\gamma_k}[\mathbb{E}_{S_k|S_{k-1},\gamma_k}[e^{2h(x_k|S_k=s_k)}]]+\Delta \\
&\geq |\lambda|^2\mathbb{E}_{S_{k-1},\gamma_k}[\frac{1}{2^{2R\gamma_k}}e^{2h(x_k|S_{k-1}=s_{k-1})}]+\Delta \\
&= |\lambda|^2\mathbb{E}_{\gamma_k}[\frac{1}{2^{2R\gamma_k}}\mathbb{E}_{S_{k-1}|\gamma_k}[e^{2h(x_k|S_{k-1}=s_{k-1})}]]+\Delta. \qquad (5.25)
\end{aligned}$$

In the above, the first equality is due to that the control input u_k is measurable w.r.t. $\sigma(S_k)$ and the translation invariance property of entropy [1]. The first inequality follows from the entropy power inequality [1] and the fact that w_k is independent of (x_k, S_k). The second inequality is due to the assumption that $\Delta \leq \inf_{k\in\mathbb{N}} e^{2h(w_k)}$, whereas the last inequality follows from Lemma 5.3.

Since

$$S_{k-1} \leftrightarrow (S_{k-2}, \gamma_{k-1}) \leftrightarrow \gamma_k$$

forms a Markov chain, we can similarly derive that

$$\begin{aligned}
&\mathbb{E}_{S_{k-1}|\gamma_k}[e^{2h(x_k|S_{k-1}=s_{k-1})}] \geq |\lambda|^2\mathbb{E}_{S_{k-1}|\gamma_k}[e^{2h(x_{k-1}|S_{k-1}=s_{k-1})}]+\Delta \\
&= |\lambda|^2\mathbb{E}_{S_{k-2},\gamma_{k-1}|\gamma_k}\left[\mathbb{E}_{S_{k-1}|S_{k-2},\gamma_{k-1}}[e^{2h(x_{k-1}|S_{k-1}=s_{k-1})}]\right]+\Delta \\
&\geq \mathbb{E}_{S_{k-2},\gamma_{k-1}|\gamma_k}\left[\frac{|\lambda|^2}{2^{2R\gamma_{k-1}}}e^{2h(x_{k-1}|S_{k-2}=s_{k-2})}\right]+\Delta \\
&= \mathbb{E}_{\gamma_{k-1}|\gamma_k}\left[\frac{|\lambda|^2}{2^{2R\gamma_{k-1}}}\mathbb{E}_{S_{k-2|\gamma_{k-1},\gamma_k}}\left[e^{2h(x_{k-1}|S_{k-2}=s_{k-2})}\right]\right]+\Delta. \qquad (5.26)
\end{aligned}$$

Inserting (5.26) into (5.25) leads to that

$$\begin{aligned}
&\mathbb{E}_{S_k}[e^{2h(x_{k+1}|S_k=s_k)}] \\
&\geq \mathbb{E}_{\gamma_{k-1},\gamma_k}\left[\frac{|\lambda|^4}{2^{2R(\gamma_{k-1}+\gamma_k)}}\mathbb{E}_{S_{k-1}|\gamma_{k-1},\gamma_k}[e^{2h(x_{k-1}|S_{k-2}=s_{k-2})}]\right] \\
&\quad+\left(\mathbb{E}_{\gamma_k}[\frac{|\lambda|^2}{2^{2R\gamma_k}}]+1\right)\Delta
\end{aligned}$$

$$\geq \mathbb{E}_{\gamma_1,\ldots,\gamma_k}\left[\frac{|\lambda|^{2k}}{2^{2R(\gamma_1+\ldots+\gamma_k)}}\mathbb{E}_{S_1|\gamma_1,\ldots,\gamma_k}[e^{2h(x_1|S_0=s_0)}]\right]$$

$$+\Delta\left(\sum_{j=2}^{k}\mathbb{E}_{\gamma_j,\ldots,\gamma_k}\left[\frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\ldots+\gamma_k)}}\right]+1\right)$$

$$\geq \mathbb{E}\left[\frac{|\lambda|^{2(k+1)}}{2^{2R(\gamma_0+\gamma_1+\cdots+\gamma_k)}}\right]e^{2h(x_0)}+\Delta\left(\sum_{j=1}^{k}\mathbb{E}\left[\frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\cdots+\gamma_k)}}\right]+1\right), \quad (5.27)$$

where (5.27) is due to that the initial state is independent of the packet loss process $\{\gamma_k\}$. To study the lower bound of (5.27), let $\xi_0 = e^{2h(x_0)}$ and consider an auxiliary system as follows:

$$\xi_{k+1} = \frac{|\lambda|^2}{2^{2R\gamma_k}}\xi_k + \Delta, \quad \forall k \in \mathbb{N}. \quad (5.28)$$

Then, ξ_{k+1} is written as

$$\xi_{k+1} = \frac{|\lambda|^{2(k+1)}e^{2h(x_0)}}{2^{2R(\gamma_0+\gamma_1+\cdots+\gamma_k)}}+\Delta\left(\sum_{j=1}^{k}\frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\cdots+\gamma_k)}}+1\right).$$

Associated with the system (5.28), we introduce the following notation:

$$\alpha_k \triangleq \begin{bmatrix}\alpha_k^0\\ \alpha_k^1\end{bmatrix}, \alpha_k^j = \mathbb{E}[\xi_k 1_{\{\gamma_k=j\}}] \geq 0, \quad \forall j \in \{0,1\}. \quad (5.29)$$

Thus, the recursive equation for α_k is written as follows:

$$\alpha_{k+1}^j = \sum_{i=0}^{1}\mathbb{E}[(\frac{|\lambda|^2}{2^{2Ri}}\xi_k+\Delta)1_{\{\gamma_{k+1}=j\bigcap\gamma_k=i\}}]$$

$$= \sum_{i=0}^{1}\mathbb{E}[\mathbb{E}[(\frac{|\lambda|^2}{2^{2Ri}}\xi_k+\Delta)1_{\{\gamma_{k+1}=j\bigcap\gamma_k=i\}}|\mathscr{F}_k]]$$

$$= \sum_{i=0}^{1}\frac{|\lambda|^2}{2^{2Ri}}\mathbb{E}[\xi_k 1_{\{\gamma_k=i\}}\mathbb{E}[1_{\{\gamma_{k+1}=j\bigcap\gamma_k=i\}}|\mathscr{F}_k]]+\Delta\mathbb{E}[1_{\{\gamma_{k+1}=j\bigcap\gamma_k=i\}}]$$

$$= \sum_{i=0}^{1}\frac{|\lambda|^2 p_{ij}}{2^{2Ri}}\alpha_k^i+\Delta\sum_{i=0}^{1}p_{ij}\pi_k^i, \quad (5.30)$$

where the distribution of γ_k is defined by $\pi_k^j \triangleq \mathbb{P}(\gamma_k = j), \forall j \in \{0,1\}$.

Let

$$\psi_k = [\psi_k^0, \psi_k^1]^T$$

with $\psi_k^j = \sum_{i=0}^{1} p_{ij}\pi_k^i$ and

$$\mathscr{A} \triangleq \begin{bmatrix} |\lambda|^2 p_{00} & \frac{|\lambda|^2}{2^{2R}} p_{10} \\ |\lambda|^2 p_{01} & \frac{|\lambda|^2}{2^{2R}} p_{11} \end{bmatrix}.$$

Rewriting (5.30) in a compact form leads to the following recursion:

$$\alpha_{k+1} = \mathscr{A}\alpha_k + \Delta\psi_k. \tag{5.31}$$

Note that

$$\mathbb{E}[\xi_k] = \sum_{j=0}^{1} \alpha_k^j = \|\alpha_k\|_1,$$

where $\|\cdot\|_1$ is the standard ℓ^1 norm in $\mathbb{R}^2$. Assume the networked system (5.6) is MS-Stabilizable, it follows that

$$\sup_{k\in\mathbb{N}} \Phi_k = \sup_{k\in\mathbb{N}} \frac{1}{2\pi e} \mathbb{E}_{S_{k-1}}[e^{2h(x_k|S_{k-1}=s_{k-1})}] \le \mathbb{E}[|x_k|^2] < \infty,$$

which, together with (5.27), implies that

$$\sup_{k\in\mathbb{N}} \|\alpha_k\|_1 < \infty.$$

Moreover, due to that Markov process $\{\gamma_k\}$ is ergodic, π_k will converge to a unique stationary distribution, that is,

$$\pi_k^j \to \frac{p^{1-j}q^j}{p+q}, \quad \forall j \in \{0, 1\} \text{ as } k \to \infty.$$

Then, there exists a $\psi^j > 0$ such that

$$\psi_k^j \to \psi^j > 0, \quad \forall j \in \{0, 1\}$$

as $k \to \infty$. In view of (5.31), it follows that the spectral radius of $\mathscr{A}$ is strictly less than one since otherwise,

$$\lim_{k\to\infty} \|\alpha_k\|_1 = \infty.$$

Thus, letting $\Delta = 0$ in (5.31), we obtain that $\lim_{k\to\infty} \|\alpha_k\|_1 = 0$. Again by (5.27), it yields that

$$\lim_{k\to\infty} \mathbb{E}[\frac{|\lambda|^{2(k+1)}}{2^{2R(\gamma_0+\gamma_1+\cdots+\gamma_k)}}] = 0. \tag{5.32}$$

As in the proof of the necessity of Theorem 5.1, a necessary condition for (5.32) to hold is that

$$\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}] < 1.$$

The rest of proof follows from that of Theorem 5.1.

5.4.2 Proof of Sufficiency

We adopt the adaptive quantizer developed in [2] to capture the unbounded noise so that the upper bound of the second moment of the R-bit quantization error decays at a rate of 2^{-2R} if the random quantizer input variable x has a bounded $(2+\delta)$th moment for some $\delta > 0$. In particular, given a parameter $\rho > 1$, the R-bit quantizer generates 2^R quantization intervals labeled from left to right by $I_R(0), \ldots, I_R(2^R - 1)$. Let $I_0(0) \triangleq (-\infty, \infty)$, $I_1(0) \triangleq (-\infty, 0]$ and $I_1(1) \triangleq (0, \infty)$. If $R \geq 2$, the quantization intervals are generated by

- partitioning the set $[-1, 1]$ into 2^{R-1} intervals of equal length,
- partitioning the sets $(-\rho^{i-1}, -\rho^{i-2}]$ and $[\rho^{i-2}, \rho^{i-1})$ respectively into 2^{R-1-i}, $i \in \{2, \ldots, R-1\}$ intervals of equal length.

The two infinite length intervals $(-\infty, -\rho^{R-2}]$ and $[\rho^{R-2}, \infty)$ are respectively the leftmost and rightmost intervals of the quantizer. Let

- $\kappa_R(\sigma)$ be the half-length of interval $I_R(\sigma)$ for $\sigma \in \{1, \ldots, 2^R - 2\}$, and be equal to $\rho^R - \rho^{R-1}$ if $\sigma = 2^R - 1$ and $-(\rho^R - \rho^{R-1})$ if $\sigma = 0$.
- $Q_R(x)$ be the midpoint of interval if $x \in I_R(\sigma), \sigma \in 1, \ldots, 2^R - 2$, and be equal to ρ^R if $\sigma = 2^R - 1$ and equal to $-\rho^R$ if $\sigma = 0$.

The above quantizer allows the quantization intervals $I_R(\cdot)$ to be generated recursively by starting from $Q_2(\cdot)$. For example, quantizer intervals for $Q_{i+1}(\cdot)$, $i \geq 2$ are produced by partitioning each bounded interval of $I_i(\sigma), \sigma \in \{1, \ldots, 2^i - 2\}$ into two uniform subintervals and the unbounded interval $I_i(0) = (-\infty, -\rho^{i-2}]$ into $I_{i+1}(0) = (-\infty, -\rho^{(i+1)-2}]$, $I_{i+1}(1) = (-\rho^{(i+1)-2}, -\rho^{i-2}]$. A similar partition is applied to the other infinite subinterval $I_i(2^i - 1)$. More details can be found in [2]. For any random variables x, r and constant real number $\delta > 0$, define the functional

$$M_\delta[x, r] = \mathbb{E}[r^2 + |x|^{2+\delta} r^{-\delta}].$$

It can be verified that $M_\delta[x, r] \geq \mathbb{E}[|x|^2]$ [2]. A fundamental property of the above quantizer is given below.

Lemma 5.4 ([2, Lemma 5.2]) *Let x and $r > 0$ be random variables with $\mathbb{E}[|x|^{2+\delta}] < \infty$ for some $\delta > 0$. Given a R bit adaptive quantizer as above and $\rho > 2^{2/\delta}$, then the quantization error $x - rQ_R(x/r)$ satisfies*

$$M_\delta[x - rQ_R(x/r), r\kappa_R(\sigma)] \leq \frac{\varsigma}{2^{2R}} M_\delta[x, r],$$

where $\sigma \in \{0, \ldots, 2^R - 1\}$ *is the index of the levels of the quantizer* $Q_R(\cdot)$ *and* ς *is a constant greater than* 2 *determined only by* δ *and* ρ.

Lemma 5.5 (C_r—inequality) *Given* $a_i \geq 0$, $i \in \mathbb{N}$, *then* $\forall n \in \mathbb{N}$,

$$(\sum_{i=0}^{n} a_i)^r \leq \begin{cases} (n+1)^{r-1}(\sum_{i=0}^{n} a_i^r), & \text{if } r \geq 1; \\ \sum_{i=0}^{n} a_i^r, & \text{if } 0 \leq r < 1. \end{cases}$$

Proof If $r > 1$, it is easy to verify that $f(x) = x^r$ is a convex function for $x > 0$. This implies that

$$f(\frac{1}{n+1}\sum_{i=0}^{n} a_i) \leq \frac{1}{n+1} f(a_i). \tag{5.33}$$

Then, it follows that

$$(\sum_{i=0}^{n} a_i)^r \leq (n+1)^{r-1}(\sum_{i=0}^{n} a_i^r).$$

For $r \geq 0$, the function $(a_0 + x)^r - a_0^r - x^r$ is increasing and has the value 0 when $x = 0$. This implies that $(a_0 + a_1)^r \leq a_0^r + a_1^r$. Inductively, it is easy to obtain that

$$(\sum_{i=0}^{n} a_i)^r \leq \sum_{i=0}^{n} a_i^r.$$

Thus, the proof is completed. □

Now, we proceed to complete the proof of sufficiency.

By (5.23), there exists a $\delta > 0$ such that

$$|\lambda|^{2+\delta}(1-q) < 1.$$

Let the adaptive quantizer parameters be $\rho > 2^{2/\delta}$ and R satisfy (5.24). Now, we use the above quantizer to approach the lower bound of (5.24). To this aim, divide the integers $j \in \mathbb{N}$ into "cycles" $\{t_{k\tau}, \ldots, t_{(k+1)\tau-1}\}$, $\forall k \in \mathbb{N}$ with length $m \in \mathbb{N}$, which is determined by the available data rate and is to be specified later. The encoder and decoder simultaneously construct an estimator of the state based on the quantized symbol and packet acknowledgement as follows:

$$\hat{x}_0 = \hat{x}_1 = 0;$$

$$\hat{x}_{j+1} = (\lambda + b\mu)\hat{x}_j, j \in \{t_{k\tau}, t_{k\tau} + 1, \ldots, t_{(k+1)\tau} - 2\};$$

$$\hat{x}_{t_{(k+1)\tau}} = \lambda^{t_{(k+1)\tau}-t_{k\tau}}[\hat{x}_{t_{k\tau}} + l_k Q_{\tau R}(\frac{x_{t_{k\tau}} - \hat{x}_{t_{k\tau}}}{l_k})] + \sum_{j=t_{k\tau}}^{t_{(k+1)\tau}-1} \lambda^{t_{(k+1)\tau}-j-1} b\mu\hat{x}_j,$$

where the stabilizing control gain μ is chosen to make $|\lambda + b\mu| < 1$ and the input is formed by $u_j = \mu\hat{x}_j$, $\forall j \in \mathbb{N}$.

The quantizer $Q_{\tau R}(\cdot)$ works as follows. At random times $t_{k\tau}$, $\forall k \in \mathbb{N}$, the encoder quantizes the normalized innovation, denoted as $\frac{x_{t_{k\tau}} - \hat{x}_{t_{k\tau}}}{l_k}$, by the above described R-bit adaptive quantizer and sends the quantizer output to the decoder via the *forward* channel. Based on the definition of random times $t_{k\tau}$, the decoder will receive the packet during the time interval $[t_{k\tau+1}-1, t_{k\tau+1})$ and send a packet reception acknowledgement to the encoder. By the assumption that the transmission of the packet and acknowledgement can be finished within one sampling interval, the encoder and decoder agree on that $\frac{x_{t_{k\tau}} - \hat{x}_{t_{k\tau}}}{l_k} \in I_R(\cdot)$ at time $t_{k\tau+1}$. Then, the encoder and decoder further divide $I_R(\cdot)$ into 2^R subintervals in the manner described above. The quantizer output related to the subinterval $I_{R+R}(\cdot)$ will be sent to the decoder to further reduce the uncertainty of the normalized innovation for the decoder. By receiving the second packet in the time interval $(t_{k\tau+2} - 1, t_{k\tau+2}]$, the encoder and decoder agree on that $\frac{x_{t_{k\tau}} - \hat{x}_{t_{k\tau}}}{l_k} \in I_{R+R}(\cdot)$. Continuing the same steps and after receiving the mth packet reception acknowledgement in the time interval $[t_{(k+1)\tau} - 1, t_{(k+1)\tau})$, the encoder and decoder agree on the fact that $\frac{x_{t_{k\tau}} - \hat{x}_{t_{k\tau}}}{l_k}$ belongs to one of the subintervals $I_{\tau R}(\cdot)$.

Since $|\lambda|^{2+\delta}(1-q) < 1$, select an $\varepsilon > 1$ such that

$$|\lambda|^{(2+\delta)\varepsilon}(1-q) < 1.$$

Then, it follows from Lemma 5.1 that $\mathbb{E}[|\lambda|^{(2+\delta)\varepsilon t_1^*}] < \infty$. Using C_r-inequality in Lemma 5.5, we obtain that $\forall r \geq 0$,

$$\begin{aligned}\mathbb{E}[(t_\tau - t_0)^r] &= \mathbb{E}[(\sum_{j=1}^{\tau} t_j^*)^r] \leq (1+\tau^{r-1})\mathbb{E}[\sum_{j=1}^{\tau}(t_j^*)^r] \\ &= \tau(1+\tau^{r-1})\mathbb{E}[(t_1^*)^r] < \infty. \end{aligned} \tag{5.34}$$

Now, define

$$f(x) = \frac{|\lambda|^{(2+\delta)x}}{x^{-1-\delta}}$$

and choose $\varepsilon' > 1$ such that $1/\varepsilon' + \frac{1}{\varepsilon} = 1$. By using the Hölder inequality it follows that

$$\begin{aligned}&\mathbb{E}[(t_\tau - t_0)f(t_\tau - t_0)] = \mathbb{E}[(t_\tau - t_0)^{2+\delta}|\lambda|^{(2+\delta)(t_\tau - t_0)}] \\ &\leq (\mathbb{E}[(t_\tau - t_0)^{(2+\delta)\varepsilon'}])^{1/\varepsilon'}(\mathbb{E}[|\lambda|^{(2+\delta)\varepsilon t_1^*}])^{\tau/\varepsilon} < \infty. \end{aligned} \tag{5.35}$$

Let

$$g_{t_{k\tau}} \triangleq \sum_{j=t_{k\tau}}^{t_{(k+1)\tau}-1} \lambda^{t_{(k+1)\tau}-j-1} w_j.$$

Then, the $(2+\delta)$th absolute moment of $g_{t_{k\tau}}$ is uniformly bounded. Precisely,

$$\mathbb{E}[|g_{t_{k\tau}}|^{2+\delta}] \leq \mathbb{E}[f(t_{(k+1)\tau} - t_{k\tau}) \sum_{j=t_{k\tau}}^{t_{(k+1)\tau}-1} |w_j|^{2+\delta}] \tag{5.36}$$

$$\leq \varpi \mathbb{E}[(t_{(k+1)\tau} - t_{k\tau}) f(t_{(k+1)\tau} - t_{k\tau})] \tag{5.37}$$

$$= \varpi \mathbb{E}[(t_\tau - t_0) f(t_\tau - t_0)] \triangleq \alpha^{2+\delta}. \tag{5.38}$$

In the above, we applied the C_r inequality of Lemma 5.5 in (5.36) and (5.37) was obtained in view of the assumption that the channel variation is independent of the noise process while (5.38) is due to that

$$t_{(k+1)\tau} - t_{k\tau} = t^*_{k\tau+1} + \cdots + t^*_{(k+1)\tau}$$

has the same distribution as that of $t_\tau - t_0 = t^*_1 + \cdots + t^*_\tau$ by Lemma 5.1. By using the Hölder inequality [13], it can be shown that

$$\mathbb{E}[|g_{t_{k\tau}}|^2] \leq (\mathbb{E}[|g_{t_{k\tau}}|^{2+\delta}])^{\frac{2}{2+\delta}} \leq \alpha^2.$$

The scaling coefficient $\{l_k\}$ with $l_0 = \alpha$ is updated as follows

$$l_{k+1} = \max\{\alpha, l_k |\lambda|^{t_{(k+1)\tau} - t_{k\tau}} \kappa_{\tau R}(\sigma_{t_{k\tau}})\}. \tag{5.39}$$

The proof of [2] is extended to our case for the proof of the stability of the error dynamics in random times $t_{k\tau}$. Define the estimation error by $\widetilde{x}_j = x_j - \hat{x}_j$, the error dynamics is governed by

$$\begin{aligned} \widetilde{x}_{t_{(k+1)\tau}} &= \lambda^{t_{(k+1)\tau} - t_{k\tau}} [\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\frac{\widetilde{x}_{t_{k\tau}}}{l_k})] + g_{t_{k\tau}}, \\ \widetilde{x}_{j+1} &= \lambda \widetilde{x}_j + w_j, j \in \{t_{k\tau}, t_{k\tau}+1, \ldots, t_{(k+1)\tau} - 2\}. \end{aligned} \tag{5.40}$$

Define

$$\theta_k = \mathbb{E}[l_k^2 + |\widetilde{x}_{t_{k\tau}}|^{2+\delta} l_k^{-\delta}]$$

and let $\phi \triangleq 2^{1+\delta} > 1$. In view of the error dynamics in (5.40), one can easily derive the following result:

$$|\widetilde{x}_{t_{(k+1)\tau}}|^{2+\delta} \leq \phi\big(|\lambda^{t_{(k+1)\tau} - t_{k\tau}}|^{2+\delta} |\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta} + |g_{t_{k\tau}}|^{2+\delta}\big). \tag{5.41}$$

Denote $\vartheta \triangleq \mathbb{E}[\lambda^{2t_1^*}]$ which is bounded by (5.23) and Lemma 5.1. Then, we have the following results:

$$\begin{aligned}\mathbb{E}[|\widetilde{x}_{t_{(k+1)\tau}}|^{2+\delta} l_{k+1}^{-\delta}] &\le \phi\mathbb{E}[\frac{\lambda^{2(t_{(k+1)\tau}-t_{k\tau})}|\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta}}{[l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})]^{\delta}} + \frac{g_{t_{k\tau}}^{2+\delta}}{l_{k+1}^{\delta}}] \\ &= \phi([\mathbb{E}[\lambda^{2t_1^*}]]^m \mathbb{E}[\frac{|\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta}}{[l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})]^{\delta}}] + \mathbb{E}[\frac{g_{t_{k\tau}}^{2+\delta}}{l_{k+1}^{\delta}}]) \\ &\le \phi(\vartheta^m \mathbb{E}[\frac{|\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta}}{[l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})]^{\delta}}] + \mathbb{E}[\frac{g_{t_{k\tau}}^{2+\delta}}{\alpha^{\delta}}]) \\ &\le \phi(\vartheta^m \mathbb{E}[\frac{|\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta}}{[l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})]^{\delta}}] + \alpha^2), \end{aligned} \tag{5.42}$$

where the first equality follows from that $t_{(k+1)\tau} - t_{k\tau}$ is independent of the sigma field $\sigma\{x_0, t_1^*, \ldots, t_{k\tau}^*\}$ by Lemma 5.1 and the fact that x_0 is independent of $\{\gamma_k\}$. Also, it follows from (5.39) that

$$\mathbb{E}[l_{k+1}^2] \le \alpha^2 + \vartheta^{\tau}\mathbb{E}[|l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})|^2].$$

By summing the above two inequalities we obtain that

$$\begin{aligned}\theta_{k+1} &\le \phi(2\alpha^2 + \vartheta^{\tau}\mathbb{E}[\frac{|\widetilde{x}_{t_{k\tau}} - l_k Q_{\tau R}(\widetilde{x}_{t_{k\tau}}/l_k)|^{2+\delta}}{[l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})]^{\delta}} + |l_k \kappa_{\tau R}(\sigma_{t_{k\tau}})|^2]) \\ &\le 2\phi\alpha^2 + \phi\vartheta^{\tau}\frac{\varsigma}{2^{2\tau R}} M_{\delta}[\widetilde{x}_{t_{k\tau}}, l_k] \qquad (5.43) \\ &\le 2\phi\alpha^2 + \phi\varsigma(\frac{\vartheta}{2^{2R}})^{\tau}\theta_k, \qquad (5.44)\end{aligned}$$

where (5.43) follows from the property of the quantizer given in Lemma 5.4. By (5.24), it is clear that

$$\frac{\vartheta}{2^{2R}} = \mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}] < 1.$$

This implies that there exists an $\tau > 0$ such that $\nu \triangleq \phi\varsigma(\frac{\vartheta}{2^{2R}})^{\tau} < 1$. In addition, it immediately follows from (5.44) that $\sup_k \theta_k \le \frac{2\phi\alpha^2}{1-\nu}$. For any given $j > 1$, it reads that

$$\begin{aligned}\mathbb{E}[|\widetilde{x}_j|^2] &= \sum_{k=0}^{\infty} \mathbb{E}[|\widetilde{x}_j|^2] 1_{\{t_{k\tau} \le j < t_{(k+1)\tau}\}} \\ &\le \sum_{k=0}^{\infty} \mathbb{E}[|\lambda^{j-t_{k\tau}} \widetilde{x}_{t_{k\tau}}|^2] 1_{\{t_{k\tau} \le j < t_{(k+1)\tau}\}} + \alpha^2\end{aligned}$$

$$\leq \vartheta^{\tau} \sup_k \theta_k + \alpha^2$$
$$\leq \frac{2\vartheta^{\tau}\phi\alpha^2}{1-\nu} + \alpha^2. \tag{5.45}$$

Observing that we have designed a stabilizing controller for the estimator, it is straightforward that $\sup_{j\in\mathbb{N}} \mathbb{E}[|\hat{x}_j|^2] < \infty$, which further implies the stabilization of the system since

$$\sup_{j\in\mathbb{N}} \mathbb{E}[|x_j|^2] \leq 2(\sup_{j\in\mathbb{N}} \mathbb{E}[|\hat{x}_j|^2] + \sup_{j\in\mathbb{N}} \mathbb{E}[|\widetilde{x}_j|^2]) < \infty. \qquad \square$$

Remark 5.3 It should be noted that in [8], an i.i.d. loss process is considered. In this case, (5.25) will directly lead to that

$$\mathbb{E}_{S_k}[e^{2h(x_{k+1}|S_k=s_k)}] \geq \mathbb{E}[\frac{|\lambda|^2}{2^{R\gamma_k}}]\mathbb{E}_{S_{k-1}}[e^{2h(x_k|S_{k-1}=s_{k-1})}]] + \Delta,$$

from which the necessary condition follows by letting

$$\mathbb{E}[\frac{|\lambda|^2}{2^{R\gamma_k}}] < 1.$$

However, due to temporal correlations of the Markov process $\{\gamma_k\}$, the above arguments are no longer applicable. To overcome this difficulty, the properties of the Markov process are further exploited in the proof of the necessity.

Remark 5.4 In comparison with the proceeding chapter, an adaptive quantizer is adopted to establish the sufficiency. What makes the current problem more challenging is that the down sampling interval, denoted by $t_{(k+1)\tau} - t_{k\tau}$, is stochastic and unbounded. While in [2, 8], it is a finite constant. This implies that for their cases, the stabilization of the periodically sampled system immediately results in the stabilization of the original system, which does not trivially hold for the randomly down sampled system. Furthermore, Lemma 5.1 plays an indispensable role in proving the MS-stabilization of randomly down sampled systems.

Remark 5.5 We have developed a tool to down sample the system with a random sampling interval t_k^* so that the data rate of the down sampled system appears as a constant. It is not difficult to verify that this approach is applicable for any i.i.d. $\{t_k^*\}$ satisfying $\mathbb{E}[|\lambda|^{2t_1^*}] < \infty$. That is, it can be directly applied to networked control systems with transmission times that are driven by an i.i.d. stochastic process. Hence, our approach can jointly address the issues of minimizing the number of transmissions between the sensor and the controller/actuator as well as reducing the size of packet at each transmission, leading to a better utilization of the bandwidth of a communication network.

5.5 Vector Systems

The main challenge in stabilizing a vector system with Markovian packet loss consists of optimally allocating bits to each unstable state variable. It is worth mentioning that even for the case of i.i.d. packet loss, there is no explicit characterization of the minimum data rate for the mean square stabilization of a general vector system [8, 9].

For brevity, we consider noise free vector systems with bounded initial support in this section. A necessary condition for mean square stabilization will be given in terms of a group of inequalities that are related to unstable open-loop poles. A sub-optimal bit-allocation scheme, which is optimal for some special cases, is provided to achieve the mean square stabilization. When specialized to the case of i.i.d. packet loss, our work naturally recovers the results in [8].

5.5.1 Real Jordan Form

As in Chap. 3, we adopt a real Jordan form for the system under investigation which is briefly reviewed below, and there is no loss of generality to assume that all the eigenvalues of A lie outside or on the unit circle [10].

By the same argument in Sect. 3.3.2, consider the vector system as follows:

$$x_{k+1} = Jx_k + Bu_k, \tag{5.46}$$

where the state vector

$$x_k = [(x_k^{(1)})^T, \ldots, (x_k^{(d)})^T]^T \in \mathbb{R}^n$$

is partitioned in conformity with the block diagonal structure of J and the pair (J, B) is controllable. Moreover, the initial state x_0 has a known bounded support, i.e., $\|x_0\|_\infty \leq l_0$ for some $l_0 > 0$ and has a probability density $P_{x_0}(\cdot)$.

5.5.2 Necessity

Theorem 5.3 *Consider the system* (5.46) *and the network configuration in Fig.* 5.1 *where the packet loss process of the* forward *channel is a time-homogeneous Markov process with the transition probability matrix* (5.2)*. Let* d_{ij}, $j \in \{1, \ldots, n_i\}$ *denote the dimension of an invariant real subspace of* J_i, $i \in \{1, \ldots, d\}$*. Then, a* necessary *condition for the* asymptotic *MS-stabilization of the networked system is that for any*

$$r_i \in \mathscr{D}_i \triangleq \{d_{i1}, \ldots, d_{in_i}\}$$

and $\Sigma_r = \sum_{i=1}^{d} r_i$, *the following conditions hold:*

(a) *The probability of the channel recovering from packet loss is large enough,*

$$q > 1 - \frac{1}{(\prod_{i=1}^{d} |\lambda_i|^{2r_i})^{1/\Sigma_r}}; \tag{5.47}$$

(b) *The data rate satisfies the following strict inequality*

$$R > \frac{\Sigma_r}{2} \log_2 \mathbb{E}[(\prod_{i=1}^{d} |\lambda_i|^{2r_i})^{t_1^*/\Sigma_r}]. \tag{5.48}$$

Remark 5.6

(a) For a lossless digital channel, i.e., $p \to 0$ and $q \to 1$, it is obvious that the sojourn time t_1^* is always equal to one, (5.47) is automatically enforced, and the inequality (5.48) reduces to

$$R > \sum_{i=1}^{d} \mu_i \log_2 |\lambda_i|$$

by selecting $r_i = \mu_i, \forall i \in \{1, \ldots, d\}$. This is the well-known minimum data rate condition for stabilizing an unstable linear system [2, 10].

(b) If it is specialized to scalar systems, (5.48) becomes

$$R > \frac{1}{2} \log_2 \mathbb{E}[|\lambda_1|^{2t_1^*}],$$

which is the same as the rate condition in (5.10).

(c) When the packet loss is i.i.d., our result recovers the one derived in [8].

Proof of Theorem 5.3

Together with the results in Sect. 5.3, the proof of [8] is extended here to establish the stability of error dynamics in random time $t_{k\tau}$. By the definition of $\mathscr{D}_i$, for any $r_i \in \mathscr{D}_i$, the block J_i has an invariant real subspace, denoted by $\mathscr{H}_i$, of dimension r_i. Denote the indices of the nonempty subspaces by $\{e_1, \ldots, e_{d_r}\}$, e.g., $\mathscr{H}_{e_i} \neq \varnothing$ and the corresponding state variables w.r.t. $\mathscr{H}_{e_i}$ by $x_k^{(e_i)}$. Consider the subspace $\mathscr{H}$ formed by taking the Cartesian product of all the nonempty invariant real subspaces, i.e.,

$$\mathscr{H} = \prod_{i=1}^{d_r} \mathscr{H}_{e_i},$$

the dimension of $\mathscr{H}$ is computed as $\Sigma_r = \sum_{i=1}^{d} r_i$. Stack the unstable state variables $x_k^{(e_i)}$ to get a new vector state

$$x_k^{\mathscr{H}} = [x_k^{(e_1)T}, \ldots, x_k^{(e_{d_r})T}]^T \triangleq P x_k,$$

where P is some transformation matrix. Thus, the new vector state $x_k^{\mathscr{H}}$ evolves as follows:

$$x_{k+1}^{\mathscr{H}} = J^{\mathscr{H}} x_k^{\mathscr{H}} + PBu_k, \tag{5.49}$$

where $|\det(J^{\mathscr{H}})| = \prod_{i=1}^{d} |\lambda_i|^{r_i}$. Similarly, a lower bound of the mean square of $x_k^{\mathscr{H}}$ is chosen as

$$\Phi_k^{\mathscr{H}} = \frac{1}{2\pi e} \mathbb{E}_{S_{k-1}}[e^{2h(x_k|S_{k-1}=s_{k-1})/\Sigma_r}] \leq \mathbb{E}[\|x_k^{\mathscr{H}}\|^2].$$

Following the proof of the necessity of Theorem 5.2, we can obtain that

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{[\det(J^{\mathscr{H}})]^{2(k+1)/\Sigma_r}}{2^{2R(\gamma_0+\cdots+\gamma_k)/\Sigma_r}}\right] = 0. \tag{5.50}$$

As in Theorem 5.1, a necessary condition for (5.50) is that

$$\mathbb{E}\left[\frac{[\det(J^{\mathscr{H}})]^{2t_1^*/\Sigma_r}}{2^{2R/\Sigma_r}}\right] < 1. \tag{5.51}$$

By substituting $|\det(J^{\mathscr{H}})| = \prod_{i=1}^{d} |\lambda_i|^{r_i}$ into the above equality and after some simple manipulations, the necessity is established. □

5.5.3 Sufficiency

To achieve the *asymptotic* MS-stabilization, we propose a sub-optimal bit allocation to each state variable. The number of bits assigned to each state variable is proportional to the magnitude of its corresponding unstable mode.

Theorem 5.4 *Consider the system* (5.46) *and the network configuration in Fig.* 5.1 *where the packet loss process of the* forward *channel is a time-homogeneous Markov process with the transition probability matrix* (5.2). *The networked system is* asymptotically *MS-Stabilizable if the following conditions hold:*

(a) *The probability of the channel recovering from packet loss is large enough,*

$$q > 1 - \frac{1}{\max_{i\in\{1,\ldots,d\}} |\lambda_i|^2}; \tag{5.52}$$

(b) *The unstable eigenvalues* $(\lambda_1, \ldots, \lambda_d)$ *are inside the convex hull determined by the following constraints*

$$R > \frac{\mu_i}{2\mathbf{a}_i(R)} \log_2(\mathbb{E}[|\lambda_i|^{2t_1^*}]), \quad \forall i \in \{1, \ldots, d\}, \tag{5.53}$$

where the rate allocation vector $\mathbf{a}(R) = [\mathbf{a}_1(R), \ldots, \mathbf{a}_d(R)]$ *satisfies*

$$\begin{cases} 0 \le \mathbf{a}_j(R) \le 1; \\ \sum_{j=1}^{d} \mathbf{a}_j(R) \le 1, \quad \forall j \in \{1, \ldots, d\}; \\ \frac{R}{\mu_j}\mathbf{a}_j(R) \in \mathbb{N}. \end{cases} \tag{5.54}$$

Remark 5.7

(a) For a Markovian lossy channel with infinite bandwidth, the data rate condition of (5.53) is automatically satisfied. The probability condition of the channel recovering from packet loss in (5.52) reduces to that of [6]. That is, the sufficient condition is also necessary in this case.

(b) The sufficient condition is optimal when the magnitudes of strictly unstable eigenvalues are the same. For example, assume that there exists $d_1 \le d$ such that $|\lambda_1| = \cdots = |\lambda_{d_1}| > 1$ and $|\lambda_j| = 1, \forall j \in \{d_1 + 1, \ldots, d\}$ for the vector system of (5.46), the transition probability and rate condition in (5.52) and (5.53) are respectively written as $q > 1 - \frac{1}{|\lambda_1|^2}$ and

$$R > \frac{\sum_{i=1}^{d_1} \mu_i}{2} \log_2(\mathbb{E}[|\lambda_1|^{2t_1^*}])$$

which are the same as the necessary conditions in (5.47) and (5.48) by choosing $r_i = \mu_i, \forall i \in \{1, \ldots, d_1\}$ and $r_i = 0, \forall i \in \{d_1 + 1, \ldots, d\}$. In particular, if all the unstable eigenvalues have the same magnitude, e.g., $|\lambda_1| = \cdots = |\lambda_d|$, the sufficient condition is necessary as well.

Proof of Theorem 5.4

For any data rate R satisfying (5.53) and (5.54), the uniform quantizer of (3.9) will be adopted. Similar to the proof of Theorem 5.2, divide the integers $j \in \mathbb{N}$ into cycles $\{k\tau, \ldots, (k+1)\tau - 1\}, \forall k \in \mathbb{N}$ with length $\tau \in \mathbb{N}$, which is determined by R and is to be specified later. The communication protocol is the same as in Theorem 5.2, except that the uniform quantizer is used here. Thus, at each time $j \in \mathbb{N}$, the encoder and decoder share a state estimator $\hat{x}_j$ based on the quantized messages and packet acknowledgement, and update the estimator as follows:

$$\begin{aligned} \hat{x}_0 = 0, \quad \hat{x}_1 &= l_0 J \sigma_0; \\ \hat{x}_{j+1} &= (J + BK)\hat{x}_j, \quad j \in \{t_{k\tau}, t_{k\tau} + 1, \ldots, t_{(k+1)\tau} - 2\}; \\ \hat{x}_{t_{(k+1)\tau}} &= J^{t_{(k+1)\tau} - t_{k\tau}}[\hat{x}_{t_{k\tau}} + L_k \sigma_{t_{k\tau}}] + \sum_{j=t_{k\tau}}^{t_{(k+1)\tau}-1} J^{t_{(k+1)\tau}-j-1} BK\hat{x}_j, \end{aligned}$$

where a stabilizing control gain K is chosen to satisfy that the spectra radius of $J+BK$ is strictly less than one. Denote the ℓth component of the state vector corresponding

to the ith unstable mode by $x_k^{(i,\ell)}$, where $i \in \{1, \ldots, d\}$ and $\ell \in \{1, \ldots, \mu_i\}$. The vector σ_0 is composed by

$$\sigma_0 = [\sigma_0^{(1,1)}, \ldots, \sigma_0^{(1,\mu_1)}, \sigma_0^{(2,1)}, \ldots, \sigma_0^{(d,\mu_d)}]^T,$$

with $\sigma_0^{(i,\ell)} = q_{R\mathbf{a}_i(R)/\mu_i}(\frac{x_0^{(i,\ell)}}{l_0})$ while

$$\sigma_{t_{k\tau}} = [\sigma_{t_{k\tau}}^{(1,1)}, \ldots, \sigma_{t_{k\tau}}^{(1,\mu_1)}, \sigma_{t_{k\tau}}^{(2,1)}, \ldots, \sigma_{t_{k\tau}}^{(d,\mu_d)}]^T,$$

with

$$\sigma_{t_{k\tau}}^{(i,\ell)} = q_{\tau R\mathbf{a}_i(R)/\mu_i}(\frac{x_{t_{k\tau}}^{(i,\ell)} - \hat{x}_{t_{k\tau}}^{(i,\ell)}}{L_k^i}),$$

where $q_{R\mathbf{a}_i(R)/\mu_i}(\cdot)$ is a uniform quantizer of (3.9) using $R\mathbf{a}_i(R)/\mu_i$ bits of precision to represent quantizer output and similarly for $q_{\tau R\mathbf{a}_i(R)/\mu_i}(\cdot)$. Note that by (5.54), we have $R\mathbf{a}_i(R)/\mu_i \in \mathbb{N}$.

Denote $I_{\mu_i} \in \mathbb{R}^{\mu_i \times \mu_i}$ an identity matrix. Then, the scaling matrix L_k is given by

$$L_k = \text{diag}(L_k^1 I_{\mu_1}, \ldots, L_k^d I_{\mu_d}).$$

Moreover, L_k^i is simultaneously updated on both sides of the channel via

$$\begin{aligned} L_0^i &= \frac{\zeta\sqrt{\mu_i}|\lambda_i|}{2^{R\mathbf{a}_i(R)/\mu_i}} l_0; \\ L_{k+1}^i &= \frac{\zeta\sqrt{\mu_i}(t_{(k+1)\tau} - t_{k\tau})^{\mu_i - 1}|\lambda_i|^{t_{(k+1)\tau} - t_{k\tau}}}{2^{\tau R\mathbf{a}_i(R)/\mu_i}} L_k^i. \end{aligned} \tag{5.55}$$

By setting the control to be $u_j = K\hat{x}_j, \forall j \in \mathbb{N}$, the estimation error, i.e., $\widetilde{x}_j = x_j - \hat{x}_j$, is recursively computed by

$$\begin{aligned} \widetilde{x}_0 &= x_0, \widetilde{x}_1 = J(x_0 - l_0\sigma_0); \\ \widetilde{x}_{j+1} &= J\widetilde{x}_j, j \in \{t_{k\tau}, t_{k\tau} + 1, \ldots, t_{(k+1)\tau} - 2\}; \\ \widetilde{x}_{t_{(k+1)\tau}} &= J^{t_{(k+1)\tau} - t_{k\tau}}(\widetilde{x}_{t_{k\tau}} - L_k\sigma_{t_{k\tau}}). \end{aligned} \tag{5.56}$$

By Lemma 3.8, the quantizer will not overflow, i.e., $|x_{t_{k\tau}}^{(i,\ell)} - \hat{x}_{t_{k\tau}}^{(i,\ell)}| \leq L_k^i, \forall k \in \mathbb{N}$. In fact it obviously holds for $k = 0$. Assume $\exists k \geq 1$ such that

$$|\widetilde{x}_{t_{k\tau}}^{(i,\ell)}| \leq L_k^i, \forall i \in \{1, \ldots, d\}, \forall \ell \in \{1, \ldots, \mu_i\},$$

then $\|\widetilde{x}_{t_{k\tau}}^{(i)}\|_\infty \leq L_k^i$. By the error dynamics in (5.56), the update recursion for L_k^i in (5.55) and Lemma 3.8, it can be established that

$$|\widetilde{x}_{t_{(k+1)\tau}}^{(i,\ell)}| \leq \|\widetilde{x}_{t_{(k+1)\tau}}^{(i)}\|_\infty \leq \|J_i^{t_{(k+1)\tau} - t_{k\tau}}\|_\infty \|\widetilde{x}_{t_{k\tau}}^{(i)} - L_k^i \sigma_{t_{k\tau}}^{(i)}\|_\infty \leq L_{k+1}^i.$$

In addition, it follows from (5.52) that

$$|\lambda_i|^2(1-q) < 1, \forall i \in \{1, \ldots, d\}.$$

That is, there exists a $g > 1$ such that

$$|\lambda_i|^{2g}(1-q) < 1. \tag{5.57}$$

Then, it is straightforward that $\mathbb{E}[|\lambda_i|^{2gt_1^*}] < \infty$ by Lemma 5.1. Moreover, given any sequence $\{g_j\}_{j\geq 0} \subset (1, g]$ such that $\lim_{j\to\infty} g_j = 1$, it is easy to show that

$$\left(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}}\right)^{g_j} \leq (|\lambda_i|^{2t_1^*})^{g_j} \leq (|\lambda_i|^{2t_1^*})^{g}.$$

By the dominated convergence theorem [13], it follows that

$$\lim_{j\to\infty} \mathbb{E}\left[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^{g_j}\right] = \mathbb{E}\left[\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}}\right] < 1 \text{ by (5.53)}.$$

Thus, there exists a $c \in (1, g]$ such that the following inequality holds:

$$\mathbb{E}\left[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^{c}\right] < 1. \tag{5.58}$$

Select a $c' > 1$ such that $1/c' + \frac{1}{c} = 1$, we obtain that

$$\begin{aligned}
&\mathbb{E}[(L_{k+1}^i)^2] \\
&= \zeta^2\mu_i\mathbb{E}[\frac{(t_{(k+1)\tau} - t_{k\tau})^{2(\mu_i-1)}|\lambda_i|^{2(t_{(k+1)\tau}-t_{k\tau})}}{2^{2\tau R\mathbf{a}_i(R)/\mu_i}}]\mathbb{E}[(L_k^i)^2] \\
&\leq \zeta^2\mu_i(\mathbb{E}[(t_{(k+1)\tau} - t_{k\tau})^{2c'(\mu_i-1)}])^{1/c'}(\mathbb{E}[(\frac{|\lambda_i|^{2(t_{(k+1)\tau}-t_{k\tau})}}{2^{2\tau R\mathbf{a}_i(R)/\mu_i}})^{c}])^{1/c}\mathbb{E}[(L_k^i)^2] \\
&\leq C_1(\mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^{c}])^{m/c}\mathbb{E}[(L_k^i)^2] \\
&\triangleq \eta_i\mathbb{E}[(L_k^i)^2].
\end{aligned} \tag{5.59}$$

Here the first equality is obtained from (5.55) and the fact that $t_{(k+1)\tau} - t_{k\tau}$ is independent of L_k^i by Lemma 5.1. The Hölder inequality was applied in the first inequality. The constant

$$C_1 \triangleq \zeta^2\mu_i(\mathbb{E}[(t_{(k+1)\tau} - t_{k\tau})^{2c'(\mu_i-1)}])^{1/c'}$$

in the second inequality is finite by (5.34). In light of (5.58), there exists an $\tau > 0$ such that

$$\eta_i = C_1(\mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^c])^{\tau/c} < 1. \tag{5.60}$$

Using the above inequality and Lemma 3.8, we can further derive that

$$\begin{aligned}\mathbb{E}[\sum_{k=1}^{\infty}\|\widetilde{x}_k^{(i)}\|_\infty^2] &\le \zeta^2\mu_i\sum_{k=0}^{\infty}\mathbb{E}[\sum_{j=t_{k\tau}}^{t_{(k+1)\tau}-1}(j-t_{k\tau})^{2(\mu_i-1)}|\lambda_i|^{2(j-t_{k\tau})}\|\widetilde{x}_{t_{k\tau}}^{(i)}\|_\infty^2] \\ &\le \zeta^2\mu_i\mathbb{E}[(t_\tau-t_0)^{2\mu_i-1}|\lambda_i|^{2(t_\tau-t_0)}]\sum_{k=0}^{\infty}\mathbb{E}[\|\widetilde{x}_{t_{k\tau}}^{(i)}\|_\infty^2] \qquad (5.61)\\ &\le C_2\sum_{k=0}^{\infty}\mathbb{E}[(L_k^i)^2] \le C_2(L_0^i)^2\sum_{k=0}^{\infty}\eta_i^k. \qquad (5.62)\end{aligned}$$

In the above, the second inequality follows from Lemma 3.8. The inequality in (5.61) is due to that for all $j \in \{t_{k\tau}, t_{k\tau}+1, \ldots, t_{(k+1)\tau}\}$, it holds

$$(j-t_{k\tau})^{2(\mu_i-1)}|\lambda_i|^{2(j-t_{k\tau})} \le (t_{(k+1)\tau}-t_{k\tau})^{2(\mu_i-1)}|\lambda_i|^{2(t_{(k+1)\tau}-t_{k\tau})}$$

and $t_{(k+1)\tau}-t_{k\tau}$ is of the same distribution as $t_\tau - t_0$.

Choosing $g' > 1$ such that $1/g' + \frac{1}{g} = 1$ and using Hölder inequality, then

$$\begin{aligned}C_2 &= \zeta^2\mu_i\mathbb{E}[(t_\tau-t_0)^{2\mu_i-1}|\lambda_i|^{2(t_\tau-t_0)}] \\ &\le \zeta^2\mu_i(\mathbb{E}[(t_\tau-t_0)^{g'(2\mu_i-1)}])^{1/g'}(\mathbb{E}[|\lambda_i|^{2gt_1^*}])^{\tau/g} \\ &< \infty \text{ by (5.34) and (5.37).}\end{aligned}$$

By (5.60) and (5.62), we obtain that

$$\mathbb{E}[\sum_{k=1}^{\infty}\|\widetilde{x}_k^{(i)}\|_\infty^2] < \infty.$$

This immediately implies that

$$\lim_{k\to\infty}\mathbb{E}[\|\widetilde{x}_k^{(i)}\|_\infty^2] = 0,$$

which further concludes that

$$\lim_{k\to\infty}\mathbb{E}[\|\widetilde{x}_k^{(i)}\|^2] \le \mu_i\lim_{k\to\infty}\mathbb{E}[\|\widetilde{x}_k^{(i)}\|_\infty^2] = 0.$$

Consequently, we have shown that $\lim_{k\to\infty}\mathbb{E}[\|\widetilde{x}_k\|^2] = 0$. Under the designed stabilizing controller, it is trivial that $\lim_{k\to\infty}\mathbb{E}[\|x_k\|^2] = 0$. The rest of the proof follows from similarly as in [8] and details are omitted. □

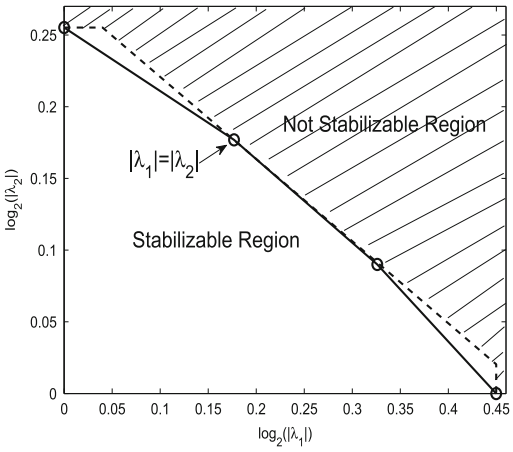

Fig. 5.3 Stabilizable and unstabilizable regions for the example in Sect. 5.5.4

5.5.4 An Example

In this section, an example is included to examine the gap between the necessary and sufficient conditions. Let the transition probabilities of the Markov process be $p = 1/2, q = 2/3$ and the data rate be $R = 1$. Consider an unstable system with distinct eigenvalues $\lambda_1 \in \mathbb{R}$, $\lambda_2 \in \mathbb{C}$ and $\mu_1 = 1, \mu_2 = 2$. The stabilizable and unstabilizable regions respectively determined by Theorems 5.3 and 5.4 are plotted in Fig. 5.3. It is clear from the figure that they are optimal for the three cases respectively corresponding to that $|\lambda_1| = 1$, $|\lambda_1| = |\lambda_2|$ and $|\lambda_2| = 1$. It also shows that the necessary condition is almost sufficient.

5.6 Summary

Packet loss and data rate constraints are two important issues of networked control systems. In this chapter, we have investigated their joint effect on the mean square stabilization of networked linear systems where the digital channel is subject to Markovian packet losses. The temporal correlations of the packet loss process posed significant challenges to the study of the minimum data rate which were overcome by converting the networked system with random packet loss into a randomly sampled system. The minimum data rate for the scalar case was then derived which is explicitly given in terms of the magnitude of the unstable mode and the transition probabilities of the Markov chain. The result exactly quantifies the joint effect of Markovian packet loss and finite communication data rate on the mean square stabilization of linear scalar systems and contains existing results on packet loss probability and data rate for stabilization as special cases.

We have also studied the mean square stabilization problem for vector systems. Necessary condition and sufficient condition were respectively derived and shown to be optimal for some special cases. This approach can also be directly applied to NCSs where transmission times are driven by an i.i.d. stochastic process with the consideration of reducing the size of the data to be transmitted at each transmission. Till now, the case of general vector systems is largely open, and deserves further consideration. We note that the theory of Markov jump linear systems is adopted to study this problem in [8].

References

1. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
2. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)
3. M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses. Automatica **43**(4), 598–607 (2007)
4. L. Xie, L. Xie, Stability analysis of networked sampled-data linear systems with Markovian packet losses. IEEE Trans. Autom. Control **54**(6), 1368–1374 (2009)
5. S. Meyn, R. Tweedie, J. Hibey, *Markov Chains and Stochastic Stability* (Springer, London, 1996)
6. V. Gupta, N. Martins, J. Baras, Optimal output feedback control using two remote sensors over erasure channels. IEEE Trans. Autom. Control **54**(7), 1463–1476 (2009)
7. K. You, L. Xie, Minimum data rate for mean square stabilization of discrete LTI systems over lossy channels. IEEE Trans. Autom. Control **55**(10), 2373–2378 (2010)
8. P. Minero, M. Franceschetti, S. Dey, G. Nair, Data rate theorem for stabilization over time-varying feedback channels. IEEE Trans. Autom. Control **54**(2), 243–255 (2009)
9. N. Martins, M. Dahleh, N. Elia, Feedback stabilization of uncertain systems in the presence of a direct link. IEEE Trans. Autom. Control **51**(3), 438–447 (2006)
10. S. Tatikonda, S. Mitter, Control under communication constraints. IEEE Trans. Autom. Control **49**(7), 1056–1068 (2004)
11. G. Nair, R. Evans, Exponential stabilisability of finite-dimensional linear systems with limited data rates. Automatica **39**(4), 585–593 (2003)
12. K. You, W. Su, M. Fu, L. Xie, Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization. Automatica **47**(1), 170–176 (2011)
13. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)

Chapter 6
Stabilization of Linear Systems Over Fading Channels

Fading channels are often encountered in wireless communications and have attracted a lot of attentions in the study of networked control recently. It has been shown in [1] that the minimum network requirement for stabilizability through controller design is nonconvex and has no explicit solution in general networked MIMO control over fading channels. Only the minimum capacity for state feedback stabilization of a single-input plant over a single fading channel is given in [1]. In this chapter, we focus on the mean square stabilization problem of MIMO plants over multiple fading channels via both state and output feedback.

The chapter is organized as follows. The mean square stabilization problem for an NCS over an input network with fading channel(s) is formulated in Sect. 6.1, followed by a preliminary lemma on stabilization over predefined fading channels. Section 6.2 presents the network requirement for stabilizability via state feedback, where necessary and sufficient conditions on the network are given as the main results of this chapter. The minimum mean square capacity is provided for stabilizability under the serial transmission strategy (STS) and the parallel transmission strategy (PTS), respectively. Section 6.3 deals with the network requirement for stabilizability via output feedback. Both SISO and triangularly decoupled MIMO cases are considered. The extension to stabilization over output fading channels is given in Sect. 6.4, followed by an application in multi-vehicle platooning. In Sect. 6.5, we analyze the possible benefits of introducing pre- and post-channel filters and channel feedback to the NCS. Section 6.6 considers the feedback stabilization and performance design of an SISO NCS with a stochastic channel contaminated by both multiplicative and additive noises. We provide a necessary and sufficient condition on the network for mean square stabilizability, a suboptimal algorithm for performance design along with a numerical example. Section 6.7 summarizes the chapter.

6.1 Problem Formulation

Consider a discrete-time NCS as depicted in Fig. 6.1, where an unreliable network with fading channel(s) is placed in the path from the controller to the plant.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_6

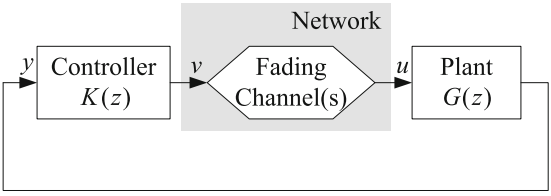

Fig. 6.1 Control over input fading channel(s)

Assume that the plant $G(z)$ is strictly proper and has a state-space representation:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k, \end{aligned} \tag{6.1}$$

where $x_k \in \mathbb{R}^n$ is the plant state, $u_k \in \mathbb{R}^m$ is the plant input, and $y_k \in \mathbb{R}^\ell$ is the measured output. Without loss of generality, assume that A is unstable, $B = [B_1\ B_2, \ldots, B_m]$ has full-column rank, $C = [C_1^T\ C_2^T, \ldots, C_\ell^T]^T$ has full-row rank, and the triplet (A, B, C) is stabilizable and detectable.

The controller is assumed to be time invariant, and has a proper transfer function $K(z)$ and a state-space realization:

$$\begin{aligned} x_{K,k+1} &= A_K x_{K,k} + B_K y_k, \\ v_k &= C_K x_{K,k} + D_K y_k, \end{aligned} \tag{6.2}$$

where $v_k \in \mathbb{R}^m$ is the controller output. The dimension of the controller state $x_{K,k}$ is not specified a priori.

The model of the fading channel(s) is given in the following memoryless multiplicative form:

$$u_k = \xi_k v_k, \tag{6.3}$$

where u_k is directly applied to the plant,[1] and $\xi_k \in \mathbb{R}^{m\times m}$ represents the channel fading and has the diagonal structure:

$$\xi_k = \operatorname{diag}\left\{\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{m,k}\right\}. \tag{6.4}$$

It is assumed that $\xi_{i,k},\ i = 1, 2, \ldots, m$, are scalar-valued white noise processes with

$$\mu_i \triangleq \mathscr{E}\{\xi_{i,k}\},\ \sigma_{ij} \triangleq \mathscr{E}\{(\xi_{i,k} - \mu_i)(\xi_{j,k} - \mu_j)\}, \tag{6.5}$$

satisfying $\mu_i \neq 0, \sigma_{ii} > 0$, and $\sigma_{ij} = \sigma_{ji}$, for all $i, j = 1, 2, \ldots, m$. We further denote $\sigma_i \triangleq \sqrt{\sigma_{ii}}$ and

$$\Sigma \triangleq [\sigma_{ij}]_{i,j=1,2,\ldots,m},\quad \Pi \triangleq \operatorname{diag}\{\mu_1, \mu_2, \ldots, \mu_m\},\quad \Lambda \triangleq \operatorname{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_m\}. \tag{6.6}$$

It is easy to see that Σ is positive semidefinite.

[1] The NCS with pre- and post-channel filters will be further discussed in Sect. 6.5.

Remark 6.1 In model (6.3)–(6.4), if the ith and jth components of v_k are sent over the same fading channel, whose channel fading is constant over each time step of the underlying discrete-time system,[2] then we can set $\xi_{i,k} = \xi_{j,k}$. The additive noise is not included in (6.3), since the main focus here is the stabilization issue.[3]

Denote the overall system state by $x'_k = [x_k^T \; x_{K,k}^T]^T$, then the closed-loop NCS in Fig. 6.1 can be written into

$$x'_{k+1} = [\tilde{A} + \tilde{B}\xi_k\tilde{C}]x'_k, \tag{6.7}$$

where

$$\tilde{A} = \begin{bmatrix} A & 0 \\ B_K C & A_K \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad \tilde{C} = [D_K C \;\; C_K]. \tag{6.8}$$

We shall concentrate on the mean square stabilization defined next.

Definition 6.1 ([1]) Consider the NCS in Fig. 6.1, and denote $\Xi_k \triangleq \mathscr{E}[x'_k x_k'^T]$. The plant (6.1) is stabilized in the mean square sense by the controller (6.2), if, for any initial state x'_0, Ξ_k is well defined for all $k \geq 0$ and $\lim_{k\to\infty} \Xi_k = 0$.

The model (6.3) and (6.4) of the fading channel(s) is said to be predefined if μ_i, σ_{ij} in (6.5) are fixed and known for all $i, j = 1, 2, \ldots, m$.

Let $\tilde{B} = [\tilde{B}_1 \; \tilde{B}_2, \ldots, \tilde{B}_m]$ and $\tilde{C} = [\tilde{C}_1^T \; \tilde{C}_2^T, \ldots, \tilde{C}_m^T]^T$. The next lemma will be used in the later developments, which extends the stability criterion for systems with independent multiplicative noises presented in [1] and Chap. 9 of [3] to that with possibly correlated multiplicative noises.

Lemma 6.1 *Consider the NCS in Fig. 6.1 with predefined fading channel(s) modeled by (6.3)–(6.4). The following statements are equivalent.*

(i) *The plant (6.1) or the triplet (A, B, C) can be stabilized in the mean square sense via controller of the form (6.2).*

(ii) *There exists a set of A_K, B_K, C_K, D_K such that the sequence $\{\Xi_k\}_{k\geq 0}$ computed by*

$$\Xi_{k+1} = (\tilde{A} + \tilde{B}\Pi\tilde{C})\Xi_k(\tilde{A} + \tilde{B}\Pi\tilde{C})^T + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\tilde{B}_j\tilde{C}_j\Xi_k\tilde{C}_i^T\tilde{B}_i^T \tag{6.9}$$

with any $\Xi_0 \geq 0$ is convergent to 0 as k approaches ∞.

(iii) *There exists a set of A_K, B_K, C_K, D_K such that $\rho(\Psi) < 1$, where*

$$\Psi = (\tilde{A} + \tilde{B}\Pi\tilde{C}) \otimes (\tilde{A} + \tilde{B}\Pi\tilde{C}) + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}(\tilde{B}_i\tilde{C}_i) \otimes (\tilde{B}_j\tilde{C}_j). \tag{6.10}$$

---

[2] This is a valid assumption when the fading is coherent over the sampling interval of the system [2].

[3] In Sect. 6.6, we shall revisit the effect of channel additive noise under the assumption that the channel input power is bounded by a predefined level.

(iv) *There exist a set of* A_K, B_K, C_K, D_K *and* $P > 0$ *such that*

$$P > (\tilde{A} + \tilde{B}\Pi\tilde{C})^T P(\tilde{A} + \tilde{B}\Pi\tilde{C}) + \tilde{C}^T J\tilde{C}, \tag{6.11}$$

where $J = \Sigma \odot (\tilde{B}^T P\tilde{B}) > 0$.

(v) *There exist a set of* A_K, B_K, C_K, D_K *and* $W > 0$ *such that*

$$W > (\tilde{A} + \tilde{B}\Pi\tilde{C})W(\tilde{A} + \tilde{B}\Pi\tilde{C})^T + \tilde{B}H\tilde{B}^T, \tag{6.12}$$

where $H = \Sigma \odot (\tilde{C}W\tilde{C}^T)$.

Proof (i)⇔(ii): It follows directly from Definition 6.1.

(ii)⇔(iii): Based on Lemma A.4, the recursion (6.9) can be written into

$$\mathrm{vec}(\Xi_{k+1}) = \Psi\,\mathrm{vec}(\Xi_k) = \Psi^{k+1}\mathrm{vec}(\Xi_0).$$

It is easy to see that $\lim_{k\to\infty} \Xi_k = 0$ is equivalent to the existence of A_K, B_K, C_K, D_K such that $\rho(\Psi) < 1$.

(i)⇔(iv): First, suppose that (iv) is true. Introduce a Lyapunov function $V(\Xi_k) = \mathrm{tr}(\Xi_k P)$, then

$$\begin{aligned} V(\Xi_{k+1}) &= \mathrm{tr}\left\{\left((\tilde{A} + \tilde{B}\Pi\tilde{C})\Xi_k(\tilde{A} + \tilde{B}\Pi\tilde{C})^T + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\tilde{B}_j\tilde{C}_j\Xi_k\tilde{C}_i^T\tilde{B}_i^T\right)P\right\} \\ &= \mathrm{tr}\left\{\Xi_k\left((\tilde{A} + \tilde{B}\Pi\tilde{C})^T P(\tilde{A} + \tilde{B}\Pi\tilde{C}) + \tilde{C}^T J\tilde{C}\right)\right\} \\ &< \mathrm{tr}(\Xi_k P) = V(\Xi_k). \end{aligned}$$

It follows from Lyapunov theory that $\lim_{k\to\infty} \Xi_k = 0$. On the other hand, assume that the closed-loop system is mean square stable. Then, $\rho(\Psi) < 1$ based on (iii), which is equivalent to $\rho(\Psi^T) < 1$. Therefore, the sequence $\{\hat{\Xi}_k\}_{k\geq 0}$ computed by

$$\hat{\Xi}_{k+1} = (\tilde{A} + \tilde{B}\Pi\tilde{C})^T\hat{\Xi}_k(\tilde{A} + \tilde{B}\Pi\tilde{C}) + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\tilde{C}_i^T\tilde{B}_i^T\hat{\Xi}_k\tilde{B}_j\tilde{C}_j \tag{6.13}$$

is convergent to zero as k approaches ∞ for any $\hat{\Xi}_0 \geq 0$. Let $\hat{\Xi}_0 > 0$ and $P_k = \sum_{s=0}^{k}\hat{\Xi}_s$, then we have

$$\begin{aligned} P_{k+1} &= \hat{\Xi}_0 + (\tilde{A} + \tilde{B}\Pi\tilde{C})^T P_k(\tilde{A} + \tilde{B}\Pi\tilde{C}) + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\tilde{C}_i^T\tilde{B}_i^T P_k\tilde{B}_j\tilde{C}_j \\ &> (\tilde{A} + \tilde{B}\Pi\tilde{C})^T P_k(\tilde{A} + \tilde{B}\Pi\tilde{C}) + \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\tilde{C}_i^T\tilde{B}_i^T P_k\tilde{B}_j\tilde{C}_j. \end{aligned} \tag{6.14}$$

It follows from the convergence of $\{\hat{\Xi}_k\}_{k\geq 0}$ and (6.14) that $P = \lim_{k\to\infty} P_k$ exists and satisfies (6.11) in (iv). It is easy to see that $P > 0$ since $\hat{\Xi}_0 > 0$. In addition, according to Lemma A.5, we have $J = \Sigma \odot (B^T P B) > 0$ since $\Sigma \geq 0$, $B^T P B > 0$, and Σ has no diagonal entry that is equal to zero.

(iv)$\Leftrightarrow$(v): Assume that (v) holds. By selecting a Lyapunov function $V(\hat{\Xi}_k) = \mathrm{tr}(\hat{\Xi}_k W)$, we can easily prove that (v) implies the convergence of $\{\hat{\Xi}_k\}_{k\geq 0}$ computed by (6.13), which further implies (iv). The converse follows by setting $W = \lim_{k\to\infty} \sum_{s=0}^{k} \Xi_s$ and $\Xi_0 > 0$. □

Remark 6.2 It is easy to show that the mean square stabilizability is invariant under similarity transformations on the triplet (A, B, C).

6.2 State Feedback Case

Consider the state feedback case with $C = I$. In view of Remark 6.2, we take (A, B) to be of the form:

$$A = \begin{bmatrix} A_\mathrm{s} & 0 \\ 0 & A_\mathrm{u} \end{bmatrix}, \quad B = \begin{bmatrix} B_\mathrm{s} \\ B_\mathrm{u} \end{bmatrix}, \tag{6.15}$$

where A_s is stable, all the eigenvalues of A_u are either on or outside the unit circle, and $(A_\mathrm{u}, B_\mathrm{u})$ is controllable. We can derive the next lemma.

Lemma 6.2 *Consider the NCS in Fig. 6.1 with state feedback and predefined fading channel(s) modeled by (6.3)–(6.4). The following statements are equivalent.*

(i) *The plant (6.1) or the pair (A, B) can be stabilized in the mean square sense via static state feedback.*
(ii) *The pair (A, B) can be stabilized in the mean square sense via dynamic state feedback.*
(iii) *There exists $P > 0$ such that*

$$P > A^T P A - A^T P B \Pi (J + \Pi B^T P B \Pi)^{-1} \Pi B^T P A, \tag{6.16}$$

where J is as defined in (6.11) with $\tilde{B} = B$.
(iv) *For any $R > 0$, there exists $P > 0$ such that*

$$P = A^T P A - A^T P B \Pi (J + \Pi B^T P B \Pi)^{-1} \Pi B^T P A + R. \tag{6.17}$$

(v) *The pair $(A_\mathrm{u}, B_\mathrm{u})$ can be stabilized in the mean square sense via static or dynamic state feedback.*

Furthermore, if any of the conditions (i)–(v) *is true, then a state feedback gain ensuring the mean square stability of the closed-loop system is given by*

$$K = -(J + \Pi B^T P B \Pi)^{-1} \Pi B^T P A, \tag{6.18}$$

where $P > 0$ is any solution to (6.16) or (6.17).

Proof (i)⇔(ii): For static state feedback, $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ in (6.8) are reduced to $\tilde{A} = A$, $\tilde{B} = B$, $\tilde{C} = K$, respectively. Obviously, the stabilizability via static state feedback always implies the stabilizability via dynamic state feedback. In what follows, we will show that the converse is also true. Suppose the closed-loop system is mean square stable under a dynamic state feedback controller, i.e., there exist a set of A_K, B_K, C_K, D_K and $W > 0$ such that (6.12) in Lemma 6.1 is true for $C = I$. Partition W in accordance with $\tilde{A}$ in (6.8) as

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix}.$$

In this case, the (1×1) block of the inequality (6.12) yields

$$\begin{aligned} W_{11} >& (A + B\Pi D_K)W_{11}(A + B\Pi D_K)^T + (A + B\Pi D_K)W_{12}C_K^T\Pi B^T \\ &+ B\left(\Sigma \odot (D_K W_{11} D_K^T + C_K W_{12}^T D_K^T + D_K W_{12} C_K^T + C_K W_{22} C_K^T)\right) B^T \\ &+ B\Pi C_K W_{22} C_K^T \Pi B^T + B\Pi C_K W_{12}^T (A + B\Pi D_K)^T. \end{aligned} \tag{6.19}$$

By setting $K = D_K + C_K W_{12}^T W_{11}^{-1}$, it follows from (6.19), the property of Hadamard product in Lemma A.5, and $W_{22} - W_{12}^T W_{11}^{-1} W_{12} > 0$ that

$$W_{11} > (A + B\Pi K)W_{11}(A + B\Pi K)^T + B\left(\Sigma \odot (K W_{11} K^T)\right) B^T. \tag{6.20}$$

From Lemma 6.1 (v), the inequality (6.20) is equivalent to the existence of a state feedback gain K such that the closed-loop system is mean square stable.

(i)⇐(iii) By substituting $\tilde{A} = A$, $\tilde{B} = B$, $\tilde{C} = K$ into (6.11), we can obtain that

$$P > (A + B\Pi K)^T P (A + B\Pi K) + K^T J K. \tag{6.21}$$

The inequality (6.16) implies (6.21) by setting K as in (6.18).

(i)⇒(iii): By taking derivative with respect to u, we can conclude that $u = Kx$ with K defined in (6.18) minimizes the function

$$f(u) = -x^T P x + (Ax + B\Pi u)^T P (Ax + B\Pi u) + u^T J u.$$

(iii)⇔(iv): By choosing the same P, (iv) obviously implies (iii). Next, suppose that (iii) holds. Define the operators

$$\begin{aligned} \mathscr{L}_1(Z) &= A^T Z A - A^T Z B \Pi (\Sigma \odot (B^T Z B) + \Pi B^T Z B \Pi)^{-1} \Pi B^T Z A, \\ \mathscr{L}_2(Z, K) &= (A + B\Pi K)^T Z (A + B\Pi K) + K^T (\Sigma \odot (B^T Z B)) K. \end{aligned}$$

It follows that $\mathscr{L}_2(Z, K)$ is affine in Z, and for any $Z \geq 0$ and K,

$$0 \leq \mathscr{L}_1(Z) = \mathscr{L}_2(Z, K^*(Z)) \leq \mathscr{L}_2(Z, K)$$

with $K^*(Z) = -(\Sigma \odot (B^T Z B) + \Pi B^T Z B \Pi)^{-1} \Pi B^T Z A$. For any $Z_2 \geq Z_1 \geq 0$, we have

$$\mathscr{L}_1(Z_1) = \mathscr{L}_2(Z_1, K^*(Z_1)) \leq \mathscr{L}_2(Z_1, K^*(Z_2)) \leq \mathscr{L}_2(Z_2, K^*(Z_2)) = \mathscr{L}_1(Z_2).$$

Let the sequence $\{Z_k\}_{k\geq 0}$ be computed by $Z_{k+1} = \mathscr{L}_1(Z_k) + R$ with any initial condition $Z_0 \geq 0$. The condition (iii) can be written into $P > \mathscr{L}_1(P)$ with $P > 0$. In this case, we can always choose $c_1 \in [0, 1)$ and $c_2 \in (0, \infty)$ such that

$$\mathscr{L}_1(P) \leq c_1 P,\ Z_0 \leq c_2 P,\ R \leq c_2 P.$$

Since $\mathscr{L}_1(c_2 P) = c_2 \mathscr{L}_1(P)$, we have

$$\begin{aligned} Z_1 &= \mathscr{L}_1(Z_0) + R \leq c_1 c_2 P + c_2 P, \\ Z_2 &= \mathscr{L}_1(Z_1) + R \leq c_1^2 c_2 P + c_1 c_2 P + c_2 P. \end{aligned}$$

By mathematical induction, we have $Z_k \leq \sum_{s=0}^{k} c_1^s c_2 P \leq \frac{c_2}{1-c_1} P$, which implies that the sequence $\{Z_k\}_{k\geq 0}$ is bounded. Let $Z_0 = 0$, then it follows from the monotonicity and boundedness of $\{Z_k\}_{k\geq 0}$ that $Z = \lim_{k\to\infty} Z_k$ exists and satisfies $Z = \mathscr{L}_1(Z) + R$. Note that $Z > 0$ since $R > 0$ and $\mathscr{L}_1(Z) \geq 0$, which completes the proof.

(i)$\Leftarrow$(v): In view of the equivalence between (i) and (ii), we limit our attention to the static state feedback case in the proof. Let (A, B) be of the form (6.15). Since A_s in (6.15) is stable, there exists $P_1 > 0$ such that $P_1 - A_s^T P_1 A_s > 0$. Also, since the pair (A_u, B_u) is mean square stabilizable, based on Lemma 6.1 (iv), there exist $P_2 > 0$ and K_u such that

$$P_2 > (A_u + B_u \Pi K_u)^T P_2 (A_u + B_u \Pi K_u) + K_u^T J_u K_u,$$

where $J_u = \Sigma \odot (B_u^T P_2 B_u)$. It follows that, for some $\beta > 0$, the inequality

$$P_2 > \beta I + (A_u + B_u \Pi K_u)^T P_2 (A_u + B_u \Pi K_u) + K_u^T J_u K_u$$

is true. Let $J_s = \Sigma \odot (B_s^T P_1 B_s)$. By choosing $P = \mathrm{diag}\{P_1, \gamma P_2\}$, $K = [0\ \ K_u]$ with a sufficiently large $\gamma > 0$ such that

$$\begin{aligned} \gamma\beta I > {} & K_u^T \Pi B_s^T P_1 B_s \Pi K_u \\ & + K_u^T J_s K_u + K_u^T \Pi B_s^T P_1 A_s (P_1 - A_s^T P_1 A_s)^{-1} A_s^T P_1 B_s \Pi K_u, \end{aligned}$$

and in view of $J = J_s + J_u$, we have (6.21) holds, i.e., the pair (A, B) is mean square stabilizable.

(i)$\Rightarrow$(v): According to the condition (iii), the mean square stabilizability of (A, B) in the form of (6.15) ensures that there exists

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix} > 0$$

such that (6.16) is true. After applying the linear coordinate transformation matrix

$$\begin{bmatrix} I & -P_{11}^{-1}P_{12} \\ 0 & I \end{bmatrix},$$

the (2×2) block of the inequality (6.16) implies that

$$P_2 > A_{\mathrm{u}}^T P_2 A_{\mathrm{u}} - A_{\mathrm{u}}^T P_2 B_{\mathrm{u}} \Pi (J_{\mathrm{u}} + \Pi B_{\mathrm{u}}^T P_2 B_{\mathrm{u}} \Pi)^{-1} \Pi B_{\mathrm{u}}^T P_2 A_{\mathrm{u}},$$

where $P_2 = P_{22} - P_{12}^T P_{11}^{-1} P_{12} > 0$ and $J_{\mathrm{u}} = \Sigma \odot (B_{\mathrm{u}}^T P_2 B_{\mathrm{u}})$. Therefore, the pair $(A_{\mathrm{u}}, B_{\mathrm{u}})$ is mean square stabilizable. □

Remark 6.3 One solution $P > 0$ to (6.17) can be obtained by solving the following optimization:

$$P = \arg\max_{\hat{P}>0} \mathrm{tr}(\hat{P}) \tag{6.22}$$

subject to the linear matrix inequality (LMI) constraint

$$\begin{bmatrix} \hat{P} - A^T \hat{P} A - R & A^T \hat{P} B \Pi \\ \Pi B^T \hat{P} A & -\Sigma \odot (B^T \hat{P} B) - \Pi B^T \hat{P} B \Pi \end{bmatrix} \leq 0.$$

However, the optimization in (6.22), despite of the ease of computation, offers no explicit condition on μ_i and σ_{ij} even when a solution $P > 0$ exists. Therefore, we turn to investigate the requirement on the network over which an unstable plant can be stabilized.

In view of Lemma 6.2, we limit our attention to the static state feedback case in the rest part of this section and assume that all the eigenvalues of A are either on or outside the unit circle without loss of generality. Based on Remark 6.2, the pair (A, B) is further assumed to have the following Wonham decomposition [4]:

$$A = \begin{bmatrix} A_1 & \star & \cdots & \star \\ 0 & A_2 & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{bmatrix}, \quad \text{and} \quad B = \begin{bmatrix} b_1 & \star & \cdots & \star \\ 0 & b_2 & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_m \end{bmatrix}, \tag{6.23}$$

where $\star$ represents terms that will not be used in the derivation, $A_i \in \mathbb{R}^{n_i \times n_i}$, $b_i \in \mathbb{R}^{n_i \times 1}$, $\sum_{i=1}^m n_i = n$, and each pair (A_i, b_i) is controllable. The next theorem which is parallel to Theorem 12.1 in the previous chapter provides necessary and sufficient conditions on the network for stabilizability via state feedback.

Theorem 6.1 *The NCS in Fig. 6.1 is mean square stabilizable via state feedback if*

$$1 + \frac{\mu_i^2}{[\Sigma_1]_{ii}} > \mathscr{M}(A_i)^2, \quad \forall i = 1, 2, \ldots, m, \tag{6.24}$$

and only if

$$\prod_{i=1}^{m}\left(1+\frac{\mu_i^2}{[\Sigma_2]_{ii}}\right) > \mathscr{M}(A)^2, \tag{6.25}$$

where Σ_1 and Σ_2 are arbitrary positive-semidefinite diagonal matrices satisfying $0 \le \Sigma_2 \le \Sigma \le \Sigma_1$.

Proof We will first prove the necessity of the condition (6.25). Suppose that the closed-loop system is mean square stable under a static state feedback controller. Based on the property of Hadamard product in Lemma A.5, we have $J \ge \hat{J}$ with $\hat{J} = \Sigma_2 \odot (B^T P B)$. The condition (6.25) obviously holds when any $[\Sigma_2]_{ii} = 0$, otherwise it follows from (6.16) that

$$\begin{aligned}
\det(P) &> \det(A^T)\det\left(I - PB\Pi(J+\Pi B^T PB\Pi)^{-1}\Pi B^T\right)\det(P)\det(A)\\
&= \det(A)^2\det(P)\det\left(I-(J+\Pi B^T PB\Pi)^{-1}\Pi B^T PB\Pi\right)\\
&= \det(A)^2\det(P)\left(\det(I+\Pi B^T PB\Pi J^{-1})\right)^{-1}\\
&= \det(A)^2\det(P)\left(\det(I+P^{1/2}B\Pi J^{-1}\Pi B^T P^{1/2})\right)^{-1}\\
&\ge \det(A)^2\det(P)\left(\det(I+P^{1/2}B\Pi \hat{J}^{-1}\Pi B P^{1/2})\right)^{-1}\\
&= \det(A)^2\det(P)\left(\det(I+\hat{J}^{-1/2}\Pi B^T PB\Pi\hat{J}^{-1/2})\right)^{-1}\\
&\ge \mathscr{M}(A)^2\det(P)\prod_{i=1}^{m}\left(\frac{[\Sigma_2]_{ii}}{[\Sigma_2]_{ii}+\mu_i^2}\right).
\end{aligned} \tag{6.26}$$

The inequality (6.26) follows from Hadamard's inequality; see Lemma A.2. The proof of necessity is completed since (6.26) implies (6.25).

Next, the sufficiency of the condition (6.24) will be shown. According to Corollary 8.4 in [1] and Lemma 6.1, if (6.24) is true, then there exist $P_i > 0$ and $K_i \in \mathbb{R}^{1\times n_i}$ such that

$$P_i > (A_i + b_i\mu_i K_i)^T P_i (A_i + b_i\mu_i K_i) + [\Sigma_1]_{ii} K_i^T b_i^T P_i b_i K_i.$$

It is direct to show that the closed-loop system is mean square stable for $m = 1$. Next, we consider the case $m = 2$. Adopting the similar technique in the proof of the equivalence between (i) and (v) in Lemma 6.2, we can show that there exist $P = \mathrm{diag}\{P_1, \gamma_1 P_2\}$ and $K = \mathrm{diag}\{K_1, K_2\}$ with sufficiently large γ_1 such that

$$P > (A + B\Pi K)^T P(A + B\Pi K) + K^T(\Sigma_1 \odot (B^T PB))K, \tag{6.27}$$

where the pair (A, B) has the form given in (6.23). By induction, for the case $m > 2$, there exist

$$P = \text{diag}\{P_1, \gamma_1 P_2, \ldots, \gamma_{m-1} P_m\}$$

and

$$K = \text{diag}\{K_1, K_2, \ldots, K_m\}$$

with sufficiently large $\gamma_1, \gamma_2, \ldots, \gamma_{m-1}$ such that (6.27) holds. It follows from $\Sigma_1 \odot (B^T P B) \geq J$ that (6.11) holds with $\tilde{A} = A, \tilde{B} = B, \tilde{C} = K$. Therefore, the closed-loop NCS is mean square stable under a static state feedback controller. □

Remark 6.4 Note that for the single-input case, the bounds presented in (6.24) and (6.25) are consistent by taking $\Sigma_1 = \Sigma_2 = \Sigma$, and Theorem 6.1 is reduced to Corollary 8.4 of [1]. For multi-input case, the tightness of (6.24) and (6.25) is affected by the selection of Σ_1, Σ_2 and thus depends on the structure of Σ. Moreover, (6.24) is also related to the particular Wonham decomposition in (6.23) which is generally not unique.

Next, we will further refine Theorem 6.1 under the following two typical transmission strategies for the network (6.3)–(6.4).

- **Parallel transmission strategy (PTS)**: Each element of v_k is sent across an individual fading channel, and $\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{m,k}$ are uncorrelated with each other.
- **Serial transmission strategy (STS)**: All elements of v_k are transmitted over the same fading channel one after another, and $\xi_{1,k} = \xi_{2,k} = \cdots = \xi_{m,k}$.

Remark 6.5 In some industrial and military applications where a group of ground, aerial, and/or underwater vehicles/robots are geographically separated from each other and from the remote controller, the control signal to each vehicle/robot is sent across an individual channel, which motivates the consideration of PTS. As mentioned in Remark 6.1, $\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{m,k}$ can be made to be uncorrelated with each other in PTS by adopting an orthogonal access scheme. For STS, the condition $\xi_{1,k} = \xi_{2,k} = \cdots = \xi_{m,k}$ is valid in applications when the fading is coherent over the sampling interval. Note that the network model with parallel channels or a single channel has been widely used in the literature on networked estimation and networked control; see, e.g., [5–9].

Under PTS or STS, define the overall mean square capacity of the network as

$$\mathscr{C}_{\text{MS}} \triangleq \sum_{i=1}^{m} \frac{1}{2} \ln\left(1 + \frac{\mu_i^2}{\sigma_i^2}\right), \tag{6.28}$$

which is reduced to the mean square capacity presented in [1] when $m = 1$. Further denote

$$\mathscr{C}_{\text{MS}i} \triangleq \frac{1}{2} \ln\left(1 + \frac{\mu_i^2}{\sigma_i^2}\right), \; g_i \triangleq 1 + \frac{\mu_i^2}{\sigma_i^2}$$

and $g \triangleq \prod_{i=1}^{m} g_i$, then it follows that

$$\mathscr{C}_{\mathrm{MS}i} = \frac{1}{2} \ln g_i, \ \mathscr{C}_{\mathrm{MS}} = \frac{1}{2} \ln g,$$

and the larger the g_i, the larger the $\mathscr{C}_{\mathrm{MS}i}$. Note that under STS, $\mathscr{C}_{\mathrm{MS}} = m\mathscr{C}_{\mathrm{MS}1}$. Next, we will present a tight lower bound on the overall mean square capacity of the network for mean square stabilizability under PTS and STS, respectively.

6.2.1 Parallel Transmission Strategy

Under PTS, the overall mean square capacity is considered as the network resource and assumed to satisfy the following assumption.

Assumption 6.1 The overall mean square capacity of the network is fixed and can be allocated among the parallel channels.

Remark 6.6 It is worth mentioning that the resource (e.g., power, code, time and frequency) allocation technique has been employed extensively for studying capacity maximization in communications [10, 11], and recently it has attracted the interests from the control community as well [5, 12].

A relationship between the overall mean square capacity and the Mahler measure of the plant for ensuring the mean square stabilizability via state feedback under PTS is characterized by the next theorem.

Theorem 6.2 *Under PTS and Assumption 6.1, the NCS in Fig. 6.1 is mean square stabilizable via state feedback if and only if*

$$\mathscr{C}_{\mathrm{MS}} > \ln \mathscr{M}(A). \tag{6.29}$$

Proof Note that under PTS, $\sigma_{ij} = 0$ for all $i \neq j$, i.e., Σ is diagonal. The necessity of (6.29) for mean square stabilizability via state feedback follows from (6.25) in Theorem 6.1 by taking $\Sigma_2 = \Sigma$. On the other hand, under Assumption 6.1 on capacity allocation and $\mathscr{C}_{\mathrm{MS}} > \ln \mathscr{M}(A)$, i.e.,

$$g > \mathscr{M}(A)^2 = \prod_{i=1}^{m} \mathscr{M}(A_i)^2,$$

we can always choose $g_i > \mathscr{M}(A_i)^2$ such that the condition (6.24) in Theorem 6.1 is true by letting $\Sigma_1 = \Sigma$. Therefore, there exist a stabilizing state feedback gain $K = \mathrm{diag}\{K_1, K_2, \ldots, K_m\}$ as in Theorem 6.1 and an allocation $\{\mathscr{C}_{\mathrm{MS}1}, \mathscr{C}_{\mathrm{MS}2}, \ldots, \mathscr{C}_{\mathrm{MS}m}\}$ satisfying $\mathscr{C}_{\mathrm{MS}} = \sum_{i=1}^{m} \mathscr{C}_{\mathrm{MS}i}$ such that the closed-loop system is mean square stable. □

Remark 6.7 Similar to the well-known data rate theorem [13] which quantifies the minimal data rate necessary for stabilizing an unstable system, Theorem 6.2 characterizes the minimal overall mean square capacity of the network in order to stabilize an unstable system in the mean square sense. The result also pinpoints a relationship between the minimal overall mean square capacity and the unstable eigenvalues of the system matrix A.

6.2.2 Serial Transmission Strategy

Based on Schur's complement in Lemma A.1, the inequality (6.11) in Lemma 6.1 (iv) under STS with $\tilde{A} = A, \tilde{B} = B, \tilde{C} = K$ is equivalent to

$$\begin{bmatrix} -S & (AS+BY)^T & (\sqrt{\frac{1}{g_1-1}}BY)^T \\ AS+BY & -S & 0 \\ \sqrt{\frac{1}{g_1-1}}BY & 0 & -S \end{bmatrix} < 0 \tag{6.30}$$

with $S = P^{-1}/\mu,\ Y = KP^{-1}$. Then, we can deduce the following result.

Proposition 6.2.1 *Under STS, the NCS in Fig. 6.1 is mean square stabilizable via state feedback if and only if*

$$\mathscr{C}_{\mathrm{MS}} > \frac{m}{2} \ln g_{1c}, \tag{6.31}$$

where

$$g_{1c} = \begin{cases} \mathscr{M}(A)^2, & \text{if } m = 1; \\ \rho(A)^2, & \text{if } m = n; \\ \inf\limits_{S>0,Y} g_1, \text{ subject to (6.30)}, & \text{otherwise}. \end{cases} \tag{6.32}$$

Furthermore, it holds that

$$\frac{m}{2} \ln g_{1c} \geq \ln \mathscr{M}(A). \tag{6.33}$$

Proof First, we will show the expression for g_{1c} in (6.32). When $m = 1$, the STS is the same as the PTS, thus the critical value g_{1c} for mean square stabilizability follows directly from Theorem 6.2. Note that (6.16) in Lemma 6.2 (iii) under STS becomes

$$P > A^T PA - \frac{\mu_1^2}{\sigma_1^2 + \mu_1^2} A^T PB(B^T PB)^{-1} B^T PA. \tag{6.34}$$

When $m = n$, the inequality (6.34) is further reduced to

$$P > \left(g_1^{-\frac{1}{2}} A\right)^T P \left(g_1^{-\frac{1}{2}} A\right). \tag{6.35}$$

In this case, the mean square stabilizability is equivalent to $\rho(g_1^{-\frac{1}{2}}A) < 1$, and thus the critical value g_{1c} is given by $\rho(A)^2$. For general cases, we can use (6.30) as the necessary and sufficient condition for mean square stabilizability. Observe that if (6.30) is true for some $g_1 = g_{1a} > 1$, then it holds for all $g_1 \geq g_{1a}$. Therefore, the critical value g_{1c} can be obtained via the minimization of g_1 over (6.30).

Next, we will prove (6.33) for an arbitrary m. Since all the eigenvalues of A are assumed to be on or outside the unit circle, it follows from (6.34) that

$$\begin{aligned}\det(P) &> \det(A)\det\left(I - \frac{\mu_1^2}{\sigma_1^2+\mu_1^2}PB(B^TPB)^{-1}B^T\right)\det(P)\det(A^T)\\ &= \det(A)^2\det(P)\det\left(I_m - \frac{\mu_1^2}{\sigma_1^2+\mu_1^2}I_m\right)\\ &= \mathscr{M}(A)^2\det(P)g_1^{-m},\end{aligned}$$

which gives (6.33). In particular, the equality in (6.33) holds for $m = 1$, while for $m = n$, we have that

$$\frac{m}{2}\ln g_{1c} = \ln\rho(A)^n \geq \ln\mathscr{M}(A). \qquad \square$$

Remark 6.8 Note that Proposition 6.2.1 is consistent with Lemma 5.4 of [14] where the packet-loss issue is considered. Except for some special cases, the critical value g_{1c} is, in general, not connected with the system matrix A explicitly. From Theorem 6.2 and Proposition 6.2.1, we can conclude that if the overall mean square capacity is allocatable among fading channels, i.e., additional flexibility of designing the communication component is added, then the optimization in (6.32) is avoided, and the requirement on overall mean square capacity may be reduced when $m > 1$.

6.3 Output Feedback Case

In the case of output feedback, we will base our analysis on both the state-space and stable coprime factorization approaches. The frequency variable z will be omitted whenever no confusion is caused. In view of the framework of fading channels in [1], the network (6.3)–(6.4) under PTS is equivalent to the structure shown in Fig. 6.2, where $z_k = v_k$, $w_k = \tilde{\Delta}_k z_k$, and the uncertainty block $\tilde{\Delta}_k$ is defined as $\tilde{\Delta}_k \triangleq \Lambda^{-1}(\xi_k - \Pi)$.

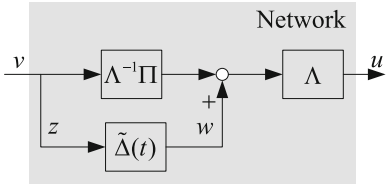

Fig. 6.2 Equivalent description of the network (6.3)–(6.4) under PTS

It is easy to verify that the mean square stability of the NCS in Fig. 6.1 with the network (6.3)–(6.4) always implies the internal stability of the corresponding closed-loop system with $w_k = 0$. Denote the set of all proper controllers achieving the above internal stability by $\mathscr{K}$. The transfer function of the closed-loop from w_k to z_k, without considering the uncertainty block $\tilde{\Delta}_k$, is given by

$$T(z) = (I - K(z)G(z)\Pi)^{-1}K(z)G(z)\Lambda. \tag{6.36}$$

The mean square norm of $G(z)$ with dimension $\ell \times m$, if exists, is defined as

$$\|G(z)\|_{\mathrm{MS}} \triangleq \sqrt{\max_{i=1,2,\ldots,\ell} \frac{1}{2\pi}\int_{-\pi}^{\pi} [G(e^{j\omega})G^H(e^{j\omega})]_{ii}d\omega}. \tag{6.37}$$

As proved in [1], under PTS with given Π and Λ in (6.6), the mean square stabilization of the NCS in Fig. 6.1 is equivalent to that

$$\inf_{K(z)\in\mathscr{K},\Theta>0,\mathrm{diag}} \|\Theta^{-1}T(z)\Theta\|_{\mathrm{MS}}^2 < 1, \tag{6.38}$$

where $T(z)$ is given in (6.36) and $\Theta \in \mathbb{R}^{m\times m}$ is a positive-definite diagonal matrix. Similarly to other robust control problems, the search for the optimal controller $K(z)$ on the left-hand side of (6.38) is generally nonconvex in Θ, and the minimal overall capacity for stabilizability has no explicit solution in general.

Let a pair of right and left coprime factorizations of $G(z)$ be given in the familiar way, namely

$$G = NM^{-1} = \tilde{M}^{-1}\tilde{N}, \tag{6.39}$$

where $N, M, \tilde{N}, \tilde{M} \in \mathscr{RH}_\infty$, and satisfy the double Bezout identity

$$\begin{bmatrix} \tilde{Y} & -\tilde{X} \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} M & X \\ N & Y \end{bmatrix} = I$$

for some $X, Y, \tilde{X}, \tilde{Y} \in \mathscr{RH}_\infty$. Then, the set of all proper controllers achieving the internal stability of the closed-loop system in Fig. 6.1 with the network in Fig. 6.2 can be parameterized as [15, p. 228]

$$\begin{aligned}\mathscr{K} = \{K(z) : K(z) &= \Pi^{-1}(X + MQ)(Y + NQ)^{-1} \\ &= \Pi^{-1}(\tilde{Y} + Q\tilde{N})^{-1}(\tilde{X} + Q\tilde{M}), Q \in \mathscr{RH}_\infty\Big\}.\end{aligned} \tag{6.40}$$

In this situation, the transfer function $T(z)$ in (6.36) becomes

$$T(z) = \Pi^{-1}(X\tilde{N} + MQ\tilde{N})\Lambda, \tag{6.41}$$

where Π and Σ are given in (6.6).

6.3.1 SISO Plants

For SISO plants, we have $m = \ell = 1$, and $\xi_k = \xi_{1,k}$ is a scalar-valued white noise signal. In this case, the STS and the PTS are the same.

Consider an SISO plant $G(z)$ with n_ϕ anti-stable poles $\phi_1, \phi_2, \ldots, \phi_{n_\phi}, n_\zeta$ distinct[4] NMP zeros $\zeta_1, \zeta_2, \ldots, \zeta_{n_\zeta}$ and relative degree $r \geq 1$. It is easy to derive that

$$\mathscr{M}(G) = \begin{cases} \prod_{i=1}^{n_\phi} |\phi_i|, & \text{if } n_\phi \geq 1; \\ 1, & \text{if } n_\phi = 0, \end{cases} \tag{6.42}$$

which is equal to $\mathscr{M}(A)$ with $\left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right]$ being any detectable and stabilizable realization of $G(z)$. The Blaschke product with respect to the anti-stable poles of $G(z)$ is given by [17]

$$B_\phi(z) \triangleq \prod_{i=1}^{n_\phi} \frac{z - \phi_i}{1 - z\bar{\phi}_i}. \tag{6.43}$$

Since $B_\phi(z)$ is inner [18, p. 66], we can derive that

$$B_\phi(z^{-1}) = \sum_{k=0}^{\infty} \beta_l z^{-k}$$

with $\beta_k \triangleq \frac{1}{k!} \frac{d^k}{dz^k} B_\phi(z)|_{z=0}$. Further denote [19]

$$\eta(G) \triangleq \sum_{l=1}^{n_\zeta} \sum_{k=1}^{n_\zeta} \frac{\bar{\gamma}_k \gamma_l}{\bar{\zeta}_k \zeta_l - 1}, \tag{6.44}$$

$$\varphi(G) \triangleq \sum_{l=1}^{r-1} \left(|\beta_l|^2 + |\psi_l|^2 \right), \tag{6.45}$$

where $\psi_l \triangleq \sum_{k=1}^{n_\zeta} \gamma_k \zeta_k^{l-1}$ and

$$\gamma_k \triangleq (1 - |\zeta_k|^2) \left(B_\phi(\zeta_k^{-1}) - \sum_{i=0}^{r-1} \beta_i \zeta_k^{-i} \right) \prod_{i=1, i \neq k}^{n_\zeta} \frac{1 - \bar{\zeta}_i \zeta_k}{\zeta_k - \zeta_i}.$$

[4] The assumption on distinct NMP zeros simplifies the subsequent analysis and may be relaxed at the expense of more complex expressions [16].

Proposition 6.3.1 *Assume that $G(z)$ is SISO and has no or distinct (if any) NMP zeros. The NCS in Fig. 6.1 is mean square stabilizable via dynamic output feedback if and only if*

$$\mathscr{C}_{\mathrm{MS}} > \frac{1}{2}\ln\left\{\mathscr{M}(G)^2 + \eta(G) + \varphi(G)\right\}, \tag{6.46}$$

where $\mathscr{M}(G)$, $\eta(G)$, $\varphi(G)$ are defined in (6.42), (6.44) and (6.45), respectively.

Proof For $m = 1$, we have that

$$\|\Theta^{-1}T(z)\Theta\|_{\mathrm{MS}}^2 = \|T(z)\|_2^2$$

with

$$T(z) = (X\tilde{N} + MQ\tilde{N})(g-1)^{-\frac{1}{2}}.$$

It follows that (6.38) is equivalent to

$$g > \inf_{Q\in\mathscr{RH}_\infty} \|X\tilde{N} + MQ\tilde{N}\|_2^2 + 1. \tag{6.47}$$

The evaluation of the right-hand side of (6.47) follows from [17, 19]. □

6.3.2 Triangularly Decoupled Plants

Motivated by the observation that Wonham decomposition plays an important role in establishing the sufficiency part of Theorem 6.1, we adopt the next definition of triangular decoupling.

Definition 6.2 ([20]) $G(z)$ is said to be triangularly decoupled by $K(z) \in \mathscr{K}$, if $T(z)$ in (6.41) is either lower or upper triangular.

For $G = NM^{-1}$ with $M, N \in \mathscr{RH}_\infty$ and $m \le \ell$, there always exist suitable unimodular matrices U_M, U_N such that

$$L_M = MU_M, \quad \begin{bmatrix} L_N \\ 0 \end{bmatrix} = U_N N U_M, \quad U_N G = \begin{bmatrix} L_N L_M^{-1} \\ 0 \end{bmatrix}, \tag{6.48}$$

where L_M, L_N are respectively square and lower (or upper) triangular with compatible dimensions. We refer to $L_G \triangleq L_N L_M^{-1}$ as the left triangular structure of $G(z)$.

Remark 6.9 As shown in [20], (i) L_M and L_N are unique up to postmultiplication and/or premultiplication by any lower (or upper) triangular unimodular matrix, and are independent of the particular factorization of $G(z)$ in (6.39); (ii) a necessary and sufficient condition for the solvability of triangular decoupling is the corprimeness between each diagonal entry of L_M and the corresponding diagonal entry of L_N.

We sum up the assumptions as follows.

Assumption 6.2 (i) The plant $G(z)$ can be triangularly decoupled and satisfies $m \leq \ell$.

(ii) The ith diagonal entry of the left triangular structure, $[L_G]_{ii}$, has no or distinct (if any) NMP zeros for every $i = 1, 2, \ldots, m$.

Note that $[L_G]_{ii}$ has relative degree $r_i \geq 1$, since $U_N(z)$ in (6.48) is unimodular and $G(z)$ is strictly proper. Suppose that $U_N(z)$ has a minimal realization

$$\left[\begin{array}{c|c} A_U & B_U \\ \hline C_U & D_U \end{array}\right],$$

then $\rho(A_U) < 1$ since $U_N(z)$ is unimodular. Under Assumption 6.2 (i),

$$\left[\begin{array}{cc|c} A & 0 & B \\ B_U C & A_U & 0 \\ \hline D_U C & C_U & 0 \end{array}\right]$$

is a stabilizable and detectable realization for $U_N(z)G(z)$. It follows that

$$\prod_{i=1}^{m} \mathscr{M}([L_G]_{ii}) = \mathscr{M}(U_N G) = \mathscr{M}\left(\left[\begin{array}{cc} A & 0 \\ B_U C & A_U \end{array}\right]\right) = \mathscr{M}(A) = \mathscr{M}(G). \tag{6.49}$$

Theorem 6.3 *Under PTS and Assumptions 6.1 and 6.2, the NCS in Fig. 6.1 is mean square stabilizable via dynamic output feedback if*

$$\mathscr{C}_{\mathrm{MS}} > \frac{1}{2} \sum_{i=1}^{m} \ln \left\{ \mathscr{M}([L_G]_{ii})^2 + \eta([L_G]_{ii}) + \varphi([L_G]_{ii}) \right\}, \tag{6.50}$$

and only if

$$\mathscr{C}_{\mathrm{MS}} > \ln \mathscr{M}(A). \tag{6.51}$$

Proof We will first show the necessity. Assume that there exists $K(z)$ such that the closed-loop system is mean square stable. Since any dynamic output feedback controller can always be constructed from a dynamic state feedback controller, the condition (6.51) follows directly from the equivalence between (i) and (ii) in Lemma 6.2 and Theorem 6.2.

Next, the sufficiency of (6.50) will be shown. Suppose that L_G is in a lower triangular form. Let a left stable coprime factorization of L_G be $L_{\tilde{M}}^{-1} L_{\tilde{N}}$. According to Assumption 6.2 (i) and Remark 6.9, $[L_M]_{ii}$ and $[L_N]_{ii}$ are corprime, and thus there always exist lower triangular $L_X, L_Y, L_{\tilde{X}}, L_{\tilde{Y}} \in \mathscr{RH}_\infty$ such that

$$\left[\begin{array}{cc} L_{\tilde{Y}} & -L_{\tilde{X}} \\ -L_{\tilde{N}} & L_{\tilde{M}} \end{array}\right] \left[\begin{array}{cc} L_M & L_X \\ L_N & L_Y \end{array}\right] = I.$$

It is easy to see that $[L_G]_{ii} = [L_N]_{ii}[L_M]_{ii}^{-1} = [L_{\tilde{M}}]_{ii}^{-1}[L_{\tilde{N}}]_{ii}$, and

$$\begin{bmatrix} [L_{\tilde{Y}}]_{ii} & -[L_{\tilde{X}}]_{ii} \\ -[L_{\tilde{N}}]_{ii} & [L_{\tilde{M}}]_{ii} \end{bmatrix} \begin{bmatrix} [L_M]_{ii} & [L_X]_{ii} \\ [L_N]_{ii} & [L_Y]_{ii} \end{bmatrix} = I.$$

If we construct right and left coprime factorizations of $U_N G$ as

$$U_N G = \begin{bmatrix} L_N \\ 0 \end{bmatrix} L_M^{-1} = \begin{bmatrix} L_{\tilde{M}} & 0 \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} L_{\tilde{N}} \\ 0 \end{bmatrix}, \tag{6.52}$$

then the set of all proper controllers achieving triangular decoupling can be parameterized as

$$\begin{aligned} \mathscr{K}_L \triangleq \{ & [L_K \ \star] U_N \in \mathscr{K} : L_K = \Pi^{-1}(L_X + L_M L_Q)(L_Y + L_N L_Q)^{-1} \\ & = \Pi^{-1}(L_{\tilde{Y}} + L_Q L_{\tilde{N}})^{-1}(L_{\tilde{X}} + L_Q L_{\tilde{M}}), \text{ triangular } L_Q \in \mathscr{RH}_\infty \}, \end{aligned} \tag{6.53}$$

where $\star$ denotes the part that will not be used in the proof. Adopt one of those triangular decoupling controllers $K(z) \in \mathscr{K}_L$, then $T(z)$ in (6.36) becomes

$$\begin{aligned} T(z) &= K(I - G\Pi K)^{-1} G\Lambda \\ &= [L_K \ \star] U_N (I - G\Pi [L_K \ \star] U_N)^{-1} G\Lambda \\ &= [L_K \ \star] (I - U_N G\Pi [L_K \ \star])^{-1} U_N G\Lambda \\ &= [\Pi^{-1}(L_X + L_M L_Q)(L_Y + L_N L_Q)^{-1} \ \star] \\ &\quad \times \left(I - \begin{bmatrix} L_{\tilde{M}} & 0 \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} L_{\tilde{N}} \\ 0 \end{bmatrix} \Pi [\Pi^{-1}(L_X + L_M L_Q)(L_Y + L_N L_Q)^{-1} \ \star] \right)^{-1} \\ &\quad \times \begin{bmatrix} L_{\tilde{M}} & 0 \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} L_{\tilde{N}} \\ 0 \end{bmatrix} \Lambda \\ &= \Pi^{-1}(L_X L_{\tilde{N}} + L_M L_Q L_{\tilde{N}})\Lambda. \end{aligned}$$

By choosing

$$\Theta_1 \triangleq \Pi\Theta = \operatorname{diag}\{\varepsilon^{m-1}, \varepsilon^{m-2}, \ldots, 1\}$$

with a small real number $\varepsilon > 0$, it follows that

$$\begin{aligned} \Theta^{-1} T(z) \Theta &= \Theta^{-1}\Pi^{-1}(L_X L_{\tilde{N}} + L_M L_Q L_{\tilde{N}})\Lambda\Theta \\ &= \Theta_1^{-1}(L_X L_{\tilde{N}} + L_M L_Q L_{\tilde{N}})\Lambda\Pi^{-1}\Theta_1 \end{aligned}$$

$$= \operatorname{diag}\left\{T_1(z)\sqrt{\frac{1}{g_1-1}}, T_2(z)\sqrt{\frac{1}{g_2-1}}, \ldots, T_m(z)\sqrt{\frac{1}{g_m-1}}\right\} + o_z(\varepsilon),$$

where $T_i(z) = [L_N]_{ii}[L_{\tilde{X}}]_{ii} + [L_M]_{ii}[L_Q]_{ii}[L_{\tilde{N}}]_{ii}$ with $[L_Q]_{ii} \in \mathscr{RH}_\infty$. According to the proof of Proposition 6.3.1, it follows that

$$\mathscr{C}_{\mathrm{MS}i} > \frac{1}{2}\ln\left\{\mathscr{M}([L_G]_{ii})^2 + \eta([L_G]_{ii}) + \varphi([L_G]_{ii})\right\}$$

is sufficient to ensure $\|T_i(z)\|_2^2/(g_i - 1) < 1$ under Assumption 6.2 (ii). Moreover, given the condition (6.50) and Assumption 6.1, we have that

$$\|\Theta^{-1}T(z)\Theta\|_{\mathrm{MS}}^2 < 1$$

by choosing a sufficiently small ε, i.e., the mean square stability of the closed-loop system is guaranteed. The result follows in a similar way by setting $\Theta_1 = \operatorname{diag}\{1, \varepsilon, \ldots, \varepsilon^{m-1}\}$ when L_G is in an upper triangular form. □

Remark 6.10 Note that the proof of the necessity does not rely on Assumptions 6.1 and 6.2, thus (6.51) actually provides a uniform lower bound on the overall capacity for stabilizability via output feedback under PTS over all proper stabilizing controllers. In view of (6.49), the gap between (6.50) and (6.51) shrinks to zero if $[L_G]_{ii}$ is minimum phase with relative degree 1 for all $i = 1, 2, \ldots, m$.

6.4 Extension and Application

6.4.1 Stabilization Over Output Fading Channels

Now, consider an NCS with an output network as depicted in Fig. 6.3, where

$$y = G(z)u, \quad u = K(z)v, \quad v_k = \xi_k y_k. \tag{6.54}$$

Motivated by applications where a sensor network is used to monitor a phenomenon or object of interest (e.g., a moving vehicle to be tracked) and the sensors are geographically distributed and are not collocated with the fusion center and/or the remote controller, we let the fading gain ξ_k in (6.54) and the overall mean square capacity of the network be given by

$$\xi_k = \operatorname{diag}\left\{\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{\ell,k}\right\}$$

and $\mathscr{C}_{\mathrm{MS}} \triangleq \sum_{i=1}^{\ell} \mathscr{C}_{\mathrm{MS}i}$, respectively. In this section, we assume that the controller only knows the first and second moments but not the exact value of ξ_k.

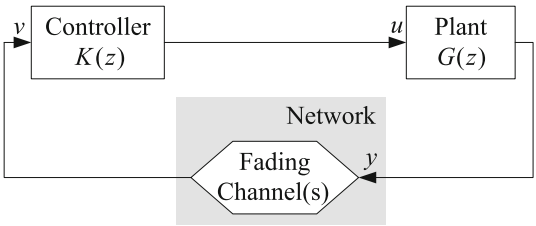

Fig. 6.3 Control over output fading channel(s)

Note that for $G = \tilde{M}^{-1}\tilde{N}$ with $\tilde{M}, \tilde{N} \in \mathscr{RH}_\infty$ and $\ell \leq m$, there exist suitable unimodular matrices $U_{\tilde{M}}, U_{\tilde{N}}$ such that

$$L_{\tilde{M}} = U_{\tilde{M}}\tilde{M}, \begin{bmatrix} L_{\tilde{N}} & 0 \end{bmatrix} = U_{\tilde{M}}\tilde{N}U_{\tilde{N}}, GU_{\tilde{N}} = \begin{bmatrix} L_{\tilde{M}}^{-1}L_{\tilde{N}} & 0 \end{bmatrix},$$

where $L_{\tilde{M}}, L_{\tilde{N}}$ are respectively square and lower (or upper) triangular with compatible dimensions. We refer to $L_{\tilde{G}} \triangleq L_{\tilde{M}}^{-1}L_{\tilde{N}}$ as the right triangular structure of $G(z)$.

The proposition that follows extends the previous results on the NCS in Fig. 6.1.

Proposition 6.4.1 *Consider the NCS in Fig. 6.3.*

(a) *Assume that $B = I$. Under PTS and Assumption 6.1, the networked system is mean square stabilizable via static output feedback if and only if*

$$\mathscr{C}_{\text{MS}} > \ln \mathscr{M}(A).$$

(b) *Assume that $G(z)$ is SISO and has no or distinct (if any) NMP zeros. The networked system is mean square stabilizable via dynamic output feedback if and only if*

$$\mathscr{C}_{\text{MS}} > \frac{1}{2} \ln \left\{ \mathscr{M}(G)^2 + \eta(G) + \varphi(G) \right\}.$$

(c) *Assume that $G(z)$ can be triangularly decoupled, $\ell \leq m$, and $[L_{\tilde{G}}]_{ii}$ has no or distinct (if any) NMP zeros. Under PTS and Assumption 6.1, the networked system is mean square stabilizable via dynamic output feedback if*

$$\mathscr{C}_{\text{MS}} > \frac{1}{2} \sum_{i=1}^{\ell} \ln \left\{ \mathscr{M}([L_{\tilde{G}}]_{ii})^2 + \eta([L_{\tilde{G}}]_{ii}) + \varphi([L_{\tilde{G}}]_{ii}) \right\},$$

and only if

$$\mathscr{C}_{\text{MS}} > \ln \mathscr{M}(A).$$

Proof The proof follows analogously to those of Theorem 6.2, Proposition 6.3.1 and Theorem 6.3. □

6.4.2 Stabilization of a Finite Platoon

In automated highway systems, one basic issue is to move a platoon of closely spaced vehicles from one place to another [21, 22]. Suppose that there are $\ell + 1$ vehicles in a platoon, and the leader, i.e., vehicle 0, generates a position trajectory $\{x_{0,k}\}_{k\geq 0}$ with $x_{0,0} = 0$ according to its local reference signal. Each follower, i.e., vehicle i, for all $i = 1, 2, \ldots, \ell$, uses its ranging sensor and local controller to keep a fixed distance behind the preceding vehicle.

Given any time-domain signal x_k, denote its z-transform by $X(z)$. Let $x_{i,k}$, $u_{i,k}$, $G_i(z)$ be the position, the input and the dynamics of the ith vehicle, and model the ith vehicle starting from rest by

$$X_i(z) = G_i(z)U_i(z) + \frac{zx_{i,0}}{z-1}, \quad i = 1, \ldots, \ell. \tag{6.55}$$

We assume that $G_i(z)$ is SISO and strictly proper. The vehicle separations $e_{i,k}$, $i = 1, 2, \ldots, \ell$, are defined as $e_{i,k} = x_{i,k} - x_{i-1,k} + \tau$, where $\tau > 0$ is a constant target separation for all followers. Suppose that the initial position of the ith vehicle, $x_{i,0}$, is $-i\tau$ indicating that the platoon starts with zero spacing errors. It follows that

$$E(z) = G(z)U(z) - [1\ 0\ \cdots\ 0]^T X_0(z)$$

with

$$E(z) = [E_1(z)\ E_2(z) \cdots E_\ell(z)]^T,\ U(z) = [U_1(z)\ U_2(z) \cdots U_\ell(z)]^T$$

and

$$G(z) = \begin{bmatrix} G_1(z) & 0 & 0 & \cdots & 0 \\ -G_1(z) & G_2(z) & 0 & \cdots & 0 \\ 0 & -G_2(z) & G_3(z) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -G_{\ell-1}(z) & G_\ell(z) \end{bmatrix}. \tag{6.56}$$

If we take the separation $e_{i,k}$ as the output of vehicle i and let its ranging sensor obtain $e_{i,k}$ through the ith channel as $v_{i,k} = \xi_{i,k}e_{i,k}$, then without considering $X_0(z)$, the platoon system fits in the model of Fig. 6.3 with $m = \ell$, $y_k = e_k$. Since the local controller on vehicle i only uses $v_{i,k}$, the distributed controller in Fig. 6.3 has the form

$$U(z) = \text{diag}\{K_1(z), K_2(z), \ldots, K_\ell(z)\}V(z), \tag{6.57}$$

where $V(z) = [V_1(z)\ V_2(z) \cdots V_\ell(z)]^T$ is the output vector of parallel channels. As in the model (12.4), each channel experiences transmission failure and Nakagami fading [23] simultaneously:

$$\xi_{i,k} = \Omega_{i,k}\Upsilon_{i,k}, \quad i = 1, 2, \ldots, \ell, \tag{6.58}$$

and $\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{\ell,k}$ are uncorrelated with each other. In (6.58), $\Omega_{i,k}$ is 0/1-valued (0 for "failure", 1 for "success") with probability distribution

$$\Pr\{\Omega_{i,k} = 0\} = \alpha_i,\ \Pr\{\Omega_{i,k} = 1\} = 1 - \alpha_i,\ 0 \le \alpha_i < 1,$$

and $\Upsilon_{i,k}$ is Nakagami distributed with ϖ_i denoting the mean channel power gain and $\rho_i \in [\frac{1}{2}, \infty)$ describing the severity of fading (i.e., the severity of fading decreases as ρ_i increases) [24]. Further assume that $\Omega_{i,k}, \Upsilon_{i,k}$ are uncorrelated, then it is easy to derive that

$$g_i = \frac{\rho_i \Gamma(\rho_i)^2}{\rho_i \Gamma(\rho_i)^2 - (1-\alpha_i)\Gamma(\rho_i + \frac{1}{2})^2}, \tag{6.59}$$

where $\Gamma(\cdot)$ is the gamma function. Based on the expression of g_i in (6.59) and $\mathscr{C}_{\mathrm{MS}i} = \frac{1}{2}\ln g_i$, it is intuitive to see that the smaller the α_i or the larger the ρ_i, the larger the g_i and $\mathscr{C}_{\mathrm{MS}i}$. The effect of transmission failure and Nakagami fading disappears as $\alpha_i \to 0$ and $\rho_i \to \infty$, respectively. In particular, we have

$$\lim_{\rho_i \to \infty} g_i = \frac{1}{\alpha_i}$$

since

$$\lim_{\rho_i \to \infty} \frac{\rho_i \Gamma(\rho_i)^2}{\Gamma(\rho_i + \frac{1}{2})^2} = 1,$$

which is reduced to the scenario considered in Corollary 5.1 of [25].

We have the next corollary on stabilization of a finite platoon.

Corollary 6.1 *Assume that $G_i(z)$, for all $i = 1, 2, \ldots, \ell$, is SISO, strictly proper and has no or distinct (if any) NMP zeros. Then, the finite platoon system described above with $\ell + 1$ vehicles is mean square stabilizable via dynamic output feedback over the network* (6.58) *if and only if*

$$\mathscr{C}_{\mathrm{MS}i} > \frac{1}{2}\ln\left\{\mathscr{M}(G_i)^2 + \eta(G_i) + \varphi(G_i)\right\},\ i = 1, 2, \ldots, \ell,$$

where $\mathscr{C}_{\mathrm{MS}i} = \frac{1}{2}\ln g_i$, and g_i is given in (6.59).

Proof Based on the diagonal structure of $K(z)$ in (6.57) and the lower triangular structure of $G(z)$ in (6.56), without considering both $X_0(z)$ and the uncertainty block, we have

$$\begin{aligned} T(z) &= G(z)(I - K(z)\Pi G(z))^{-1}K(z)\Lambda \\ &= \begin{bmatrix} T_1 & 0 & 0 & \cdots & 0 \\ -T_1S_2 & T_2 & 0 & \cdots & 0 \\ T_1T_2S_3 & -T_2S_3 & T_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \prod_{i=1}^{\ell-1}(-T_i)S_\ell & \cdots & T_{\ell-2}T_{\ell-1}S_\ell & -T_{\ell-1}S_\ell & T_\ell \end{bmatrix} \Pi^{-1}\Lambda, \end{aligned} \tag{6.60}$$

where $S_i = (1 - K_i\mu_i G_i)^{-1}$ and $T_i = G_i(1 - K_i\mu_i G_i)^{-1}K_i\mu_i$. Denote the set of all proper controllers achieving the internal stability by $\mathscr{K}$. It follows from the proofs of Propositions 6.4.1 (b) and 6.3.1 that

$$\inf_{K(z)\in\mathscr{K}} \|T_i(z)\|_2^2 = \mathscr{M}(G_i)^2 + \eta(G_i) + \varphi(G_i) - 1. \tag{6.61}$$

For any diagonal scaling matrix Θ, the condition (6.38) implies that

$$g_i > \mathscr{M}(G_i)^2 + \eta(G_i) + \varphi(G_i),$$

which shows the necessity.

For the sufficiency, a constructive proof will be given. By choosing $K_i(z)$ based on (6.61), we can ensure that $G_i(1 - K_i\mu_i G_i)^{-1}$, $(1 - K_i\mu_i G_i)^{-1}K_i\mu_i$ and S_i are all stable and proper. It is then easy to verify that

$$K(z) = \text{diag}\{K_1(z), K_2(z), \ldots, K_\ell(z)\} \in \mathscr{K}$$

by invoking the special structures of $G(z)$ and $K(z)$, i.e., the internal stability of the overall closed loop in Fig. 6.3 is guaranteed. Thus, by setting

$$\Theta = \text{diag}\{\varepsilon^{\ell-1}, \varepsilon^{\ell-2}, \ldots, 1\}$$

with sufficiently small ε, we have that (6.38) is true. This completes the proof. □

The numerical example as below demonstrates the usefulness of Corollary 6.1.

Example 6.1 Consider a homogeneous platoon with 5 vehicles and

$$G_i(z) = \frac{1}{(z-1)^2}, \quad i = 1, 2, 3, 4.$$

Suppose that the leader has a constant speed: 1 meter/second. It is easy to check that the infimum in (6.61) is 0 which is not achievable by any stabilizing controller due to the double marginally stable poles of $G_i(z)$. If we approximate $z = 1$ by $z = 1 - \varepsilon$ with a small real number $\varepsilon > 0$, then the design procedure in [19] yields the controller:

$$K_i = \frac{2\varepsilon - (2\varepsilon + \varepsilon^2)z}{(z + 2\varepsilon)\mu_i}, \tag{6.62}$$

which can approach the infimum in (6.61) with any desired accuracy by choosing a sufficiently small ε.

Let $\alpha_i = 0.2$, $\varpi_i = 2$, $\rho_i = 2$ for the ith channel, then g_i in (6.59) is 3.4113. By setting $\varepsilon = 0.1$ in (6.62), we have

$$\|T_i(z)\|_2^2 = \|G_i(1 - K_i\mu_i G_i)^{-1}K_i\mu_i\|_2^2 = 0.1489.$$

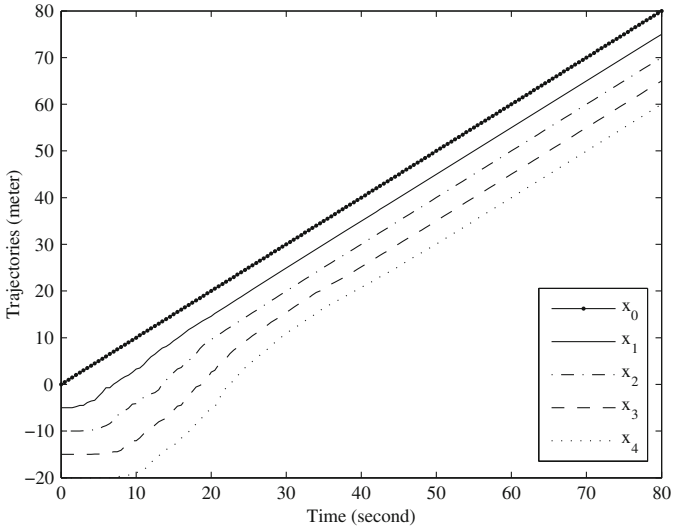

Fig. 6.4 Vehicle trajectories in the platoon for one sample of simulation

Since $\|T_i(z)\|_2^2/(g_i-1) = 0.0436 < 1$ for all $i = 1, 2, 3, 4$, Corollary 6.1 implies that the platoon can be stabilized in the mean square sense by the controller (6.62) with $\varepsilon = 0.1$. The trajectories of the vehicles in the platoon for one sample of simulation are shown in Fig. 6.4.

6.5 Channel Processing and Channel Feedback

As pointed out in Remark 6.8, additional degree of freedom may benefit the NCS in reducing the capacity requirement for stabilizability. Here, we further analyze the possible advantages of introducing pre- and post-channel filters and channel feedback into the NCS in Fig. 6.1. To this end, we consider the NCS in Fig. 6.5.

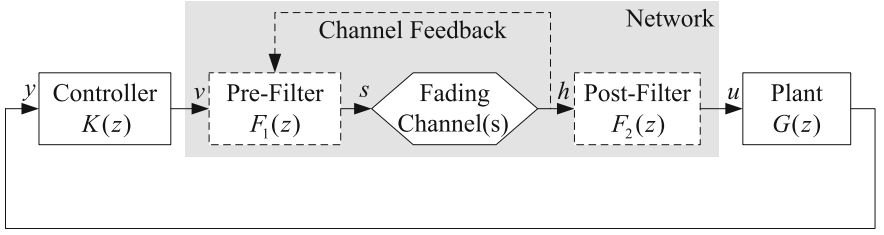

Fig. 6.5 Control over input fading channel(s) with channel processing and channel feedback

The system interconnections are described as follows:

$$y = G(z)u,\ v = K(z)y,\ s = F_{1v}(z)v + F_{1h}(z)h,\ u = F_2(z)h,\ h_k = \xi_k s_k, \quad (6.63)$$

where $F_1(z) = [F_{1v}(z)\ F_{1h}(z)]$, $s_k \in \mathbb{R}^m$ is the channel input, $h_k \in \mathbb{R}^m$ is the channel output, and ξ_k has the form as in (6.4). We assume that the channel feeds h back to the pre-filter with one-step delay, therefore $F_{1h}(z)$ is assumed to be strictly proper. The channel feedback in Fig. 6.5 is equivalent to sending the channel state information (CSI) back to the pre-filter. We can derive the following result.

Theorem 6.4 *Consider the NCS in Fig. 6.5, where $K(z), F_1(z), F_2(z)$ are to be designed. Under PTS and Assumption 6.1, the networked system is mean square stabilizable via state feedback or dynamic output feedback if and only if*

$$\mathscr{C}_{\mathrm{MS}} > \ln \mathscr{M}(A).$$

Proof First, we will show the necessity for the dynamic state feedback case, which also implies the necessity for the static state feedback and dynamic output feedback cases. Note that Lemma 6.1 is still true with the overall system state

$$x_k' = [x_k^T\ x_{K,k}^T\ x_{F_{1,k}}^T\ x_{F_{2,k}}^T]^T,$$

where $x_{F_{1,k}}$ and $x_{F_{2,k}}$ are the states of the pre-filter and the post-filter, respectively. Suppose that $F_2(z)$ has a state-space realization

$$\left[\begin{array}{c|c} A_{F_2} & B_{F_2} \\ \hline C_{F_2} & D_{F_2} \end{array}\right],$$

then the interconnection between the plant and the post-filter yields

$$G(z)F_2(z) = \left[\begin{array}{cc|c} A & BC_{F_2} & BD_{F_2} \\ 0 & A_{F_2} & B_{F_2} \\ \hline C & 0 & 0 \end{array}\right].$$

Considering $G(z)F_2(z)$ as an augmented plant and following similar lines of the proof of Theorem 6.3, we have

$$\prod_{i=1}^{m}\left(1 + \frac{\mu_i^2}{\sigma_i^2}\right) > \mathscr{M}\left(\left[\begin{array}{cc} A & BC_{F_2} \\ 0 & A_{F_2} \end{array}\right]\right)^2 = \mathscr{M}(A)^2\mathscr{M}(A_{F_2})^2 \geq \mathscr{M}(A)^2,$$

i.e., $\mathscr{C}_{\mathrm{MS}} > \ln \mathscr{M}(A)$.

To show the sufficiency, a constructive proof will be given. Let $F_{1v}(z) = I$, $F_2(z) = I$, then we have $s = K(z)y + F_{1h}(z)u$. Construct an observer-based realization for $[K(z)\ F_{1h}(z)]$ as

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + L(y_k - C\hat{x}_k) + Bu_k, \\ s_k &= K\hat{x}_k, \end{aligned} \quad (6.64)$$

where L and K are respectively the observer gain and the state feedback gain to be designed. Then, $F_{1h}(z)$ is strictly proper, and the closed loop with $x'_k = [x_k^T \ \hat{x}_k^T]^T$ is given by

$$\begin{aligned} x'_{k+1} &= \left(\begin{bmatrix} A & 0 \\ LC & A - LC \end{bmatrix} + \begin{bmatrix} B \\ B \end{bmatrix} \xi_k [0 \ \ K] \right) x'_k \\ &= \begin{bmatrix} I & 0 \\ I & I \end{bmatrix} \left(\begin{bmatrix} A & 0 \\ 0 & A - LC \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} \xi_k [K \ \ K] \right) \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}^{-1} x'_k. \end{aligned}$$

By using the similar technique in the proof of the equivalence between (i) and (v) in Lemma 6.2, we can show that the mean square stability of the closed-loop system is ensured by selecting L such that $\rho(A - LC) < 1$ and K as in the sufficiency part of Theorem 6.2. □

Remark 6.11 As we can see from Theorem 6.2 and Theorem 6.4, when the full state is available to the controller, both the filtering and the channel feedback of the network are redundant in stabilization. By comparing Proposition 6.3.1 and Theorem 6.3 with Theorem 6.4 and in view of the proof of Theorem 6.4, we can conclude that, in the case of dynamic output feedback, (1) $F_{1v}(z)$ and $F_2(z)$ are redundant parameters in stabilizing the NCS in Fig. 6.5; (2) the channel feedback and $F_{1h}(z)$ can completely eliminate the limitation on stabilization of the NCS induced by the NMP zeros and high relative degree of the plant; (3) the post-filter should be chosen to be stable otherwise the network requirement would be more stringent; (4) the assumptions on the triangular decoupling of the plant and the distinctness of NMP zeros are not needed in Theorem 6.4.

6.6 Power Constraint

In this section, we further analyze the effect of channel additive noise and channel input power bound on stability and performance of the NCS. In the following, a random variable or process is said to be uncorrelated, without specifying with respect to which variables or processes, if it is uncorrelated with any other random variable or process. If a scalar-valued discrete-time stochastic process

$$x = \{x_k\}_{k=0}^{\infty}$$

is bounded and convergent in the mean square sense, then the power norm of x is defined as

$$\|x\|_{\mathscr{P}} \triangleq \sqrt{\lim_{k \to \infty} \mathscr{E}\{x_k^2\}}.$$

Both the time index k and the frequency variable z will be omitted whenever no confusion is caused.

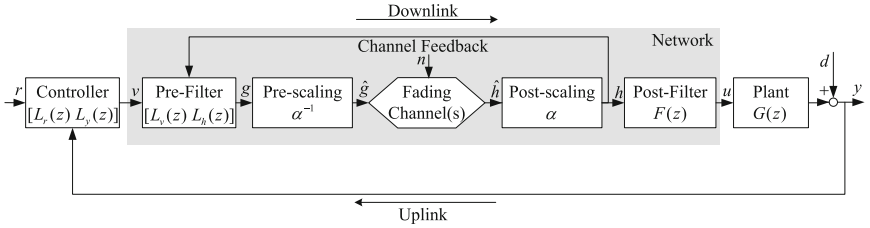

Fig. 6.6 Control over an input fading channel with channel additive noise and network scaling

We limit our attention to the SISO LTI case and consider the NCS as depicted in Fig. 6.6. The SISO plant has a transfer function $G(z)$, and the measured output $y \in \mathbb{R}$ is given by

$$y = G(z)u + d, \tag{6.65}$$

where $u \in \mathbb{R}$ is the plant input, and $d \in \mathbb{R}$ is the output disturbance. The controller and the pre-filter have the form

$$v = L_r(z)r + L_y(z)y, \tag{6.66}$$

$$g = L_v(z)v + L_h(z)h, \tag{6.67}$$

where $g \in \mathbb{R}$, $r \in \mathbb{R}$, $h \in \mathbb{R}$ are respectively the output of the pre-filter, the reference signal, and the input of the post-filter. The post-filter has a transfer function $F(z)$ as

$$u = F(z)h. \tag{6.68}$$

Let the pre-scaling and post-scaling in Fig. 6.6 be α^{-1} and α respectively, then we have

$$\hat{g} = \alpha^{-1}g,\ h = \alpha\hat{h}, \tag{6.69}$$

and α is called the scaling factor of the network. The model of the single fading channel is described as

$$\hat{h}_k = \xi_k\hat{g}_k + n_k, \tag{6.70}$$

where $\xi_k \in \mathbb{R}$ denotes the channel fading, and $n_k \in \mathbb{R}$ represents the channel additive noise.

Denote the state of the overall system in Fig. 6.6 by x'_k. The assumption that follows will be adopted throughout this section.

Assumption 6.3 Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70).

(i) [Constraints on LTI components] The plant transfer function $G(z)$ is unstable and strictly proper. The transfer functions

$$L_h(z) \in \mathscr{R}_{\mathrm{sp}}, L_r(z), L_y(z), L_v(z), F(z) \in \mathscr{R}_\ell$$

are to be designed. The scaling factor of the network satisfies $0 < \alpha < \infty$.

(ii) [Constraints on the randomness] The initial state of the whole system, x_0', is an uncorrelated second-order random variable with mean $\mu_{x',0}$ and second-order moment matrix Ξ_0. The exogenous signals r, d, n are uncorrelated white noise processes with zero means and bounded variances $\sigma_r^2, \sigma_d^2, \sigma_n^2$. The channel fading ξ_k is an uncorrelated white noise process with mean $\mu_\xi \neq 0$ and bounded variance σ_ξ^2.

(iii) [Constraint on the power of channel input] The channel input $\hat{g}$ satisfies

$$\|\hat{g}\|_{\mathscr{P}}^2 < \mathscr{P}, \tag{6.71}$$

and $\mathscr{P} > 0$ is the predefined power level.

Remark 6.12 Note that the strict properness of $G(z)$ and the constraints on the randomness lead to simpler formulae and can be relaxed at the expense of more involved expressions. For communication systems, a power bound on the channel input as in (6.71) is usually introduced in order to avoid the interference to other communication users and/or due to the hardware limitations of the transmitter.

6.6.1 Feedback Stabilization

The concept of stability used in this section is presented as follows.

Definition 6.3 Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70). The closed-loop system is said to be mean square stable, if, for every x_0', r, d, n, ξ satisfying Assumption 6.3 (ii), there exists a bounded matrix $\Xi_{x'} \geq 0$ independent of x_0' such that $\Xi_k \triangleq \mathscr{E}\{x_k' x_k'^T\}$ is well defined for every $k \geq 0$, and $\lim_{k\to\infty} \Xi_k = \Xi_{x'}$.

Since the controller is collocated with the pre-filter, thus we can lump (6.66) and (6.67) together as

$$g = L_v(z)L_r(z)r + L_v(z)L_y(z)y + L_h(z)h. \tag{6.72}$$

Under Assumption 6.3, the NCS in Fig. 6.6 with Eqs. (6.65)–(6.70) is equivalent to the NCS in Fig. 6.7, where $z_k = \hat{g}_k$, $w_k = \Delta_k z_k$, and

$$L(z) \triangleq [L_v(z)L_r(z)\ L_v(z)L_y(z)\ L_h(z)],$$
$$\Delta_k \triangleq \sigma_\xi^{-1}\left(\xi_k - \mu_\xi\right).$$

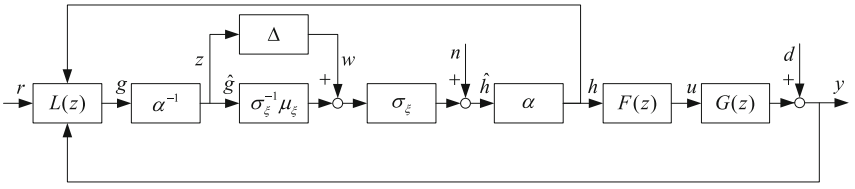

Fig. 6.7 Representation of the NCS in Fig. 6.6 with channel model (6.70) under Assumption 6.3

It is easy to verify that the mean square stability of the NCS in Fig. 6.7 always implies the internal stability of the corresponding closed-loop system in the absence of r, d, n, w. Let

$$K(z) \triangleq \begin{bmatrix} L_v(z)L_r(z) & L_v(z)L_y(z) & L_h(z) \\ 0 & 0 & F(z) \end{bmatrix}, \tag{6.73}$$

and denote the set of $K(z)$ satisfying Assumption 6.3 (i) and achieving the above internal stability by $\mathscr{K}$.

The theorem that follows presents a necessary and sufficient condition for mean square stabilizability.

Theorem 6.5 *Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70). Under Assumption 6.3, the plant $G(z)$ can be mean square stabilized by designing $K(z)$ and α if and only if*

$$\frac{\mu_\xi^2 \mathscr{P}}{\sigma_\xi^2 \mathscr{P} + \sigma_n^2} + 1 > \mathscr{M}(G)^2. \tag{6.74}$$

Proof To show the necessity of (6.74), we first assume that the closed-loop system is mean square stable. In this case, all signals that appear in the loop of Fig. 6.7 are convergent in the mean square sense. Therefore, without loss of generality we can assume that $x_0' = 0$ and consider directly the limiting second-order moments of the relevant signals. When an SISO plant is controlled over a single fading channel, we have $\|T_{w\hat{g}}(z)\|_2^2 < 1$ from Theorem 6.4 in [1], where $T_{w\hat{g}}(z)$ denotes the transfer function of the closed-loop from w to $\hat{g}$ without considering the uncertainty block Δ. It is easy to see from Fig. 6.7 that

$$\hat{g} = T_{r\hat{g}} r + T_{d\hat{g}} d + T_{n\hat{g}} n + T_{w\hat{g}} w, \tag{6.75}$$

where

$$T_{r\hat{g}} = \alpha^{-1}(1 - L_v L_y G F \mu_\xi - L_h \mu_\xi)^{-1} L_v L_r, \tag{6.76}$$

$$T_{d\hat{g}} = \alpha^{-1}(1 - L_v L_y G F \mu_\xi - L_h \mu_\xi)^{-1} L_v L_y, \tag{6.77}$$

$$T_{n\hat{g}} = (1 - L_v L_y G F \mu_\xi - L_h \mu_\xi)^{-1}(L_v L_y G F + L_h), \tag{6.78}$$

$$T_{w\hat{g}} = T_{n\hat{g}} \sigma_\xi. \tag{6.79}$$

In view of Assumption 6.3 (iii), (6.75), $w = \Delta\hat{g}$ and $1 - \|T_{w\hat{g}}(z)\|_2^2 > 0$, we have

$$\begin{aligned} \mathscr{P} > \|\hat{g}\|_{\mathscr{P}}^2 &= \|T_{r\hat{g}}\|_2^2 \sigma_r^2 + \|T_{d\hat{g}}\|_2^2 \sigma_d^2 + \|T_{w\hat{g}}\|_2^2 \left(\frac{\sigma_n^2}{\sigma_\xi^2} + \|\hat{g}\|_{\mathscr{P}}^2 \right) \\ &= \frac{\|T_{r\hat{g}}\|_2^2 \sigma_r^2 + \|T_{d\hat{g}}\|_2^2 \sigma_d^2 + \|T_{w\hat{g}}\|_2^2 \frac{\sigma_n^2}{\sigma_\xi^2}}{1 - \|T_{w\hat{g}}\|_2^2} \end{aligned} \tag{6.80}$$

$$\geq \frac{\|T_{w\hat{g}}\|_2^2 \frac{\sigma_n^2}{\sigma_\xi^2}}{1 - \|T_{w\hat{g}}\|_2^2}, \tag{6.81}$$

i.e.,

$$\frac{\sigma_\xi^2 \mathscr{P}}{\sigma_\xi^2 \mathscr{P} + \sigma_n^2} > \|T_{w\hat{g}}\|_2^2. \tag{6.82}$$

Rewrite $T_{w\hat{g}}$ in (6.79) as

$$T_{w\hat{g}} = (1 - K_1 G - K_2)^{-1}(K_1 G + K_2)\mu_\xi^{-1}\sigma_\xi$$

with

$$K_1 = L_v L_y F \mu_\xi,\ K_2 = L_h \mu_\xi. \tag{6.83}$$

Let

$$\breve{G} = \begin{bmatrix} G \\ z^{-1} \end{bmatrix}, \breve{K} = [K_1 \ \ K_2 z].$$

Under Assumption 6.3 (i), we have

$$\begin{aligned} \inf_{K \in \mathscr{K}, \alpha \in (0,\infty)} \|T_{w\hat{g}}\|_2^2 &= \inf_{\text{IS}, K_1 \in \mathscr{R}_\ell, K_2 \in \mathscr{R}_{\text{sp}}} \left\| (1 - K_1 G - K_2)^{-1}(K_1 G + K_2)\frac{\sigma_\xi}{\mu_\xi} \right\|_2^2 \\ &= \frac{\sigma_\xi^2}{\mu_\xi^2} \inf_{\text{IS}, \breve{K} \in \mathscr{R}_\ell} \|(1 - \breve{K}\breve{G})^{-1}\breve{K}\breve{G}\|_2^2 \qquad (6.84) \\ &= \frac{\sigma_\xi^2}{\mu_\xi^2}(\mathscr{M}(G)^2 - 1), \qquad (6.85) \end{aligned}$$

where the abbreviation "IS" stands for the constraint on the internal stability of the closed-loop system, and (6.85) follows from the $\mathscr{H}_2$ optimization and the Residue theorem; see, e.g., Theorem 4.1 in [26]. It follows from (6.85) that

$$\|T_{w\hat{g}}\|_2^2 \geq \frac{\sigma_\xi^2}{\mu_\xi^2}(\mathscr{M}(G)^2 - 1). \tag{6.86}$$

By combining the inequalities (6.82) and (6.86), we can conclude that (6.74) is true.

To show the converse, we will prove that if (6.74) is true, then there exists a set of $K(z)$ and α satisfying Assumption 6.3 such that the closed-loop system is mean square stable. Note that it is always possible to select $\breve{K}$ in (6.84) appropriately such that

$$\|T_{w\hat{g}}\|_2^2 \leq \frac{\sigma_\xi^2}{\mu_\xi^2}(\mathscr{M}(G)^2 - 1) + \varepsilon_1 \tag{6.87}$$

for any small positive real number ε_1. Choose $\varepsilon_1 \leq \frac{\sigma_n^2}{\mu_\xi^2 \mathscr{P}}$, then $\|T_{w\hat{g}}\|_2^2 < 1$ is true under the condition (6.74). It is direct to derive a set of

$$L_v \in \mathscr{R}_\ell, L_y \in \mathscr{R}_\ell, F \in \mathscr{R}_\ell, L_h \in \mathscr{R}_{\text{sp}}$$

according to

$$\breve{K} = [L_v L_y F \mu_\xi \;\; L_h \mu_\xi z] \tag{6.88}$$

for the given $\breve{K} \in \mathscr{R}_\ell$, and the internal stability of the closed-loop system in Fig. 6.7 is ensured after choosing a proper and stable L_r (e.g., $L_r = 0$). The mean square stability of the closed loop follows from Theorem 6.4 in [1]. The only thing left to be shown is that the power constraint $\|\hat{g}\|_{\mathscr{P}}^2 < \mathscr{P}$ can be satisfied along with the above mean square stability. Since the closed-loop system has been stabilized in the mean square sense, in the following we can merely focus on the limiting second-order moments of relevant signals. According to the expressions of $T_{r\hat{g}}, T_{d\hat{g}}$ in (6.76) and (6.77), we can ensure that, for any small positive real number ε_2,

$$\|T_{r\hat{g}}\|_2^2 \sigma_r^2 + \|T_{d\hat{g}}\|_2^2 \sigma_d^2 \leq \varepsilon_2 \tag{6.89}$$

by choosing a sufficiently large α. It follows from (6.87), (6.89) and (6.80) that

$$\|\hat{g}\|_{\mathscr{P}}^2 \leq \frac{\varepsilon_2 + \left[\frac{\sigma_\xi^2}{\mu_\xi^2}(\mathscr{M}(G)^2 - 1) + \varepsilon_1\right] \frac{\sigma_n^2}{\sigma_\xi^2}}{1 - \left[\frac{\sigma_\xi^2}{\mu_\xi^2}(\mathscr{M}(G)^2 - 1) + \varepsilon_1\right]}. \tag{6.90}$$

Thus, $\mathscr{P} > \|\hat{g}\|_{\mathscr{P}}^2$ if

$$\frac{\mu_\xi^2 \mathscr{P}}{\sigma_\xi^2 \mathscr{P} + \sigma_n^2} \geq \mathscr{M}(G)^2 - 1 + \frac{\varepsilon_1 (\frac{\mu_\xi^2}{\sigma_\xi^2}\sigma_n^2 + \mu_\xi^2 \mathscr{P}) + \varepsilon_2 \mu_\xi^2}{\sigma_\xi^2 \mathscr{P} + \sigma_n^2}. \tag{6.91}$$

We note that (6.74) implies (6.91) if both ε_1 and ε_2 are made to be sufficiently small. Therefore, $\|\hat{g}\|_{\mathscr{P}}^2 < \mathscr{P}$ can be satisfied by designing K and α, which completes the proof of sufficiency. □

Remark 6.13 Theorem 6.5 recovers the scenario with a pure additive noise channel by setting $\mu_\xi = 1$ and $\sigma_\xi^2 = 0$, and the inequality (6.74) comes to be

$$\frac{\mathscr{P}}{\sigma_n^2} + 1 > \mathscr{M}(G)^2, \text{ i.e., } \frac{1}{2}\ln\left(\frac{\mathscr{P}}{\sigma_n^2} + 1\right) > \ln \mathscr{M}(G),$$

as identified in Theorem 4.1 of [26], where $\frac{1}{2}\ln\left(\frac{\mathscr{P}}{\sigma_n^2} + 1\right)$ is the communication capacity of an additive white Gaussian noise channel. Theorem 6.5 also recovers the

scenario with a pure multiplicative noise channel by letting $\sigma_n^2 = 0$ or $\mathscr{P} \to \infty$, where α becomes redundant and can be chosen arbitrarily from $(0, \infty)$ (e.g., $\alpha = 1$), and the condition (6.74) is reduced to

$$\frac{\mu_\xi^2}{\sigma_\xi^2} + 1 > \mathscr{M}(G)^2, \text{ i.e., } \frac{1}{2}\ln\left(\frac{\mu_\xi^2}{\sigma_\xi^2} + 1\right) > \ln \mathscr{M}(G), \tag{6.92}$$

where

$$\frac{1}{2}\ln\left(\frac{\mu_\xi^2}{\sigma_\xi^2} + 1\right)$$

is the mean square capacity of a multiplicative noise channel as defined in (6.28). As we can see from (6.74) and (6.92), when the power of channel input is limited, a more stringent requirement on the quality of the channel is needed for mean square stabilizability.

Remark 6.14 Note that L_v, L_y, F play the same role in Theorem 6.5, and it is thus sufficient to use any of them in stabilizing the NCS. For example, for any $\check{K} \in \mathscr{R}_\ell$, (6.88) can be easily satisfied by designing L_y and selecting $L_v, F \in \mathscr{U}_\infty$ (e.g., $L_v = F = 1$) without affecting the degree of freedom and the internal stability. In addition, the scaling factor of the network α is an essential design parameter in satisfying the power constraint on the channel input.

Different from the model (6.70), another channel model is given by

$$\hat{h}_k = \xi_k(\hat{g}_k + n_k), \tag{6.93}$$

where ξ_k and n_k are as defined in (6.70). The model (6.93) is suitable for describing the digital erasure channel, where n_k is used to model uncertainties arising from quantization, and ξ_k is 0-1 binary valued and represents the packet-loss process. In this case, the NCS in Fig. 6.6 with Eqs. (6.65)–(6.69) and (6.93) can be rewritten into the NCS in Fig. 6.8, and Theorem 6.5 can be readily extended as follows.

Corollary 6.2 *Consider the NCS described by Fig. 6.8 and Eqs. (6.65)–(6.69) and (6.93). Under Assumption 6.3, the plant $G(z)$ can be mean square stabilized by*

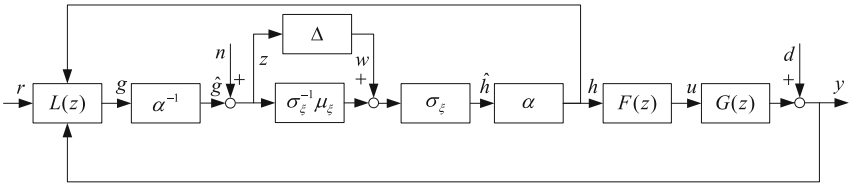

Fig. 6.8 Representation of the NCS in Fig. 6.6 with channel model (6.93) under Assumption 6.3

designing $K(z)$ and α if and only if

$$\frac{\mu_\xi^2 \mathscr{P}}{\sigma_\xi^2 \mathscr{P} + \left(\mu_\xi^2 + \sigma_\xi^2\right)\sigma_n^2} + 1 > \mathscr{M}(G)^2. \tag{6.94}$$

6.6.2 Performance Design

Denote the tracking error by $e = r - y$, and use its power norm $J \triangleq \|e\|_{\mathscr{P}}^2$ as the performance index to be minimized. First of all, we have the following result.

Proposition 6.6.1 *Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70). Under Assumption 6.3, J is bounded if $G(z)$ is stabilized in the mean square sense by $K(z)$ and α.*

Proof If the closed-loop system is mean square stable, then we can easily deduce that

$$e = r - (T_{ry} r + T_{dy} d + T_{ny} n + T_{wy} w), \tag{6.95}$$

where

$$\begin{aligned}
T_{ry} &= GF(1 - \mu_\xi L_v L_y GF - \mu_\xi L_h)^{-1} \mu_\xi L_v L_r, \\
T_{dy} &= GF(1 - \mu_\xi L_v L_y GF - \mu_\xi L_h)^{-1} \mu_\xi L_v L_y + 1, \\
T_{wy} &= GF(1 - \mu_\xi L_v L_y GF - \mu_\xi L_h)^{-1} \alpha \sigma_\xi, \\
T_{ny} &= T_{wy} \sigma_\xi^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
J &= \|T_{wy}\|_2^2 \left(\frac{\sigma_n^2}{\sigma_\xi^2} + \|\hat{g}\|_{\mathscr{P}}^2 \right) + \|1 - T_{ry}\|_2^2 \sigma_r^2 + \|T_{dy}\|_2^2 \sigma_d^2 \\
&= \|T_{wy}\|_2^2 \left(\frac{\sigma_n^2}{\sigma_\xi^2} + \frac{\|T_{r\hat{g}}\|_2^2 \sigma_r^2 + \|T_{d\hat{g}}\|_2^2 \sigma_d^2 + \|T_{w\hat{g}}\|_2^2 \frac{\sigma_n^2}{\sigma_\xi^2}}{1 - \|T_{w\hat{g}}\|_2^2} \right) \\
&\quad + \|1 - T_{ry}\|_2^2 \sigma_r^2 + \|T_{dy}\|_2^2 \sigma_d^2,
\end{aligned} \tag{6.96}$$

where the expression for $\|\hat{g}\|_{\mathscr{P}}^2$ is given in (6.80). Since the mean square stability of the closed-loop system always implies the internal stability and the inequality $\|T_{w\hat{g}}\|_2^2 < 1$, the performance index J in (6.96) is bounded. □

After introducing the notation:

$$\begin{aligned}
\beta_1 &\triangleq \|T\mu_\xi^{-1}\|_2^2,\\
\beta_2 &\triangleq \|1 - GFS\mu_\xi L_v L_r\|_2^2\sigma_r^2 + \|GFS\mu_\xi L_v L_y + 1\|_2^2\sigma_d^2,\\
\beta_3 &\triangleq \|GFS\|_2^2,\\
\beta_4 &\triangleq \|SL_v L_r\|_2^2\sigma_r^2 + \|SL_v L_y\|_2^2\sigma_d^2
\end{aligned}$$

with $S = (1 - \mu_\xi L_v L_y GF - \mu_\xi L_h)^{-1}$ and $T = 1 - S$, we have

$$\|\hat{g}\|_{\mathscr{P}}^2 = \frac{\beta_1\sigma_n^2 + \alpha^{-2}\beta_4}{1 - \beta_1\sigma_\xi^2}, \tag{6.97}$$

$$\begin{aligned}
J &= \beta_2 + \alpha^2\beta_3\sigma_\xi^2\left(\frac{\sigma_n^2}{\sigma_\xi^2} + \frac{\beta_1\sigma_n^2 + \alpha^{-2}\beta_4}{1 - \beta_1\sigma_\xi^2}\right)\\
&= \beta_2 + \frac{\beta_3\beta_4}{1 - \beta_1\sigma_\xi^2}\sigma_\xi^2 + \alpha^2\frac{\beta_3\sigma_n^2}{1 - \beta_1\sigma_\xi^2}.
\end{aligned} \tag{6.98}$$

For a given set of $G(z)$, σ_r^2, σ_d^2, σ_n^2, σ_ξ^2, μ_ξ, $\mathscr{P}$ satisfying (6.74), the globally optimal solution to

$$(K, \alpha) = \arg \inf_{K\in\mathscr{K},\ \alpha\in(0,\infty)} J$$

is hard to obtain due to the coupling among the design parameters. Therefore, we turn to a suboptimal design and adopt the perfect reconstruction constraint (see, e.g., [26, 27]) on the network.

Assumption 6.4 The network in Fig. 6.6 has a unit transfer function from v to u in the absence of n and w, i.e.,

$$F\mu_\xi(1 - L_h\mu_\xi)^{-1}L_v = 1. \tag{6.99}$$

Under Assumption 6.4, we have

$$\begin{aligned}
S &= (1 - L_y G)^{-1}(1 - L_h\mu_\xi)^{-1},\\
\beta_2 &= \|1 - G(1 - L_y G)^{-1}L_r\|_2^2\sigma_r^2 + \|(1 - L_y G)^{-1}\|_2^2\sigma_d^2.
\end{aligned}$$

From (6.98), J is an increasing function in β_1, β_2, β_3 and β_4. First of all, we can choose $L_h\mu_\xi$ and $L_v L_y F\mu_\xi$ according to (6.88) and let $L_r \in \mathscr{RH}_\infty$, such that the mean square stability of the closed-loop system is ensured and β_1 is minimized. In this

situation, L_h and L_y are determined under Assumption 6.4. Secondly, $L_r \in \mathscr{RH}_\infty$ can be designed to minimize

$$\|1 - G(1 - L_y G)^{-1} L_r\|_2^2,$$

which amounts to minimize β_2 since $\|(1 - L_y G)^{-1}\|_2^2$ is fixed with a given L_y. In the last step, the design of both α and the pair F, L_v is given in the theorem that follows.

Theorem 6.6 *Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70). Further assume that (6.74) is true, and L_y, L_h, L_r are chosen as above. Under Assumptions 6.3 and 6.4,*

$$\begin{aligned} J_{opt} &\triangleq \inf_{\text{IS}, L_v \in \mathscr{U}_\infty, F \in \mathscr{U}_\infty, \alpha \in (0,\infty)} J \\ &= \beta_2 + \frac{\beta_5}{1 - \beta_1 \sigma_\xi^2} \left\{ \sigma_\xi^2 + \frac{\sigma_n^2}{(1 - \beta_1 \sigma_\xi^2)\mathscr{P} - \beta_1 \sigma_n^2} \right\} \end{aligned}$$

with

$$\begin{aligned} \beta_5 \triangleq \Bigg(& \frac{1}{2\pi} \int_{-\pi}^{\pi} |G(e^{j\omega}) S(e^{j\omega})^2 (1 - L_h(e^{j\omega})\mu_\xi) \mu_\xi^{-1}| \\ & \sqrt{|L_r(e^{j\omega})\sigma_r|^2 + |L_y(e^{j\omega})\sigma_d|^2} d\omega \Bigg)^2. \end{aligned}$$

Moreover, J_{opt} can be approached by choosing $L_v \in \mathscr{U}_\infty$ such that, for any $\omega \in [-\pi, \pi]$,

$$|L_v(e^{j\omega})|^4 \approx \gamma \frac{|(1 - L_h(e^{j\omega})\mu_\xi) G(e^{j\omega})|^2}{\mu_\xi^2 \left(|L_r(e^{j\omega})\sigma_r|^2 + |L_y(e^{j\omega})\sigma_d|^2 \right)}, \tag{6.100}$$

$F = (1 - L_h \mu_\xi) L_v^{-1} \mu_\xi^{-1}$, *and*

$$\alpha = \sqrt{\frac{\beta_4}{(1 - \beta_1 \sigma_\xi^2)\mathscr{P} - \beta_1 \sigma_n^2}} + \varepsilon_3,$$

where $\gamma > 0$ is arbitrary, and $\varepsilon_3 > 0$ is sufficiently small.

Proof Based on the expression (6.97), the power constraint $\|\hat{g}\|_{\mathscr{P}}^2 < \mathscr{P}$ indicates that

$$\alpha > \sqrt{\frac{\beta_4}{(1 - \beta_1 \sigma_\xi^2)\mathscr{P} - \beta_1 \sigma_n^2}}. \tag{6.101}$$

By combining (6.98) and (6.101), we have

$$J > \beta_2 + \frac{\beta_3\beta_4}{1-\beta_1\sigma_\xi^2}\left\{\sigma_\xi^2 + \frac{\sigma_n^2}{(1-\beta_1\sigma_\xi^2)\mathscr{P}-\beta_1\sigma_n^2}\right\}.$$

It follows from the Cauchy-Schwartz inequality that

$$\begin{aligned}\beta_3\beta_4 &= \|GFS\|_2^2\|SL_v[L_r\sigma_r \;\; L_y\sigma_d]\|_2^2\\ &\geq \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|G(e^{j\omega})F(e^{j\omega})S(e^{j\omega})^2L_v(e^{j\omega})|\sqrt{|L_r(e^{j\omega})\sigma_r|^2+|L_y(e^{j\omega})\sigma_d|^2}d\omega\right)^2\\ &= \beta_5, \end{aligned} \tag{6.102}$$

where the last equality follows from the perfect reconstruction condition (6.99), and (6.102) can approach the equality with any desired accuracy by choosing L_v according to (6.100); see, e.g., [26, 28]. The remaining proof is straightforward. □

To sum up, we propose the next algorithm for a suboptimal performance design of the NCS.

Algorithm 6.6.1 Consider the NCS described by Fig. 6.6 and Eqs. (6.65)–(6.70). Assume that (6.74) is true, and Assumptions 6.3 and 6.4 hold.

(a) Choose L_h and L_y according to (6.88) in Theorem 6.5 with $L_h = K_2\mu_\xi^{-1}$ and $L_y = (1-L_h\mu_\xi)^{-1}K_1$ such that β_1 is minimized.
(b) Choose L_r according to

$$L_r = \arg\inf_{L_r\in\mathscr{R}\mathscr{H}_\infty}\|1-G(1-L_yG)^{-1}L_r\|_2^2.$$

(c) Choose L_v, F and α based on Theorem 6.6.

6.6.3 Numerical Example

The following numerical example illustrates the results of this section.

Example 6.2 Consider the NCS in Fig. 6.6 with channel model (6.70), where $G(z) = 1/(z-2)$, and $\sigma_r^2 = \sigma_d^2 = \sigma_n^2 = 1$. Suppose that Assumptions 6.3 and 6.4 hold. Based on Theorem 6.5, we can obtain the stabilizability regions as shown in Fig. 6.9. After applying Algorithm 6.6.1 with $\mu_\xi = 2$ and $\sigma_\xi^2 = 0.5$, the relationship between J_{opt} and $\mathscr{P}$ is given in Fig. 6.10.

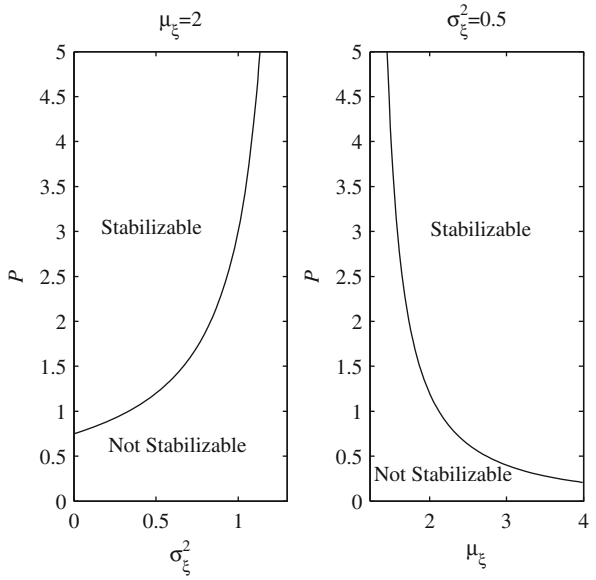

Fig. 6.9 The stabilizability regions on the $\mathscr{P}$-σ_ξ^2 plane with $\mu_\xi = 2$ (*left*) and the $\mathscr{P}$-μ_ξ plane with $\sigma_\xi^2 = 0.5$ (*right*)

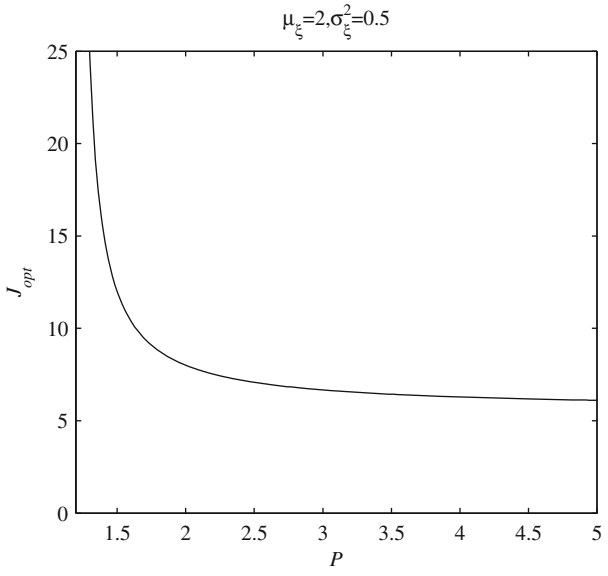

Fig. 6.10 The relationship between J_{opt} and $\mathscr{P}$ using Algorithm 6.6.1

6.7 Summary

It has been shown in this chapter that for a discrete-time LTI plant with LTI feedback over input fading channels under PTS and under the assumption on capacity allocation, the minimal overall mean square capacity of the network for mean square stabilizability can be given in terms of the Mahler measure of the plant in the case of state feedback. The minimal capacity for stabilizability via state feedback under STS in general can only be computed by optimization. We have also investigated the mean square stabilization of SISO plants or MIMO plants via dynamic output feedback. For SISO plants, the corresponding minimal mean square capacity for stabilizability is given in terms of the anti-stable poles, NMP zeros and relative degree of the plant. For triangularly decoupled MIMO plants, necessary and sufficient conditions have been provided. In addition, the results have been extended to the case with output fading channels and applied to vehicle platooning. It also has been shown that the channel feedback plays a key role in eliminating the limitation on stabilization induced by the NMP zeros and high relative degree of the plant. For an SISO LTI plant with LTI feedback over a fading channel with channel feedback, the minimal requirement on the network for mean square stabilizability can be exactly characterized in terms of the statistics of channel fading and channel additive noise, the power bound of channel input, and the Mahler measure of the plant. When a power bound is applied on the channel input, the network requirement for stabilizability becomes more stringent, and the scaling of the network is an essential parameter in satisfying the power constraint. Under the assumption on perfect reconstruction of the network, a suboptimal method has been given for performance design of the NCS.

The results in this chapter are based mainly on [25, 29, 30]. As a special case of this chapter, the minimum capacity for state feedback stabilization of a single-input plant over a single fading channel is given in [1]. In [31], a unitary matrix is introduced to encode and decode the control signal at the two ends of fading channels based on which a resource allocation method is developed to mitigate the channel fading.

References

1. N. Elia, Remote stabilization over fading channels. Syst. Control Lett. **54**(3), 237–249 (2005)
2. D. Tse, P. Viswanath, *Fundamentals of Wireless Communication* (Cambridge University Press, Cambridge, 2005)
3. S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory* (Society for Industrial Mathematics, Philadelphia, 1994)
4. W. Wonham, On pole assignment in multi-input controllable linear systems. IEEE Trans. Autom. Control **12**(6), 660–665 (1967)
5. G. Gu, L. Qiu, Networked stabilization of multi-input systems with channel resource allocation, in *Proceedings of the 17th IFAC World Congress*, pp. 625–630 (2008)
6. S. Hu, W. Yan, Stability of networked control systems under a multiple-packet transmission policy. IEEE Trans. Autom. Control **53**(7), 1706–1711 (2008)

7. Y. Li, E. Tuncel, J. Chen, W. Su, Optimal tracking performance of discrete-time systems over an additive white noise channel, in *Proceedings of the 48th IEEE Conference on Decision and Control*, pp. 2070–2075 (2009)
8. S. Hu, W. Yan, Stability robustness of networked control systems with respect to packet loss. Automatica **43**(7), 1243–1248 (2007)
9. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
10. D. Tse, Optimal power allocation over parallel Gaussian broadcast channels, in *Proceedings of IEEE International Symposium on Information Theory*, p. 27 (1997)
11. L. Li, A. Goldsmith, Capacity and optimal resource allocation for fading broadcast channels—I: Ergodic capacity. IEEE Trans. Inf. Theory **47**(3), 1083–1102 (2001)
12. M. Tabbara, A. Rantzer, D. Nesic, On controller and capacity allocation co-design for networked control systems. Syst. Control Lett. **58**(9), 672–676 (2009)
13. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)
14. L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks. Proc. IEEE **95**(1), 163–187 (2007)
15. K. Zhou, J. Doyle, *Essentials of Robust Control* (Prentice Hall, New Jersey, 1998)
16. O. Toker, J. Chen, L. Qiu, Tracking performance limitations in LTI multivariable discrete-time systems. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **49**(5), 657–670 (2002)
17. J. Braslavsky, R. Middleton, J. Freudenberg, Feedback stabilization over signal-to-noise ratio constrained channels. IEEE Trans. Autom. Control **52**(8), 1391–1403 (2007)
18. B. Francis, *A Course in H_∞ Control Theory* (Springer, New York, 1987)
19. A. Rojas, Comments on feedback stabilization over signal-to-noise ratio constrained channels. IEEE Trans. Autom. Control **54**(6), 1425–1426 (2009)
20. G. Gomez, G. Goodwin, An algebraic approach to decoupling in linear multivariable systems. Int. J. Control **73**(7), 582–599 (2000)
21. P. Seiler, A. Pant, K. Hedrick, Disturbance propagation in vehicle strings. IEEE Trans. Autom. Control **49**(10), 1835–1842 (2004)
22. R. Middleton, J. Braslavsky, String instability in classes of linear time invariant formation control with limited communication range. IEEE Trans. Autom. Control **55**(7), 1519–1530 (2010)
23. A. Goldsmith, *Wireless Communications* (Cambridge University Press, Cambridge, 2005)
24. S. Dey, A. Leong, J. Evans, Kalman filtering with faded measurements. Automatica **45**(10), 2223–2233 (2009)
25. N. Xiao, L. Xie, L. Qiu, Mean square stabilization of multi-input systems over stochastic multiplicative channels, in *Proceedings of the 48th IEEE Conference on Decision and Control*, pp. 6893–6898 (2009)
26. E. Silva, A unified framework for the analysis and design of networked control systems, PhD thesis, Callaghan, Australia: The University of Newcastle (2009)
27. M. Derpich, E. Silva, D. Quevedo, G. Goodwin, On optimal perfect reconstruction feedback quantizers. IEEE Trans. Signal Process. **56**(8), 3871–3890 (2008)
28. W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1987)
29. N. Xiao, L. Xie, L. Qiu, Feedback stabilization of discrete-time networked systems over fading channels. IEEE Trans. Autom. Control **57**(9), 2176–2189 (2012)
30. N. Xiao, L. Xie, Analysis and design of discrete-time networked systems over fading channels, in *Proceedings of the 30th Chinese Control Conference*, pp. 6562–6567 (2011)
31. G. Gu, L. Qiu, Networked feedback control over fading channels and the relation to H2 control, in *Proceedings of the International Conference on Information and Automation*, pp. 247–252 (2012)

Chapter 7
Stabilization of Linear Systems via Infinite-Level Logarithmic Quantization

This chapter studies a number of quantized feedback design problems for linear systems. We consider the case where quantizers are static (memoryless). The common aim of these design problems is to stabilize the given system or to achieve certain performance with the coarsest quantization density. The main result is that the classical sector bound approach is non-conservative for studying these design problems. Consequently, many quantized feedback design problems are converted to well-known robust control problems with sector bound uncertainties.

The chapter is organized as follows. In Sect. 7.1, we first review the key result in [1] which is on quadratic stabilization of SISO linear systems using quantized state feedback. We show that coarsest quantization density can be simply obtained using the sector bound method. This not only gives a simpler interpretation of their result, but also provides the basis for its generalization. Further, the coarsest quantization density is directly related to a H_∞ optimization problem, which is better than relating it to an "expensive" control problem as in [1]. In Sect. 7.2, two cases of the output feedback stabilization of SISO systems are considered: observer-based quantized state feedback and dynamic feedback using quantized output. We show that the coarsest quantization density in the former case is the same as in quantized state feedback, whereas the latter case is related to a different H_∞ optimization problem and in general requires a finer quantization density.

In Sect. 7.3, the quadratic stabilization problem is generalized to MIMO systems. It is shown that quadratic stabilization with a set of logarithmic quantizers is the same as quadratic stabilization for an associated system with sector-bounded uncertainty. Because the latter problem has been well studied, the technical difficulty for the first problem is clearly revealed. A sufficient condition is then given, in terms of an H_∞ optimization problem, for the quantizers to render a quadratic stabilizer. As in the SISO case, both state feedback and output feedback are considered.

Finally, we discuss the results on performance control problems. Both linear quadratic performance (Sect. 7.4) and H_∞ performance (Sect. 7.5) problems are studied and conditions are given for a set of quantizers to render a given performance level. Section 7.6 summarizes the chapter.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_7

7.1 State Feedback Case

In this section, we revisit the work of Elia and Mitter [1] on stabilization using quantized state feedback and reinterpret their result using the sector bound method.

The simplest and most fundamental case considered in [1] is the problem of quadratic stabilization for the following system:

$$x_{k+1} = Ax_k + Bu_k \tag{7.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, x is the state and u is a quantized state feedback in the following form:

$$u_k = Q^\infty(v_k), \tag{7.2}$$

$$v_k = Kx_k. \tag{7.3}$$

In the above, $K \in \mathbb{R}^{1\times n}$ is the feedback gain, and $Q^\infty(\cdot)$ is a quantizer which is assumed to be symmetric, i.e., $Q^\infty(-v) = -Q^\infty(v)$. Note that the quantizer is static and time-invariant.

7.1.1 Logarithmic Quantization

The set of (distinct) quantized levels is described by

$$\mathscr{U} = \{\pm u_i, i = 0, \pm 1, \pm 2, \ldots\} \cup \{0\} \tag{7.4}$$

Each of the quantization levels (say u_i) corresponds to a segment (say V_i) such that the quantizer maps the whole segment to this quantization level. In addition, these segments form a partition of $\mathbb{R}$, i.e., they are disjoint and their union equals to $\mathbb{R}$.

Denote by $\#g[\varepsilon]$ the number of quantization levels in the interval $[\varepsilon,\ 1/\varepsilon]$. The density of the quantizer $Q^\infty(\cdot)$ is defined as follows:

$$\eta_Q = \limsup_{\varepsilon\to 0} \frac{\#g[\varepsilon]}{-\ln \varepsilon} \tag{7.5}$$

With this definition, the number of quantization levels of a quantizer with a nonzero, finite quantization density grows logarithmically as the interval $[\varepsilon,\ 1/\varepsilon]$ increases. A small η_Q corresponds to a coarse quantizer. A finite quantizer (i.e., a quantizer with a finite number of quantization levels) has $\eta_Q = 0$, and a linear quantizer has $\eta_Q = \infty$.

A quantizer is called *logarithmic* if it has the form:

$$\begin{aligned} \mathscr{U} = \{&\pm u^{(i)} : u^{(i)} = \rho^i u^{(0)}, i = \pm 1, \pm 2, \ldots\}, \\ &\cup\{\pm u^{(0)}\} \cup \{0\}, \quad 0 < \rho < 1, \qquad u^{(0)} > 0. \end{aligned} \tag{7.6}$$

The associated quantizer Q^∞ is defined as follows:

$$Q^\infty(v) = \begin{cases} u_i, & \text{if } \frac{1}{1+\delta}u_i < v \le \frac{1}{1-\delta}u_i, \quad v > 0 \\ 0, & \text{if } v = 0 \\ -Q^\infty(-v), & \text{if } v < 0. \end{cases} \tag{7.7}$$

where

$$\delta = \frac{1-\rho}{1+\rho}. \tag{7.8}$$

It is easily verified that $\eta_Q = 2/\ln(1/\rho)$ for the logarithmic quantizer. This means that the smaller the ρ, the smaller the η_Q. For this reason, we will abuse the terminology by calling ρ (instead of η_Q) the *quantization density* in the rest of the chapter. The logarithmic quantizer is illustrated in Fig. 7.1. In contrast, a non-logarithmic quantizer is illustrated in Fig. 7.2.

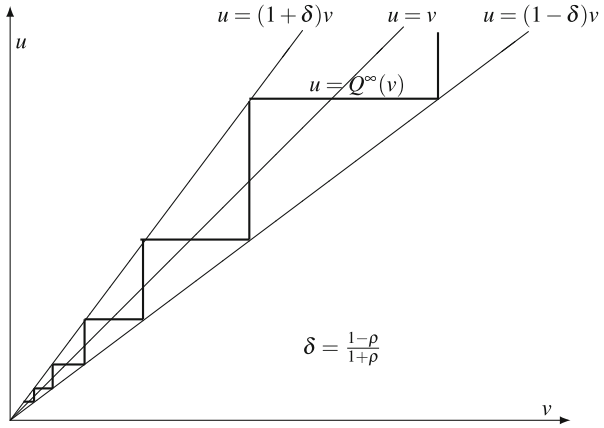

Fig. 7.1 Logarithmic quantizer

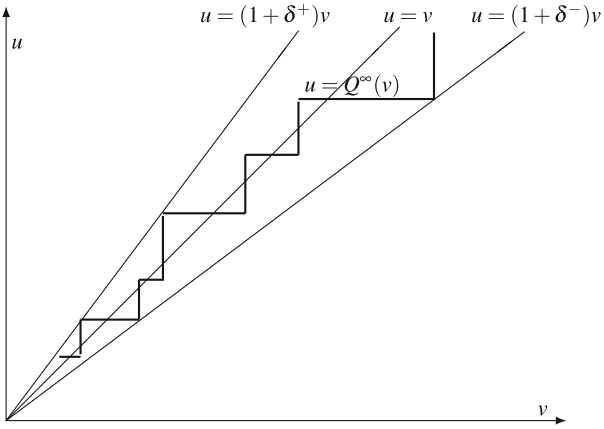

Fig. 7.2 Non-logarithmic quantizer

For the quadratic stabilization problem, a quadratic Lyapunov function $V(x) = x^T Px, P = P^T > 0$, is used to assess the stability of the feedback system. That is, the quantizer must satisfy

$$\nabla V(x) = V(Ax + BQ^{\infty}(Kx)) - V(x) < 0, \quad \forall x \neq 0. \tag{7.9}$$

The coarsest quantizer is the one which minimizes η_Q subject to (7.9). But the coarsest quantizer is in general not attainable because the constraint in (7.9) is a strict inequality.

The required density of the quantizer depends on $V(x)$ (or P) and K. This raises the key question: What is the coarsest density, ρ_{inf}, among all possible P and K? In [1], under the assumption that

$$K = K_{GD} = -\frac{B^T PA}{B^T PB} \tag{7.10}$$

the answer for ρ_{inf} is given as

$$\rho_{\text{inf}} = \frac{\prod_i |\lambda_i^u| - 1}{\prod_i |\lambda_i^u| + 1}, \tag{7.11}$$

where λ_i^u are the unstable eigenvalues of A.

We see from Figs. 7.1 and 7.2 that a quantizer can be bounded by a sector. For a logarithmic quantizer, the sector bound is described by a single parameter δ which is related to the quantization density by (7.8). In contrast, for a non-logarithmic quantizer, two parameters, δ^- and δ^+, are needed to describe the sector in general. For both finite quantizers and linear quantizers, a default output value, u_0, is needed when the input is smaller than some minimal threshold (in magnitude). If $u_0 = 0$, then $\delta^- = -1$; otherwise, $\delta^+ = \infty$.

7.1.2 Sector Bound Approach

In the following, we use the sector bound method to establish three results: (1) Given any quantizer, the quantized state feedback stabilization problem above is equivalent to a state feedback quadratic stabilization problem with an appropriately defined sector bound uncertainty; (2) The optimal quantizer structure is logarithmic; (3) For a logarithmic quantizer, the quadratic stabilization problem with the sector bound uncertainty has a simple explicit solution which leads to (7.11). It turns out that the result for ρ_{inf} remains the same even when K is without the constraint (7.10). These results are given below:

Theorem 7.1 *Consider the linear system in* (7.1) *and the quantized state feedback* (7.2) and (7.3). *Given a quantizer with a sector bound* $[\delta^-, \delta^+]$, *the system* (7.1) *is quadratically stabilizable via quantized state feedback if and only if the following uncertain system:*

$$x_{k+1} = Ax_k + B(1 + \Delta)v_k, \quad \Delta \in [\delta^-,\ \delta^+] \tag{7.12}$$

is quadratically stabilizable via (7.3)*. If the quantizer is logarithmic with density* ρ*, then the largest sector bound for* (7.12) *to be quadratically stabilizable is given by*

$$\delta_{\sup} = \frac{1}{\prod_i |\lambda_i^u|}. \tag{7.13}$$

Consequently, $\rho_{\inf}$ *is given by* (7.11)*. Finally, the logarithmic quantizer with* $\delta_{\sup}$ (*or* $\rho_{\inf}$) *is the coarsest among all quantizers for quadratically stabiling the system* (7.1) *via quantized state feedback.*

Remark 7.1 For any quantizer, there exists a sector bound for its quantization errors; see Figs. 7.1 and 7.2. The last part of Theorem 7.1 means that this sector bound must be within the largest sector bound, $\delta_{\sup}$, for the quantized feedback system to be quadratically stable. Then, within the maximum sector bound, it is not difficult to see from Figs. 7.1 and 7.2 that the logarithmic quantizer with $\rho_{\inf}$ has the coarsest quantization density possible.

Three lemmas are needed for the proof of Theorem 7.1.

Lemma 7.1 *Given a constant vector* $K \in \mathbb{R}^{1\times n}$*, a constant matrix* $\Omega_0 \in \mathbb{R}^{n\times n}$*, a vector function* $\Omega_1(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times 1}$*, scalars* $\delta^- \le \delta^+$*, and a scalar function*

$$\Delta(\cdot) : \mathbb{R} \to [\delta^-,\ \delta^+]$$

with the following property. For any $\Delta_0 \in [\delta^-,\ \delta^+]$*, there exists* $v_0 \neq 0$ *such that* $\Delta(v_0) = \Delta_0$*. Define the following matrix function:*

$$\Omega(\cdot) = \Omega_0 + \Omega_1(\cdot)K + K^T\Omega_1^T(\cdot), \tag{7.14}$$

it holds that

$$x^T\Omega(\Delta(Kx))x < 0, \quad \forall x \neq 0, x \in \mathbb{R}^n \tag{7.15}$$

if and only if

$$\Omega(\Delta) < 0, \quad \forall \Delta \in [\delta^-,\ \delta^+]. \tag{7.16}$$

Proof It is obvious that (7.16) implies (7.15). To see the converse, we assume (7.15) holds but (7.16) fails. Then, there exist some $x_0 \neq 0$ and $\Delta_0 \in [\delta^-,\ \delta^+]$ such that

$$x_0^T\Omega(\Delta_0)x_0 \ge 0. \tag{7.17}$$

We claim that $Kx_0 \neq 0$. Indeed, if $Kx_0 = 0$, then

$$x_0^T\Omega(\Delta(Kx_0))x_0 = x_0^T\Omega_0x_0 = x_0^T\Omega(\Delta_0)x_0 \ge 0 \tag{7.18}$$

by (7.14) and (7.17), which contradicts (7.15). So, $Kx_0 \neq 0$. Because of the property of $\Delta(\cdot)$, there exists a scalar $\alpha \neq 0$ such that $\Delta(\alpha K x_0) = \Delta_0$. Define $x_1 = \alpha x_0 \neq 0$. Then,

$$x_1^T \Omega(\Delta(Kx_1))x_1 = \alpha^2 x_0^T \Omega(\Delta_0)x_0 \geq 0$$

which violates (7.15). Hence, (7.15) implies (7.16). □

Lemma 7.2 *Consider the uncertain system in* (7.12)*. Define*

$$G_c(z) = K(zI - A - BK)^{-1}B. \tag{7.19}$$

Then, the supreme of δ *for which quadratic stabilization is achievable is given by*

$$\delta_{\sup} = \frac{1}{\inf_K \|G_c(z)\|_\infty}. \tag{7.20}$$

Proof It follows from [2–4] that the quadratic stabilization for (7.12) is achievable if and only if

$$\delta < \frac{1}{\inf_K \|G_c(z)\|_\infty}.$$

Taking the limit in the above yields (7.20). □

Lemma 7.3 *The solution to* (7.20) *is given by* (7.13)*.*

Proof The key to solving (7.20) is that the optimal solution is such that $G_c(z)$ is either an all-pass stable function or approaching to such a one.

Without loss of generality, we take (A, B) to be of the form:

$$A = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \tag{7.21}$$

where $A_s \in \mathbb{R}^{n_1 \times n_1}$ has all its eigenvalues inside the unit disk and $A_u \in \mathbb{R}^{n_2 \times n_2}$ has all its eigenvalues either on or outside the unit circle, and (A_u, B_2) is of a controllable canonical form.

We first claim the following.

Claim 1 *Suppose A is unstable and there exists a K such that A+BK is stable and*

$$\gamma > \|K(zI - A - BK)^{-1}B\|_\infty. \tag{7.22}$$

Then, $\gamma > \det(A_u) = \Pi_i |\lambda_i^u|$.

To prove the claim, we first note that $A+BK$ is stable and (7.22) holds if and only if [5]

$$I - \gamma^{-2} B^T P B > 0;$$

$$(A+BK)^T P(A+BK) - P + \gamma^{-2}(A+BK)^T PB \cdot (I-\gamma^{-2}B^T PB)^{-1}B^T P(A+BK) + K^T K < 0$$

for some $P = P^T > 0$. The two inequalities above are equivalent to

$$P^{-1} - \gamma^{-2}BB^T > 0;$$

$$(A+BK)^T(P^{-1} - \gamma^{-2}BB^T)^{-1}(A+BK) - P + K^T K < 0.$$

Defining $Q = (P^{-1} - \gamma^{-2}BB^T)^{-1}$, the last two inequalities above become $Q > 0$ and

$$(A+BK)^T Q(A+BK) - P + K^T K < 0. \tag{7.23}$$

Denoting

$$\Phi = (B^T QB + I)^{-1}B^T QA + K; \quad \Pi = \Phi^T(B^T QB + I)\Phi \geq 0,$$

the inequality (7.23) can be rewritten as

$$A^T QA - A^T QB(B^T QB + I)^{-1}B^T QA - P + \Pi < 0.$$

It follows that

$$A^T QA - A^T QB(B^T QB + I)^{-1}B^T QA - P < 0,$$

which is equivalent to

$$A^T(Q^{-1} + BB^T)^{-1}A - P < 0.$$

Using Schur complement, the above is equivalent to

$$AP^{-1}A^T - (Q^{-1} + BB^T) < 0.$$

Using $Q = P^{-1} - \gamma^{-2}BB^T$ and defining $X = P^{-1}$, we get $X > \gamma^{-2}BB^T$ and

$$AXA^T < X + (1-\gamma^{-2})BB^T. \tag{7.24}$$

Note that if A is stable, the above inequality exists a solution $X = X^T > \gamma^{-2}BB^T$ for any $\gamma > 0$.

Now, let X be partitioned in conformity with (7.21):

$$X = \begin{bmatrix} X_1 & X_{12} \\ X_{12}^T & X_2 \end{bmatrix}.$$

Then, (7.24) with $X > \gamma^{-2}BB^T$ implies

$$A_u X_2 A_u^T - X_2 - (1-\gamma^{-2})B_2 B_2^T < 0 \tag{7.25}$$

and $X_2 > \gamma^{-2}B_2 B_2^T$.

Now, using the fact that for any two symmetric matrices U and V with $0 \leq U < V$, $\det(U) < \det(V)$, then (7.25) leads to

$$\det(A_u X_2 A_u^T) < \det(X_2 + (1-\gamma^{-2})B_2 B_2^T).$$

Since $B_2 = [0\ 0\ \ldots\ 1]^T$, it follows that

$$|\det(A_u)|^2 \det(X_2) < \det(X_2) + (1-\gamma^{-2}) \det(X_{2,(n_2-1)}) \tag{7.26}$$

where $X_{2,(n_2-1)}$ is the upper left $(n_2-1) \times (n_2-1)$ block of X_2.

Also, since $X_2 > \gamma^{-2} B_2 B_2^T$, we have

$$\det(X_2) - \gamma^{-2} \det(X_{2,(n_2-1)}) > 0$$

or equivalently,

$$\det(X_{2,(n_2-1)}) < \gamma^2 \det(X).$$

Substituting the above into (7.26) and noting that $\det(X_2) > 0$, we have

$$|\det(A_u)|^2 < 1 + (1-\gamma^{-2})\gamma^2 = \gamma^2.$$

Thus, we have verified Claim 1.

We now show that the solution to (7.20) is indeed given by (7.13). To this end, we consider three cases.

Case 1: $a(z)$ is strictly anti-stable. In this case, simply take (A, B) be of a controllable canonical form, it is clear that

$$G_c(z) = \frac{k(z)}{a(z) - k(z)} \tag{7.27}$$

where

$$a(z) = a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n = |zI - A|$$

and

$$k(z) = k_0 + k_1 z + \cdots + k_{n-1} z^{n-1}$$

is the control polynomial.

We claim that choosing

$$K = \frac{a_0^2}{a_0^2 - 1} \left[a_0 - \frac{1}{a_0}, a_1 - \frac{a_{n-1}}{a_0}, \ldots, a_{n-1} - \frac{a_1}{a_0} \right] \tag{7.28}$$

leads to

$$\|G_c(z)\|_\infty = |a_0| = \prod_i |\lambda_i^u| > 1.$$

This together with Claim 1 implies that the solution in (7.28) is the optimal solution. The second claim above holds because (7.28) comes from solving the all-pass requirement for $G_c(z)$:

$$a(z) - k(z) = \alpha z^n k(z^{-1}) \tag{7.29}$$

for some α. Replacing z by z^{-1}, (7.29) becomes

$$a(z^{-1}) - k(z^{-1}) = \alpha z^{-n} k(z). \tag{7.30}$$

Combining (7.29) and (7.30) yields

$$k(z) = \frac{a(z) - \alpha z^n a(z^{-1})}{1 - \alpha^2}. \tag{7.31}$$

Setting the nth order coefficient of $k(z)$ to zero results in $\alpha = 1/a_0$. It is straightforward to verify that (7.31) is the same as (7.28). It remains to show that $G_c(z)$ is stable, which is the same as showing that $k(z)$ is strictly anti-stable. To see this, we rewrite (7.31) as

$$k(z) = \frac{a(z)}{1 - \alpha^2}\left(1 - \alpha \frac{z^n a(z^{-1})}{a(z)}\right).$$

Because $a(z)$ is antistable, $|\alpha| < 1$ and $|z^n a(z^{-1})/a(z)| \leq 1$ for any $|z| \leq 1, k(z) \neq 0$ for any $|z| \leq 1$. Hence, $k(z)$ is strictly anti-stable.

Case 2: $a(z)$ is marginally anti-stable. In this case, we first replace $a(z)$ by $\tilde{a}(z)$ which is a strictly anti-stable polynomial obtained from $a(z)$ by slightly perturbing the marginal zeros. From Case 1, we can choose $k(z)$ such that $\|\tilde{G}_c(z)\|_\infty = |\tilde{a}_0|$, where $\tilde{a}_0$ and $\tilde{G}_c(z)$ are the perturbed versions of a_0 and $G_c(z)$. By the continuity of $\tilde{a}_0$ (with respect to the perturbation), it is clear that $\inf_K \|G_c(z)\| = |a_0|$.

Case 3: $a(z)$ has a stable factor. In this case, we can write $a(z) = a_s(z)a_u(z)$, where $a_s(z)$ and $a_u(z)$ are the stable and unstable factors. Taking $k(z) = a_s(z)k_u(z)$ yields

$$G_c(z) = \frac{k_u(z)}{a_u(z) - k_u(z)}.$$

Then, we have reverted to Case 2. Again, we obtain

$$\inf_K \|G_c(z)\|_\infty = \prod_i |\lambda_i^u|.$$

This completes the proof. □

Proof of Theorem 7.1 Define the quantization error by

$$e = u - v = Q^\infty(v) - v = \Delta(v)v. \tag{7.32}$$

Then,

$$\Delta(v) \in [\delta^-,\ \delta^+]. \tag{7.33}$$

We can model the quantized feedback system as the following uncertain system:

$$x_{k+1} = Ax_k + B(1 + \Delta(Kx))Kx_k \tag{7.34}$$

and the corresponding quadratic stabilization condition becomes

$$\nabla V(x) = V((A + B(1 + \Delta(Kx))K)x) - V(x) < 0, \quad \forall x \neq 0. \tag{7.35}$$

Define

$$\begin{aligned} \nabla P(\Delta) &= (A + B(1 + \Delta)K)^T P(A + B(1 + \Delta)K) - P \\ &< 0, \quad \forall \Delta \in [\delta^-,\ \delta^+] \end{aligned} \tag{7.36}$$

where Δ is independent of the state. Note that the inverse mapping of $\Delta(v)$ in (7.32) is a multi-branch continuous function (except at $v = 0$). Hence, for any $\Delta_0 \in [\delta^-,\ \delta^+]$, there exists some $v_0 \neq 0$ such that $\Delta(v_0) = \Delta_0$. By Lemma 7.1, (7.35) is equivalent to (7.36). But the latter is the condition for the system (7.12) to be quadratically stabilizable.

For a logarithmic quantizer where $-\delta^- = \delta^+ = \delta$, the above means that the problem of coarsest quantization is equivalent to finding the maximum δ for (7.12) to be quadratically stabilizable. By Lemmas 7.2 and 7.3, the solution to the latter is given by (7.13). Hence, the solution to $\rho_{\inf}$ is given by (7.11).

Now consider a non-logarithmic quantizer. Define a scaled input $\hat{v} = \alpha v$, $\alpha \neq 0$, and interpret the quantizer $Q^\infty(v)$ as $\hat{f}(\hat{v})$. This results in $\hat{\Delta}(\hat{v})$, $\hat{\delta}^-$ and $\hat{\delta}^+$. If $\delta^- > -1$ and $\delta^+ < \infty$, we can always find α such that $-\hat{\delta}^- = \hat{\delta}^+$. Such a scaling does not change the quadratic stabilizability. It is clear from the analysis above that the system (7.1) cannot be quadratically stabilized via quantized state feedback if $\hat{\delta}^+ > \delta_{\sup}$. If $\delta^- = -1$ or $\delta^+ = \infty$, then α can always be found to make

$$[-\delta_{\sup},\ \delta_{\sup}] \subset [\hat{\delta}^-,\ \hat{\delta}^+].$$

Again, the system (7.1) cannot be quadratically stabilized via quantized state feedback. Hence, we conclude that a logarithmic quantizer with $\rho_{\sup}$ (or $\rho_{\inf}$) is the coreset possible. □

Remark 7.2 It is shown in [1] that the coarsest quantization density is related to the solution to the so-called "expensive" linear quadratic control problem:

$$\begin{array}{rl} \min_K & \sum_{k=0}^{\infty} |u_k|^2 \\ \text{subject to} & \text{closed-loop stability with} \\ & u_k = Kx_k. \end{array} \tag{7.37}$$

More specifically, the optimal ρ can be solved using the solution to the Riccati equation for the "expensive" control problem. However, the optimal control gain K for the quantization problem is different from the optimal control gain for the

"expensive" control problem (This is also pointed out in [1]). From the proofs above, we see that it is better to interpret the coarsest quantization problem as an H_∞ optimization problem (7.20).

Remark 7.3 We have seen that logarithmic quantizers are essential for quadratic stabilization via quantized feedback if a coarse quantization density is required. Non-logarithmic quantizers such as finite quantizers and linear quantizers are unsuitable. For this reason, we will consider logarithmic quantizers only in the rest of this chapter.

7.2 Output Feedback Case

We now show how to generalize the technique for state feedback to quantized output feedback. Consider the following system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k \end{aligned} \tag{7.38}$$

where A and B are the same as before and $C \in \mathbb{R}^{1\times n}$.

We consider two possible basic configurations for quantized output feedback which may lead to other more complicated settings.

Configuration I: The control signal is quantized but the measurement is not;
Configuration II: The measurement is quantized but the control signal is not.

It turns out that they result in different quantization density requirements.

7.2.1 Quantized Control

Configuration I. This is an easy case which has an interesting result below.

Theorem 7.2 *Consider the system* (7.38) *with quantized control input. Suppose* (A, C) *is an observable pair. The coarsest quantization density for quadratic stabilization by state feedback can also be achieved by output feedback. In particular, the corresponding output feedback controller can be chosen as an observer-based controller below:*

$$\begin{aligned} x_{c,k+1} &= Ax_{c,k} + Bu_k + L(y_k - Cx_{c,k}) \\ v_k &= Kx_{c,k} \\ u_k &= Q^\infty(v_k) \end{aligned} \tag{7.39}$$

where $Q^\infty(\cdot)$ *is the quantizer as before,* K *is the state feedback gain designed for any achievable quantization density via quantized state feedback, and* L *is any gain which yields* (7.39) *a deadbeat observer.*

Proof Let K be any state feedback gain that achieves any given quantization density. Choose L such that the observer is deadbeat, i.e., $e_k = x_k - x_{c,k} \neq 0$ only for a finite number of steps N. This can be always done because (A, C) is observable. Then, after N steps, the output feedback controller is the same as state feedback controller. Hence, the system is quadratically stabilized after N steps. Finally, it is a simple fact (although we do not give the details) that if a (nonlinear) system is quadratically stable after N steps and that the state is bounded in the first N steps (which clearly holds for the system (7.39)), it is quadratically stable. □

7.2.2 Quantized Measurements

Configuration II. In this case, the controller is in the form

$$\begin{aligned} x_{c,k+1} &= A_c x_{c,k} + B_c Q^\infty(y_k) \\ u_k &= C_c x_{c,k} + D_c Q^\infty(y_k) \end{aligned} \tag{7.40}$$

where $Q^\infty(\cdot)$ is the quantizer as before.

It is straightforward to verify that the closed-loop system is given by

$$\bar{x}_{k+1} = \mathscr{A}(\Delta(y_k)\bar{x}_k \tag{7.41}$$

where $\bar{x} = [x^T, x_c^T]^T$, $\Delta(\cdot)$ is the same as in (7.33) and

$$\begin{aligned} \bar{A} &= \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 0 & B \\ I & 0 \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0 & I \\ C & 0 \end{bmatrix} \\ \hat{I} &= \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} C & 0 \end{bmatrix}, \quad \bar{K} = \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} \end{aligned} \tag{7.42}$$

and

$$\mathscr{A}(\Delta) = \bar{A} + \bar{B}\bar{K}(\bar{C} + \hat{I}\Delta\hat{C}). \tag{7.43}$$

The problem of concern is to find the coarsest quantizer for quadratic stabilization of the closed-loop system. This can be solved by generalizing the idea for the state feedback case. The result is given below.

Theorem 7.3 *Consider the system* (7.38). *For a given quantization density* $\rho > 0$, *the system is quadratically stabilizable via a quantized controller* (7.40) *if and only if the following auxiliary system:*

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ v_k &= (1 + \Delta)Cx_k, \quad |\Delta| \leq \delta \end{aligned} \tag{7.44}$$

is quadratically stabilizable via:

$$\begin{aligned} x_{c,k+1} &= A_c x_{c,k} + B_c v_k \\ u_k &= C_c x_{c,k} + D_c v_k \end{aligned} \tag{7.45}$$

where δ and ρ are related by (7.8).

The largest sector bound $\delta_{\sup}$ (which gives $\rho_{\inf}$) is given by

$$\delta_{\sup} = \frac{1}{\inf_{\bar{K}} \|\bar{G}_c(z)\|_\infty} \tag{7.46}$$

where $\bar{K}$ is defined in (7.42) *and*

$$\bar{G}_c(z) = (1 - H(z)G(z))^{-1} H(z)G(z) \tag{7.47}$$

where $G(z) = C(zI - A)^{-1}B$ and $H(z) = D_c + C_c(zI - A_c)^{-1}B_c$.

Further, if $G(z)$ has relative degree equal to one and no unstable zeros, then the coarsest quantization density for quantized state feedback can be reached via quantized output feedback.

Proof The proof is similar to the proof of Theorem 7.1. The sector bound for the quantization error is done as in (7.32) and (7.33). For the given ρ, the quadratic stability of the closed-loop system (7.38)–(7.40) requires the existence of some $\bar{P} = \bar{P}^T > 0$ such that

$$\bar{x}^T [\mathscr{A}(\Delta(y))^T \bar{P} \mathscr{A}(\Delta(y)) - \bar{P}] \bar{x} < 0 \tag{7.48}$$

for all $\bar{x} \neq 0$ and $y = Cx = \hat{C}\bar{x}$. Using Lemma 7.1, the above is equivalent to

$$\mathscr{A}(\Delta)^T \bar{P} \mathscr{A}(\Delta) - \bar{P} < 0, \quad \forall |\Delta| \leq \delta.$$

The latter is the same as requiring the system (7.44) and (7.45) to be quadratically stable. Since the transfer function of (7.44) is $G(z)(1 + \Delta)$ and that of (7.45) is $H(z)$, the closed loop system (7.44) and (7.45) is the same as a closed-loop system with the open-loop block equal to

$$\bar{G}_c(z) = (1 - H(z)G(z))^{-1} H(z)G(z)$$

and feedback block equal to Δ. It follows that the solution to $\delta_{\sup}$ comes from the equivalence between quadratic stability and H_∞ optimization [2, 4].

Suppose $G(z)$ has relative degree one and no unstable zeros. Write $G(z) = b(z)/a(z)$. From the proof of Theorem 7.1, we know that the state feedback case corresponds to H_∞ optimization of $G_c(z)$ in (7.27). If we choose $H(z) = k(z)/b(z)$. Then, $\bar{G}_c(z)$ in (7.47) becomes $G_c(z)$. Hence, the quantization density for the quantized state feedback can be achieved by quantized output feedback. □

Now we give an example to show that using quantized output requires a higher quantization density than using quantized state feedback.

Example 7.1 The system is given by (7.38) with

$$G(z) = C(zI - A)^{-1}B = (z - 3)/z(z - 2).$$

Using quantized state feedback, $\delta = 2$ and $\rho = (2 - 1)/(2 + 1) = 0.3333$. For quantized output feedback, computing (7.46) yields $\delta = 10$ and

$$\rho = (10 - 1)/(10 + 1) = 0.8182.$$

Remark 7.4 In [1], output feedback control design is done in two steps. In Step 1, coarsest quantization is solved for state estimation, which is a dual problem to the state feedback stabilization problem. In Step 2, the separation principle is applied, i.e., optimal state feedback is combined with optimal state estimation. The main result is that logarithmic quantization is sufficient for output feedback stabilization.

The drawback of the approach in [1] is that the physical meaning of the state estimation quantizer is not clear. Indeed, the problem of quantized state estimation is formulated to be:

$$e_{k+1} = Ae_k + LQ^e(Ce_k) \tag{7.49}$$

where $e_k = x_k - x_{c,k}$ is the state estimation error and $Q^e(\cdot)$ is the state estimation quantizer. What is unsatisfactory in this formulation is that the quantizer needs to know both y_k and its estimate $Cx_{c,k}$. If the control signal is generated at the measurement end, there is obviously no need to use quantized y_k. If the control signal is generated elsewhere using a quantized y_k, it is difficult to imagine why its estimate needs to be sent back to the measurement end to form Ce_k for quantization. Hence, the validity of this formulation seems to be questionable.

7.3 Stabilization of MIMO Systems

In this section, we generalize the quantization results in Sects. 7.1 and 7.2 to MIMO systems with multiple quantizers. For simplicity, the number of quantizers is assumed to be equal to the number of inputs, although this can be easily relaxed. The quantizers are assumed to be static and independent. As in the SISO case, two configurations are treated. **Configuration I** deals with state feedback or observer-based state feedback with quantized inputs, whereas in **Configuration II** quantized outputs are used.

7.3.1 Quantized Control

Configuration I. The system is still as in (7.38) (or (7.1) for state feedback) except that we now allow $u \in \mathbb{R}^m$, $y \in \mathbb{R}^r$. Suppose quantized state feedback (7.2) and (7.3) is used, where $K \in \mathbb{R}^{m\times n}$ and

$$Q^\infty(v) = \text{diag}\{Q^1(v_1), Q^2(v_2), \ldots, Q^m(v_m)\} \tag{7.50}$$

where v_j is the jth component of v and $Q^j(\cdot)$ is a quantizer of the form (7.6) but with quantization level $0 < \rho_j < 1$.

Because we have more than one quantizer, the notion of coarsest quantization is not well-defined. Instead, we ask the following question: Given a vector of quantization levels $\rho = [\rho_1 \ \rho_2 \ \cdots \ \rho_m]$, does there exist an quantized feedback controller that quadratically stabilizes the system (7.38)? The main result is given below:

Theorem 7.4 *Given the system in* (7.38) *and a quantization level vector* ρ*, consider the following auxiliary system:*

$$x_{k+1} = Ax_k + B(I + \Delta_k)v_k \tag{7.51}$$

where $|\Delta_{j,k}| \leq \delta_j$ *for all* $j = 1, 2, \ldots, m$ *and* k*, and* δ_j *are converted from* ρ_j *using* (7.8)*, and* v_k *is a control input. Suppose the auxiliary system is quadratically stablizable via state feedback* (7.3)*, then* (7.38) *is quadratically stabilizable via quantized state feedback. Conversely, suppose the system* (7.38) *is quadratically stabilizable via quantized state feedback and, in addition, suppose* $\ln \rho_i / \ln \rho_j$ *are irrational numbers for all* $i \neq j$ *when* $m > 1$*. Then, for any (arbitrarily small)* $\varepsilon > 0$*, the auxiliary system* (7.51) *with* $|\Delta_{j,k}| \leq \delta_j - \varepsilon$ *is quadratically stabilizable via state feedback* (7.3)*.*

Further, the auxiliary system is quadratically stabilizable via state feedback (7.3) *if the following state feedback* H_∞ *control has a solution* K *for some diagonal scaling matrix* $\Gamma > 0$*:*

$$\|\Lambda\Gamma K(zI - A - BK)^{-1}B\Gamma^{-1}\|_\infty < 1 \tag{7.52}$$

where

$$\Lambda = \text{diag}\{\delta_1, \ldots, \delta_m\}. \tag{7.53}$$

In particular, any K *that renders* (7.52) *is a solution to either quadratic stabilization problem.*

Finally, if (C, A) *is an observable pair and* (7.38) *is quadratically stabilizable via quantized state feedback for the given* ρ*, then it is also quadratically stabilizable via observer-based quantized state feedback* (7.39) *for the same* ρ*.*

Remark 7.5 It is easy to see that if a given ρ does not satisfy the condition that $\ln \rho_i / \ln \rho_j$ are irrational for $i \neq j$, we can make it so by perturbing the ρ_j slightly.

Three technical lemmas are required for the proof of the result above.

Lemma 7.4 *For the quantizer* (7.6) *and any* $|\Delta| \leq \delta$*, the inverse function for* $\Delta(v)$ *is not unique, and is given by*

$$\ln \frac{v}{u^{(0)}} = i \ln \rho - \ln(\Delta + 1), \quad i = 0, \pm 1, \pm 2, \ldots \tag{7.54}$$

Proof The results follow directly from the definition of $\Delta(v)$ in (7.33). □

Lemma 7.5 *Let $Q^j(\cdot), j = 1, 2, \ldots, m$ be a set of quantizers as in (7.6) but with (possibly different) values $u_j^{(0)}$ and $0 < \rho_j < 1$. Suppose the ratios $\ln \rho_i / \ln \rho_j$ are irrational numbers for all $1 \le i, j \le m, i \ne j$ (This condition is void if $m = 1$). Then, given any pairs of vectors (v, Δ^0) with $v_j \ne 0$ and $|\Delta_j^0| \le \delta_j, j = 1, 2, \ldots, m$, and any scalar $\varepsilon > 0$ (arbitrarily small), there exists a scalar $\alpha > 0$ such that*

$$|\Delta_j(\alpha v_j) - \Delta_j^0| < \varepsilon, \quad j = 1, 2, \ldots, m \tag{7.55}$$

where $\Delta_j(\cdot)$ is as defined in (7.32) and (7.33). That is, as α varies from 0 to ∞, the vector $[\Delta_1(\alpha v_1) \ \ldots \ \Delta_m(\alpha v_m)]^T$ covers the hyperrectangle

$$[-\delta_1, \delta_1] \oplus \cdots \oplus [-\delta_m, \delta_m]$$

densely.

Proof Note that each $\Delta_j(v)$ is periodic in $\ln(v/u_j^{(0)})$ with the period $\ln \rho_j$ and that within each period the mapping between $\ln(v/u_j^{(0)})$ and $\Delta_j(v)$ is one-to-one. Therefore, it suffices to show that as α varies,

$$[\mathrm{mod}(\ln \alpha v_1/u_1^{(0)}, \ln \rho_1) \ \cdots \ \mathrm{mod}(\ln \alpha v_m/u_m^{(0)}, \ln \rho_m)]^T$$

covers

$$B = [0, \ln \rho_1] \oplus \cdots \oplus [0, \ln \rho_m]$$

densely. This is equivalent to that

$$\gamma = [\mathrm{mod}(\ln \alpha, \ln \rho_1) \ \cdots \ \mathrm{mod}(\ln \alpha, \ln \rho_m)]^T$$

covers B densely.

Let $\beta = [\beta_1, \ldots, \beta_m]^T \in B$ be any given vector. We need to find α such that γ is arbitrarily close to β. The assumption that $\ln \rho_i / \ln \rho_j$ are irrational implies that quantizers $Q^i(\cdot)$ and $Q^j(\cdot), i \ne j$, do not share a common period (in the logarithmic scale), which is the key to the analysis below. If $m = 1$, we can simply take

$$\ln \alpha = \beta_1 + i_1 \ln \rho_1 \tag{7.56}$$

as a solution with any integer i_1. If $m = 2$, we keep $\ln \alpha$ as in (7.56) but let i_1 vary. Because $Q^1(\cdot)$ and $Q^2(\cdot)$ do not share a common period, as the integer i_1 varies from $-\infty$ to ∞, $\mathrm{mod}(\ln \alpha, \ \ln \rho_2)$ will cover the set $[0, \ln \rho_2]$ densely. Let I_1 and I_2 be the infinite sequences of i_1 and the corresponding i_2, respectively, which make the corresponding set of $\mathrm{mod}(\ln \alpha, \ \ln \rho_2)$ sufficiently close to β_2. For $m = 3$, because $Q^1(\cdot)$, $Q^2(\cdot)$ and $Q^3(\cdot)$ do not share a common period pair-wise, there is an infinite sequence $\tilde{I}_1$ for i_1 (a subsequence of I_1) which generates the corresponding infinite sequence $\tilde{I}_2$ for i_2 (a subsequence of I_2) and infinite sequence I_3 for i_3 such that

$\text{mod}(\ln\alpha,\ \ln\rho_3)]^T$ is also sufficiently close to β_3. This process can continue for $m > 3$. Hence, we have proved the needed result. □

Lemma 7.6 *Let $Q^j(\cdot), j = 1, \ldots, m$, be a set of quantizers satisfying the conditions in Lemma 7.5. Given constant matrices $K \in \mathbb{R}^{m\times n}$ and $\Omega_0 = \Omega_0^T \in \mathbb{R}^{n\times n}$, and a matrix function $\Omega_1(\cdot) : \mathbb{R}^m \to \mathbb{R}^{n\times m}$, define*

$$\Omega(\cdot) = \Omega_0 + \Omega_1(\cdot)K + K^T\Omega_1^T(\cdot). \tag{7.57}$$

Suppose $\Omega(\cdot)$ is strictly convex. Then,

$$x^T\Omega(\Delta(Kx))x < 0, \quad \forall x \neq 0, \qquad x \in \mathbb{R}^n \tag{7.58}$$

if

$$\Omega(\Delta) < 0, \quad \forall|\Delta_j| \leq \delta_j, \qquad j = 1, \ldots, m. \tag{7.59}$$

Conversely, (7.58) implies

$$\Omega(\Delta) < 0, \quad \forall|\Delta_j| \leq \delta_j - \varepsilon, \qquad j = 1, \ldots, m \tag{7.60}$$

for any $\varepsilon > 0$.

Proof It is obvious that (7.59) implies (7.58). To see the converse, we assume (7.58) holds but (7.59) fails. Then, there exists some $x_0 \neq 0$ and $\Delta^0 = (\Delta_1^0 \ \cdots \ \Delta_m^0)$ with $|\Delta_j^0| \leq \delta_j, j = 1, \ldots, m$, such that

$$x_0^T\Omega(\Delta^0)x_0 \geq 0. \tag{7.61}$$

If such Δ^0 is only a boundary point, i.e., $|\Delta_i^0| = \delta_i$ for some i, then, (7.60) holds for any $\varepsilon > 0$. In the sequel, we assume that Δ^0 is an interior point.

We claim that $Kx_0 \neq 0$. Indeed, if $Kx_0 = 0$, then

$$x_0^T\Omega(\Delta(Kx_0))x_0 = x_0^T\Omega_0x_0 = x_0^T\Omega(\Delta^0)x_0 \geq 0 \tag{7.62}$$

by (7.57) and (7.61), which contradicts (7.58). So, $Kx_0 \neq 0$.

Because of the strict convexity of $\Omega(\cdot)$, there exists Δ^1 with

$$|\Delta_j^1| \leq \delta_j - \varepsilon_1, j = 1, \ldots, m,$$

for some small $\varepsilon_1 > 0$ such that

$$x_0^T\Omega(\Delta^1)x_0 > 0. \tag{7.63}$$

Because the above is continuous in x_0, we may perturb x_0 slightly such that (7.63) still holds and every element of Kx_0 is nonzero.

Now using Lemma 7.5, we know that $\Delta(\alpha K x_0)$ covers

$$[-\delta_1, \delta_1] \oplus \cdots \oplus [-\delta_m, \delta_m]$$

densely as α varies from $-\infty$ to ∞. Hence, there exists $\alpha \neq 0$ such that

$$x_0^T \Omega(\Delta(\alpha K x_0)) x_0 > 0.$$

Define $x_1 = \alpha x_0$, we get

$$x_1^T \Omega(\Delta(K x_1)) x_1 > 0$$

which contradicts (7.58). That is, Δ^0 cannot be an interior point. Hence, (7.58) implies (7.60). □

Proof of Theorem 7.4 The "equivalence" between the quantized feedback problem and the quadratic stabilization problem for the auxiliary system (7.51) follows from Lemma 7.6. The H_∞ condition for the latter comes from [3]. The result on observer-based feedback is identical to Theorem 7.2. □

7.3.2 Quantized Measurements

Configuration II. When quantized measurements are available, we have the following result:

Theorem 7.5 *Given the system in* (7.38) *and a quantization level vector* ρ*, consider the following auxiliary system:*

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k \\ v_k &= (I + \Delta_k) y_k \end{aligned} \tag{7.64}$$

where $|\Delta_{j,k}| \leq \delta_j$ *for all* $j = 1, 2, \ldots, m$ *and* k*, and* δ_j *are converted from* ρ_j *using* (7.8)*, and* v_k *is the output available for feedback. Suppose the auxiliary system is quadratically stablizable, then* (7.38) *is quadratically stabilizable via* (7.40)*. Conversely, suppose the system* (7.38) *is quadratically stabilizable via* (7.40) *and, in addition, suppose* $\ln \rho_i / \ln \rho_j$ *are irrational numbers for all* $i \neq j$ *when* $m > 1$*. Then, for any (arbitrarily small)* $\varepsilon > 0$*, the auxiliary system* (7.64) *with* $|\Delta_{j,k}| \leq \delta_j - \varepsilon$ *is quadratically stabilizable.*

Further, the auxiliary system is quadratically stabilizable if the following state feedback H_∞ *control has a solution* $H(z)$ *for some diagonal scaling matrix* $\Gamma > 0$*:*

$$\| \Lambda \Gamma (I - G(z)H(z))^{-1} G(z) H(z) \Gamma^{-1} \|_\infty < 1 \tag{7.65}$$

where Λ *is given in* (7.53)*. In particular, any* $H(z)$ *that renders* (7.52) *is a solution to either quadratic stabilization problem.*

Proof The "equivalence" between the quantized feedback problem and the quadratic stabilization problem for the auxiliary system (7.64) follows from Lemma 7.6. The proof for the relation to H_∞ optimization is similar to the proof of Theorem 7.3. □

7.4 Quantized Quadratic Performance Control

The purpose of this section is to extend the results in the previous sections to include a quadratic performance objective.

Consider the system in (7.38). Suppose the output y_k needs to be quantized. We now want to design a controller in (7.40) such that the following performance cost function

$$J(x_0) = \sum_{k=0}^{\infty} x_k^T Q x_k + u_k^T R u_k, \, Q = Q^T \geq 0, \quad R = R^T > 0 \tag{7.66}$$

is minimized in the sense below:

$$\min \mathbb{E}[J(x_0)] \tag{7.67}$$

In the above, x_0 is assumed to be a white noise with covariance $\mathbb{E}[x_0 x_0^T] = \sigma^2 I$ for some $\sigma > 0$.

Because the state of the closed-loop system is $\bar{x}_k$, we may rewrite the performance cost as

$$J(\bar{x}_0) = \sum_{k=0}^{\infty} \bar{x}_k^T \bar{Q} \bar{x}_k + u_k^T R u_k \tag{7.68}$$

where

$$\bar{x}_0 = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}; \quad \bar{Q} = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{7.69}$$

Suppose we want the closed-loop system to be quadratically stable. Let

$$V(\bar{x}) = \bar{x}^T \bar{P} \bar{x}, \bar{P} = \bar{P}^T > 0,$$

be the associated Lyapunov function. Define

$$\nabla V(\bar{x}_k) = V(\bar{x}_{k+1}) - V(\bar{x}_k). \tag{7.70}$$

Then, using (7.41), the performance cost is given by

$$\begin{aligned} J(\bar{x}_0) &= \bar{x}_0^T \bar{P} \bar{x}_0 + \sum_{k=0}^{\infty} \nabla V(\bar{x}_k) + \bar{x}_k^T \bar{Q} \bar{x}_k + u_k^T R u_k \\ &= \bar{x}_0^T \bar{P} \bar{x}_0 + \sum_{k=0}^{\infty} \bar{x}_k^T \bar{\Omega}(\Delta(y_k)) \bar{x}_k \end{aligned} \tag{7.71}$$

where

$$\bar{\Omega}(\Delta) = \mathscr{A}(\Delta)^T \bar{P} \mathscr{A}(\Delta) - \bar{P} + \bar{Q} + (\bar{C} + \hat{I}\Delta)\hat{C})^T \bar{K}^T \hat{I} R \hat{I}^T \bar{K} (\bar{C} + \hat{I}\Delta)\hat{C}). \quad (7.72)$$

For the case without quantization, i.e., $\Delta(\cdot) = 0$, it is well-known (and easy to see from above) that the optimal solution for $\bar{K}$ is such that $\bar{x}_k^T \bar{\Omega}(0)\bar{x}_k = 0$ for all k, which leads to $J(\bar{x}_0) = \bar{x}_0^T \bar{P} \bar{x}_0$ and minimization of $\mathrm{tr}\bar{P}$. In the presence of the quantizer, we can formulate the performance control problem as follows: Given a performance bound $\gamma > 0$ and $\rho > 0$, find $\bar{P}$, $\bar{K}$, if exist, such that

$$\mathrm{tr}(\bar{P}) < \gamma \quad (7.73)$$

subject to

$$\bar{x}^T \bar{\Omega}(\Delta(\hat{C}\bar{x}))\bar{x} < 0, \quad \forall \bar{x} \neq 0. \quad (7.74)$$

This problem will be called *Quantized Quadratic Performance Control (QQPC)* problem. The solution to this problem is related to the so-called *guaranteed-cost control (GCC)* problem for the auxiliary system (7.38) and (7.64), i.e., we want to find $\bar{P}$, $\bar{K}$ such that (7.73) holds subject to

$$\bar{\Omega}(\Delta) < 0, \quad \forall\, |\Delta_j| \leq \delta_j \quad (7.75)$$

where δ_j and ρ_j are related by (7.8).

Theorem 7.6 *Consider the system in* (7.38)*, the performance cost in* (7.66)*, the controller structure in* (7.40)*, some performance bound* $\gamma > 0$ *and quantization level vector* $0 < \rho < 1$*. Suppose the GCC problem has a solution. Then, there exists a solution to the QQPC problem.*

Conversely, if the QQPC problem has a solution and in addition (when $m > 1$*),* $\ln \rho_i / \ln \rho_j$ *are irrational numbers for all* $i \neq j$*, then, given any (arbitrarily small* $\varepsilon > 0$*), the GCC problem for* (7.75) *has a solution for* $|\Delta_{j,k}| \leq \delta_j - \varepsilon$*.*

Proof The proof is similar to that of Theorem 7.4. The key is to show the relationship between (7.74) and (7.75). Obviously, (7.75) implies (7.74). The fact that (7.74) implies (7.75) but with $|\Delta_j| \leq \delta_j - \varepsilon$ is proved using Lemma 7.6. The details are omitted here. □

When quantized state feedback is used instead, we have the following result:

Theorem 7.7 *Consider the system* (7.1) *with* $B \in \mathbb{R}^{n\times m}$ *and quantized state feedback as in* (7.2) *and* (7.3)*, where*

$$Q^\infty(\cdot) = [Q^1(\cdot), \ldots, Q^m(\cdot)]^T$$

with given quantization levels $0 < \rho_1, \ldots, \rho_m < 1$*. Given the performance cost function in* (7.66) *and a performance bound* $\gamma > 0$*, the QQPC problem becomes to*

finding $P = P^T > 0$ *and* K*, if exist, such that*

$$\mathrm{tr}(P) < \gamma \tag{7.76}$$

subject to

$$x^T \Omega(\Delta(v))x < 0, \quad \forall x \neq 0 \tag{7.77}$$

where $v = Kx$ *and*

$$\begin{aligned}\Omega(\Delta) = &(A + B(I + \Delta)K)^T P(A + B(I + \Delta)K) \\ &-P + Q + K^T(I + \Delta)R(I + \Delta)K.\end{aligned} \tag{7.78}$$

The related GCC problem becomes to finding $P = P^T > 0$ *and* K*, if exist, such that* (7.76) *holds subject to*

$$\Omega(\Delta) < 0, \quad \forall |\Delta_j| \leq \delta_j. \tag{7.79}$$

Further, the GCC problem has a solution if the following linear matrix inequalities

$$\mathrm{tr}\,\tilde{P} < \gamma, \quad \begin{bmatrix} -\tilde{P} & I \\ I & -S \end{bmatrix} \leq 0 \text{ and} \tag{7.80}$$

$$\begin{bmatrix} -S & * & * & * & * \\ AS + BW & -S + B\Lambda\Gamma\Lambda B^T & * & * & * \\ W & \Lambda\Gamma\Lambda B^T & -\tilde{R} & * & * \\ W & 0 & 0 & -\Gamma & * \\ Q^{1/2}S & 0 & 0 & 0 & -I \end{bmatrix} < 0 \tag{7.81}$$

have a solution for some $\tilde{P} = \tilde{P}^T, S = S^T$*,* W *and a diagonal scaling matrix* Γ*, where* $\tilde{R} = R^{-1} - \Lambda S \Lambda$*,* Λ *is given in* (7.53)*, and* $*$ *denotes the symmetric part in the matrix. Also,* P *and* K *are related to* S *and* W *as follows:*

$$P = S^{-1}, \quad K = WP. \tag{7.82}$$

For the single-input case, (7.80) *and* (7.81) *are also necessary.*

Proof The simplification of the QQPC and GCC problems is easy to check. We proceed to verify (7.81) as a sufficient condition for the GCC problem. Indeed, (7.79) holds if and only if

$$\begin{bmatrix} -P + Q & * & * \\ A + B(I + \Delta)K & -P^{-1} & * \\ (I + \Delta)K & 0 & -R^{-1} \end{bmatrix} < 0 \tag{7.83}$$

for all $|\Delta_j| \le \delta_j$. Using (7.82), the above becomes

$$\begin{bmatrix} -S+SQS & * & * \\ AS+B(I+\Delta)W & -S & * \\ (I+\Delta)W & 0 & -R^{-1} \end{bmatrix} < 0 \tag{7.84}$$

which is equivalent to

$$\begin{bmatrix} -S+SQS & * & * \\ AS+BW & -S & * \\ W & 0 & -R^{-1} \end{bmatrix} + \begin{bmatrix} 0 \\ B \\ I \end{bmatrix} \Delta [W\ 0\ 0] + \begin{bmatrix} W^T \\ 0 \\ 0 \end{bmatrix} \Delta [0\ B^T\ I] < 0. \tag{7.85}$$

Taking $\Gamma > 0$ to be any diagonal scaling matrix, (7.85) holds if

$$\begin{bmatrix} -S+SQS & * & * \\ AS+BW & -S & * \\ W & 0 & -R^{-1} \end{bmatrix} + \begin{bmatrix} 0 \\ B\Lambda \\ \Lambda \end{bmatrix} \Gamma [0\ \Lambda B^T\ \Lambda] + \begin{bmatrix} W^T \\ 0 \\ 0 \end{bmatrix} \Gamma^{-1} [W\ 0\ 0] < 0 \tag{7.86}$$

which is equivalent to (7.81) using Schur complement. Note that the conversion from (7.85) to (7.86) is lossless in the single-input case. □

7.5 Quantized H_∞ Control

Here we extend the quantization results to H_∞ control. For simplicity, only quantized state feedback is considered. This problem in the SISO setting has been studied in [6]. Our main purpose is to show that the sector bound approach can be easily generalized to quantized feedback H_∞ control.

The system of interest is as follows:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + B_1 w_k \\ z_k &= Cx_k + Du_k + D_1 w_k \end{aligned} \tag{7.87}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $w \in \mathbb{R}^{m_1}$, $z \in \mathbb{R}^\ell$ and the control signal is in the form of (7.2) and (7.3). Given a quantization level vector ρ and H_∞ performance bound $\gamma > 0$, the design objective is to find K such that the induced ℓ^2-gain from w to z is less than γ.

It is easy to verify that the closed-loop system is given by

$$\begin{aligned} x_{k+1} &= [A + B(I + \Delta(v))K]x_k + B_1 w_k, \\ z_k &= [C + D(I + \Delta(v))K]x_k + D_1 w_k. \end{aligned} \tag{7.88}$$

As in the quadratic performance control problem, we consider the following relaxed H_∞ control problem: Find $P = P^T > 0$ and K such that

$$x^T \Pi(\Delta(Kx))x < 0, \quad \forall x \neq 0 \tag{7.89}$$

where

$$\begin{aligned} \Pi(\Delta) &= A_\Delta^T P A_\Delta - P + \gamma^{-2}(A_\Delta^T P B_1 + C_\Delta^T D_1) \\ &\quad \times [I - \gamma^{-2}(D_1^T D_1 + B_1^T P B_1)]^{-1} \\ &\quad \times (B_1^T P A_\Delta + D_1^T C_\Delta) + C_\Delta^T C_\Delta, \end{aligned} \tag{7.90}$$

$$A_\Delta = A + B(I + \Delta)K, \quad C_\Delta = C + D(I + \Delta)K. \tag{7.91}$$

The motivation for formulating the problem above is as follows: When $\Delta = 0$, (7.91) is a necessary and sufficient condition for the H_∞ norm of the transfer function from w to z to be less than γ. When $\gamma \to \infty$, (7.91) recovers the condition for stabilization. When γ is finite and Δ represents a sector bound uncertainty, the condition $\Pi(\Delta) < 0$ corresponds to the robust H_∞ control problem studied in [3].

Theorem 7.8 *Consider the given system* (7.87)*, controller structure* (7.2) *and* (7.3)*, quantization level vector ρ and a H_∞ performance bound $\gamma > 0$. Suppose there exist $P = P^T > 0$ and K such that* (7.89) *holds, then the induced ℓ^2-norm from w to z is less than γ.*

Further, for any $P = P^T > 0$ and K, (7.89) *holds if $\Pi(\Delta) < 0$ for all $|\Delta_j| \leq \delta_j$, where δ_j are related to ρ_j by* (7.6)*. Conversely, if* (7.89) *holds, $\Pi(\Delta) < 0$ for all $|\Delta_j| \leq \delta_j - \varepsilon$, where $\varepsilon > 0$ is arbitrarily small.*

In addition, there exist $P = P^T > 0$ and K such that $\Pi(\Delta) < 0$ for all $|\Delta_j| \leq \delta_j$ if the following linear matrix inequality (LMI)

$$\begin{bmatrix} -S + B\Lambda\Gamma\Lambda B^T & * & * & * & * \\ (AS + BW)^T & -S & * & * & * \\ B_1^T & 0 & -\gamma I & * & * \\ D\Lambda\Gamma\Lambda B^T & CS + D_1 W & D_1 & -\gamma I & * \\ 0 & W & 0 & 0 & -\Gamma \end{bmatrix} < 0 \tag{7.92}$$

has a solution for $S = S^T, W$ and diagonal scaling matrix Γ, where Λ and the relationship between (S, W) and (P, K) are the same as in Theorem 7.7*. In the single-input case, the LMI in* (7.92) *is also necessary.*

Proof The proof is similar to that of Theorem 7.7. The details are omitted. □

Corollary 7.1 *As $\gamma \to \infty$, the LMI condition* (7.92) *is equivalent to the condition* (7.52) *in Theorem* 7.4.

Proof When $\gamma \to \infty$, (7.92) implies

$$\begin{bmatrix} -S + B\Lambda\Gamma\Lambda B^T & * & * \\ (AS + BW)^T & -S & * \\ 0 & W & -\Gamma \end{bmatrix} < 0.$$

By Schur complement and letting $K = WS^{-1}$, we have

$$(A + BK)(S^{-1} - K^T\Gamma^{-1}K)^{-1}(A + BK)^T - S + B\Lambda\Gamma\Lambda B^T < 0$$

i.e.,

$$(A + BK)^T(S - B\Lambda\Gamma\Lambda B^T)^{-1}(A + BK) - S^{-1} + K^T\Gamma^{-1}K < 0.$$

The above is equivalent to [5], i.e.,

$$\|\Gamma^{-1/2}K(zI - A - BK)^{-1}B\Lambda\Gamma^{1/2}\|_\infty < 1.$$

Letting $\Gamma_1^{-1} = \Gamma^{1/2}\Lambda$ and noting that both Λ and Γ are diagonal matrices, the above can be rewritten as

$$\|\Lambda\Gamma_1 K(zI - A - BK)^{-1}B\Gamma_1^{-1}\|_\infty < 1$$

which is (7.52) in Theorem 7.4. □

Remark 7.6 As we mentioned earlier, the quantized H_∞ control problem has been studied in [6]. We now comment on the connection between Theorem 7.8 and a related result in [6] (Theorem 5.1: Discrete-time). The problem formulation in [6] is more restrictive because it treats the single input case and assumes $C^TD = 0$, $D^TD = I$ and $D_1 = 0$. The coarsest quantization density ρ given in [6] can be written as

$$\rho = \frac{\alpha - 1}{\alpha + 1}$$

where α is the optimal solution to the following problem:

$$\inf \alpha \text{ subject to } \Sigma(X) > \Sigma_0, \quad \alpha > 1, \quad X = X^T > 0 \tag{7.93}$$

where

$$\Sigma(X) = \begin{bmatrix} \alpha^2 X - AXA^T & -AXA^T & AXC^T \\ -AXA^T & \frac{\alpha^2}{\alpha^2-1}X - AXA^T & AXC^T \\ CXA^T & CXA^T & -CXC^T \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} \alpha^2 B_1 B_1^T & 0 & 0 \\ 0 & \frac{\alpha^2}{\alpha^2-1}(B_1 B_1^T - \gamma^2 B_2 B_2^T) & 0 \\ 0 & 0 & -\gamma^2 I \end{bmatrix}.$$

Although we do not provide the details (which are quite involved), it can be shown that, by setting $\alpha = \delta^{-1}$, this condition is equivalent to the condition in Theorem 7.8 (when specialized under the assumptions on C, D and D_1). That is, Theorem 7.8 generalizes the result in [6].

7.6 Summary

We have shown that the classical sector bound method can be used to study quantized feedback control problems in a non-conservative manner. Various cases have been considered: quantized state feedback control, quantized output feedback control, MIMO systems, and control with performances. In all these problems, the key result is that quantization errors can be converted into sector bound uncertainties without conservatism. By doing so, quantized feedback control problems become well-known robust control problems.

For quadratic stabilization of SISO systems (using either quantized state feedback or quantized output feedback), complete solutions are available by solving related H_∞ optimization problems. For MIMO systems or SISO systems with a performance control objective, the resulting robust control problems usually do not have simple solutions, thus sufficient conditions on quantization densities are derived. These conditions are expressed either in terms of H_∞ optimization or linear matrix inequalities. Note that these conditions are for a given set of quantization densities. But because these conditions are convex in the sector bounds associated with the quantization densities, optimal quantization densities can be easily computed numerically.

Finally, we note that the use of the sector bound method also explains why it is difficult to find the coarsest quantization densities in the cases of MIMO stabilization and/or performance control problems. More precisely, the difficulties are the same as finding non-conservative solutions to the related robust control problems, which are known to be very difficult.

The results in this chapter are based mainly on [7, 8].

References

1. N. Elia, S. Mitter, Stabilization of linear systems with limited information. IEEE Trans. Autom. Control **46**(9), 1384–1400 (2001)
2. A. Packard, J. Doyle, Quadratic stability with real and complex perturbations. IEEE Trans. Autom. Control **35**(2), 198–201 (1990)
3. C. de Souza, M. Fu, L. Xie, H_∞ analysis and synthesis of discrete-time systems with time-varying uncertainty. IEEE Trans. Autom. Control **38**(3), 459–462 (1993)
4. L. Xie, Y. Soh, Guaranteed cost control of uncertain discrete-time systems. Control Theory Adv. Technol. **10**(4), 1235–1251 (1995)
5. C. de Souza, L. Xie, On the discrete-time bounded real lemma with application in the characterization of static state feedback h_∞ controller. Syst. Control Lett. **18**(1), 61–71 (1992)
6. N. Elia, Design of hybrid systems with guaranteed performance, in *Proceedings of the 39th IEEE Conference on Decision and Control*, vol. 1 (2000)
7. M. Fu, L. Xie, The sector bound approach to quantized feedback control. IEEE Trans. Autom. Control **50**(11), 1698–1711 (2005)
8. M. Fu, L. Xie, Finite-level quantized feedback control for linear systems. IEEE Trans. Autom. Control **54**(5), 1165–1170 (2009)

Chapter 8
Stabilization of Linear Systems via Finite-Level Logarithmic Quantization

The logarithmic quantizer is shown in the last chapter to give the coarsest quantization density for quadratic stabilization of an unstable single input linear system. However, it requires an infinite data rate. An interesting problem is whether an unstable linear system can be stabilized using a finite-level logarithmic quantizer with a dynamic scaling. In addition, it is unclear whether a logarithmic quantizer can approach the well-known minimum data rate required for stabilizing an unstable linear system.

To be specific, we ask the following question in this chapter: (1) Can we achieve the quadratic stabilization with a finite-level logarithmic quantizer? (2) Does a logarithmic quantizer require an average data rate higher than the minimum average data rate in (2.10) for stabilization of linear systems? This chapter shows that the answer to the first question is positive while it is negative for the second question. The results are confirmed by showing that the use of a dynamical finite-level logarithmic quantizer with a variable data rate. From this viewpoint, the logarithmic quantizer is optimal.

The chapter is organized as follows. In Sect. 8.1, we prove that the quadratic stabilization of linear systems is achievable by using a finite-level quantization. In Sect. 8.2, the attainability of the minimum average data rate via logarithmic quantization is proved. Concluding remarks are drawn in Sect. 8.3.

8.1 Quadratic Stabilization via Finite-level Quantization

8.1.1 Finite-level Quantizer

A logarithmic quantizer (7.8) has an infinite number of quantization levels and is not implementable practically. One simple approach is to truncate the quantizer using a

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_8

large saturator and a small dead zone. That is, we use a $2N$-level logarithmic quantizer with quantization density $\rho > \rho_{\inf}$:

$$Q(y) = \begin{cases} \rho^i \mu_0, & \text{if } \frac{1}{1+\delta}\rho^i\mu_0 < y \le \frac{1}{1-\delta}\rho^i\mu_0,\ 0 < i < N-1, \\ \rho^{N-1}\mu_0, & \text{if } 0 \le y \le \frac{1}{1-\delta}\rho^{N-1}\mu_0, \\ \mu_0, & \text{if } y > \frac{1}{1+\delta}\mu_0, \\ -Q(-y), & \text{if } y < 0. \end{cases} \tag{8.1}$$

This quantization scheme will allow the state of the system to converge to a small neighborhood, provided that the initial state is within a known bound.

Our main objective here is to show that it is possible to dynamically scale the input-output signals of the quantizer so that asymptotic stabilization can be achieved using a finite-level logarithmic quantizer, even without knowing the bound for the initial state.

The basic idea of dynamic scaling is very simple: When the signal y_k is outside of the quantization range, we scale it back by a *scaling factor* (or *gain*) $g_k > 0$ before quantization. The quantized signal is then scaled back by g_k^{-1}. That is, we use

$$v_k = g_k^{-1} Q(g_k y_k). \tag{8.2}$$

The key problem with dynamic scaling is how to design g_k. The main technical difficulty is that there is no separate feedback channel to communicate the gain value. One approach is that both sides of the feedback channel compute the same g_k independently. This is possible only when the gain g_k can be computed using only the quantized signal because this signal is available to both sides of the feedback channel, assuming no packet losses and transmission errors. In the sequel, we introduce a very simple dynamic scaling method.

The closed-loop system of (7.1)–(7.2), (7.5)–(7.9) and (8.2) is given by

$$\bar{x}_{k+1} = \bar{A}\bar{x}_k + \bar{B} g_k^{-1} Q(g_k \bar{C}\bar{x}_k), \tag{8.3}$$

where $\bar{x} = [x^T \ \hat{x}^T]^T$, and

$$\bar{A} = \begin{bmatrix} A & BC_c \\ 0 & A_c \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} BD_c \\ B_c \end{bmatrix}, \quad \bar{C} = [C \ \ 0]. \tag{8.4}$$

For the moment, we assume that an infinite-level logarithmic quantizer with density $\rho > \rho_{\inf}$ is adopted. Then, g_k has no effect. Following the sector bound approach [1], we can write (8.3) as

$$\bar{x}_{k+1} = \bar{A}(\Delta_k)\bar{x}_k, \tag{8.5}$$

where $\bar{A}(\Delta_k) = \bar{A} + \bar{B}(1+\Delta_k)\bar{C}$ and Δ_k represents the quantization error defined by $\Delta_k y_k = Q(y_k) - y_k$ with $-\delta \le \Delta_k \le \delta$.

Because (8.5) is quadratically stable, we have a quadratic Lyapunov function $V(\bar{x}) = \bar{x}^T P \bar{x}$ with $P = P^T > 0$ such that [1]

$$\bar{A}(\Delta)^T P \bar{A}(\Delta) < P, \quad \forall |\Delta| \le \delta. \tag{8.6}$$

Using the continuity argument, the above is equivalent to that

$$\bar{A}(\Delta)^T P \bar{A}(\Delta) \le (1-\eta) P, \quad \forall |\Delta| \le \delta, \tag{8.7}$$

for some $0 < \eta < 1$.

We now assume that a $2N$-level logarithmic quantizer with the same density ρ and dynamic scaling (8.1)–(8.2) is applied instead. Let γ_1 and γ_2 be two positive scaling factors such that $0 < \gamma_1 < 1$, $\sqrt{1-\eta} < \gamma_2 < 1$, and

$$\gamma_1^2 \bar{A}^T P \bar{A} < (1-\eta) P. \tag{8.8}$$

Note that (8.8) is always possible by taking γ_1 sufficiently small.

We initialize g_0 to be any positive value and define g_{k+1} for any $k \ge 0$ as follows:

$$g_{k+1} = \begin{cases} g_k \gamma_1, & \text{if } |Q(g_k y_k)| = \mu_0, \\ g_k / \gamma_2, & \text{if } |Q(g_k y_k)| = \rho^{N-1} \mu_0, \\ g_k, & \text{otherwise.} \end{cases} \tag{8.9}$$

Because of the flexibility in g_0, there is no loss of generality to normalize $\mu_0 = 1$ in (8.1). We will also denote $\varepsilon = \rho^{N-1}$. The choice of g_0 does not affect stabilizability, but choosing it according to an estimate of $\|x_0\|$ helps improve the transient performance; see Example 8.2 in Sect. 8.1.4.

Consider the scaled state defined by

$$z_k = g_k \bar{x}_k \tag{8.10}$$

and the associated Lyapunov function $V(z) = z^T P z$. We have the following result:

Lemma 8.1 *Consider the closed-loop system (8.3) with a scaled $2N$-level logarithmic quantizer (8.1) and (8.9), where ρ in (8.1) is such that $\rho > \rho_{\inf}$, and γ_1, γ_2 and η are chosen according to (8.7)–(8.8). Then, for any initial state x_0 and any $k \ge 0$,*

$$V(z_{k+1}) \le \begin{cases} (1-\eta) V(z_k), & \text{if } \varepsilon < |Q(\bar{C} z_k)| \le 1, \\ (1-\eta_1) V(z_k) + \eta_2 \varepsilon^2, & \text{if } |Q(\bar{C} z_k)| = \varepsilon, \end{cases} \tag{8.11}$$

where

$$\eta_1 = 1 - \gamma_2^{-2}(1+\tau)(1-\eta), \tag{8.12}$$

$$\eta_2 = \gamma_2^{-2}(1+\tau^{-1}) \bar{B}^T P \bar{B} \tag{8.13}$$

with τ being any positive constant satisfying $\eta_1 > 0$.

Proof The result for the case of $\varepsilon < |Q(\bar{C}z_k)| < 1$ follows directly from (8.5), (8.7) and $g_{k+1} = g_k$. For the case of $|Q(\bar{C}z_k)| = 1$, we obtain $g_{k+1} = g_k\gamma_1$. It follows that

$$V(z_{k+1}) = \gamma_1^2(\bar{A}z_k + \bar{B}\sigma_k)^T P(\bar{A}z_k + \bar{B}\sigma_k),$$

where $\sigma_k = \text{sign}(\bar{C}z_k)$. Denote

$$f(u) = \gamma_1^2(\bar{A}z_k + \bar{B}u)^T P(\bar{A}z_k + \bar{B}u).$$

From (8.8), it follows that $f(0) \le (1-\eta)V(z_k)$.

Since $|Q(\bar{C}z_k)| = 1$, we have $\sigma_k = \theta u_1$ for some $0 < \theta \le 1$, where $u_1 = (1+\Delta_k)\bar{C}z_k$ with $|\Delta_k| \le \delta$ is the unsaturated output of the quantizer. Also from (8.7), we get

$$f(u_1) = \gamma_1^2 z_k^T \bar{A}(\Delta_k)^T P\bar{A}(\Delta_k)z_k \le \gamma_1^2(1-\eta)V(z_k).$$

Since $f(u)$ is quadratic and convex (because $f(u) \to \infty$ when $|u| \to \infty$), it is clear that

$$V(z_{k+1}) = f(\sigma_k) \le \max\{f(0),\ f(u_1)\} \le (1-\eta)V(z_k).$$

For the case of $|Q(\bar{C}z_k)| = \varepsilon$, it follows that $g_{k+1} = g_k/\gamma_2$. From (7.52) and (8.5), we can write

$$\bar{x}_{k+1} = \bar{A}(\Delta_k)\bar{x}_k + \bar{B}g_k^{-1}\varepsilon_k,$$

where $|\varepsilon_k| \le \varepsilon$. It follows that

$$\begin{aligned}
V(z_{k+1}) &= \gamma_2^{-2}(\bar{A}(\Delta_k)z_k + \bar{B}\varepsilon_k)^T P(\bar{A}(\Delta_k)z_k + \bar{B}\varepsilon_k) \\
&= \gamma_2^{-2} z_k^T \bar{A}(\Delta_k)^T P\bar{A}(\Delta_k)z_k + \gamma_2^{-2}(2\varepsilon_k\bar{B}^T P\bar{A}(\Delta_k)z_k + \varepsilon_k^2\bar{B}^T P\bar{B}) \\
&\le \gamma_2^{-2}(1+\tau) z_k^T \bar{A}(\Delta_k)^T P\bar{A}(\Delta_k)z_k + \gamma_2^{-2}(1+\tau^{-1})\varepsilon^2\bar{B}^T P\bar{B} \\
&\le \gamma_2^{-2}(1+\tau)(1-\eta) z_k^T P z_k + \eta_2\varepsilon^2 \\
&= (1-\eta_1)V(z_k) + \eta_2\varepsilon^2.
\end{aligned}$$

The above holds for any $\tau > 0$. Since $\sqrt{1-\eta} < \gamma_2 < 1$, we can choose τ sufficiently small to ensure $\eta_1 > 0$. □

From Lemma 8.1, it is clear that $V(z_k)$ converges to a bounded region. This bound can be computed by solving $-\eta_1 V_\infty + \eta_2\varepsilon^2 = 0$, which gives

$$V_\infty = \eta_1^{-1}\eta_2\varepsilon^2. \tag{8.14}$$

Lemma 8.1 leads to the following result:

Corollary 8.1 *Suppose the scaled 2N-level logarithmic quantizer* (8.1), (8.2) *and* (8.9) *is applied. Then, for any initial state* x_0, $z_k = g_k\bar{x}_k$ *converges exponentially to the ellipsoid*

$$Z_\infty = \{z : z \in \mathbb{R}^{2n}, V(z) \le V_\infty\}. \tag{8.15}$$

From (8.14) and the corollary above, it is clear that we can choose N to be sufficiently large so that, when k is sufficiently large, $Q(\bar{C}z_k)$ will no longer be saturated. This is achieved by choosing N such that

$$|\bar{C}z| < 1, \quad \forall z^T P z \le \eta_1^{-1}\eta_2\rho^{2(N-1)}.$$

Since $\bar{C}z$ is a scalar, the above implies that

$$\bar{C}zz^T\bar{C}^T < 1, \; zz^T \le \eta_1^{-1}\eta_2\rho^{2(N-1)}P^{-1}.$$

By substituting the second matrix inequality into the first one, we obtain $N > N_0$, where

$$N_0 = 1 + \frac{\log(\eta_1^{-1}\gamma_2^{-2}(1+\tau^{-1})\bar{B}^T P\bar{B}\bar{C}P^{-1}\bar{C}^T)}{2\log(\rho^{-1})}. \tag{8.16}$$

The analysis above yields the following main result:

Theorem 8.1 *Suppose the scaled $2N$-level logarithmic quantizer (8.1), (8.2) and (8.9) is applied with $N > N_0$ in (8.16). Then, the state $\bar{x}_k$ converges to zero asymptotically.*

Proof From Corollary 8.1, it follows that z_k converges to Z_∞ exponentially. This property and the choice of N_0 imply that $Q(\bar{C}z_k)$ will no longer be saturated after a finite number of steps, say k_0 steps. This means that g_k will be non-decreasing for $k \ge k_0$. Note that whenever $g_{k+1} = g_k$, $V(z_k)$ decreases exponentially. If this continues for enough number of steps, $|\bar{C}z_k|$ will be less than ε, forcing g_{k+1} to increase by factor of $1/\gamma_2$. Thus, g_k cannot converge to a constant. Hence, $g_k \to \infty$ as $k \to \infty$. Since z_k is bounded for $k > k_0$, we conclude that $\bar{x}_k \to 0$ as $k \to \infty$. □

Remark 8.1 A typical behavior of the system is as follows. If the initial state is very large, the feedback signal tends to be saturated, forcing g_k to decrease fast. This would result in a period of overshoot. Once g_k is sufficiently small, saturation will stop and the state decays exponentially. When the state is sufficiently small, g_k will increase gradually, causing the quantizer to bounce back and forth between the dead zone and logarithmic region. During this phase, the state also decays exponentially, but at a lower rate.

8.1.2 Number of Quantization Levels

In this section, we try to analyze the number of quantization levels needed for stabilization. Recall that for a given controller (7.5)–(7.9) with an infinite-level logarithmic quantizer with density $\rho > \rho_{\inf}$ that quadratically stabilizes the system (7.1)–(7.2),

a sufficient number of quantization levels is given by N_0 of (8.16). However, this formula is complicated because N_0 depends on a number of design parameters (η, γ_2, ρ, P, τ, and the controller). In the sequel, we consider how to choose these parameters.

We first minimize N_0 of (8.16) with respect to τ by assuming that other parameters are fixed. From (8.16), it is clear that minimizing N_0 is equivalent to

$$\min_{\tau>0} \eta_1^{-1}\gamma_2^{-2}(1+\tau^{-1}), \tag{8.17}$$

where η_1 is given in (8.12). The solution to (8.17) is simply given by

$$\tau = \frac{\gamma_2}{\sqrt{1-\eta}} - 1, \quad \eta_1 = 1 - \gamma_2^{-1}\sqrt{1-\eta}, \tag{8.18}$$

$$\min_{\tau>0} \eta_1^{-1}\gamma_2^{-2}(1+\tau^{-1}) = (\gamma_2 - \sqrt{1-\eta})^{-2}. \tag{8.19}$$

Applying the above to (8.16) and noting $\log(\rho^{-1}) = -\log(\rho)$,

$$N_0 = 1 + \frac{2\log(\gamma_2 - \sqrt{1-\eta}) - \log(\bar{B}^T P \bar{B} \bar{C} P^{-1} \bar{C}^T)}{2\log(\rho)}. \tag{8.20}$$

We next discuss the effect of γ_2 on N_0. Since $\gamma_2 < 1$ is required, it is clear from (8.20) that N_0 is minimized by taking γ_2 very close to 1, which, however, makes g_k increase very slowly, as seen from (8.9), resulting in that $\bar{x}_k$ converges to 0 very slowly. A good choice for γ_2 should balance the convergence rate of $\bar{x}_k$ and the number of quantization levels; see Example 8.1.

With $\gamma_2 < 1$ chosen, we shall now minimize N_0 with respect to ρ, η, the controller and its associated P. Observe from (8.20) that N_0 can be reduced by increasing η and δ (or decreasing ρ). However, we can see from (8.7) that a larger η requires δ to be small. Furthermore, the choice of δ and η affect P and the controller. This implies that η and δ need to be optimized jointly. To this end, we return to (8.7) and provide the following relationship between δ and η.

Theorem 8.2 *For any given* $0 < \delta < \delta_{\sup}$, $0 < \eta < 1$ *and* $\sqrt{1-\eta} < \gamma_2 < 1$, N_0 *in* (8.20) *is minimized by solving the following optimization problem:*

$$\lambda_{\min} = argmin_{\{X,Y,R,S,W,D_c,\lambda_1,\lambda_2\}} \lambda_1\lambda_2, \tag{8.21}$$

subject to the following linear matrix inequalties:

$$\begin{bmatrix} (\eta-1)Y & * & * & * & * & * \\ (\eta-1)I & (\eta-1)X & * & * & * & * \\ AY+BW & A+BD_cC & -Y & * & * & * \\ R & XA+SC & -I & -X & * & * \\ 0 & 0 & D_c^T B^T & S^T & -1 & * \\ CY & C & 0 & 0 & 0 & -\delta^{-2} \end{bmatrix} < 0, \tag{8.22}$$

$$\begin{bmatrix} -Y & -I & BD_c \\ -I & -X & S \\ D_c^T B^T & S^T & -\lambda_1 \end{bmatrix} < 0, \quad CYC^T < \lambda_2, \tag{8.23}$$

where $X = X^T, Y = Y^T \in \mathbb{R}^{n\times n}, R \in \mathbb{R}^{n\times n}, W \in \mathbb{R}^{1\times n}, S \in \mathbb{R}^{n\times 1}$ *and* λ_1 *and* λ_2 *are scalars. The optimal* N_0 *is given by*

$$N_0 = 1 + \frac{2\log(\gamma_2 - \sqrt{1-\eta}) - \log(\lambda_{\min})}{2\log(\rho)} \tag{8.24}$$

and the optimal controller (7.5)–(7.9) *is given by the solution of* D_c *together with*

$$C_c = (W - D_c CY)\Psi^{-T}, \tag{8.25}$$

$$B_c = M^{-1}(S - XBD_c), \tag{8.26}$$

$$A_c = M^{-1}(R - XAY - XBD_c CY - MB_c CY)\Psi^{-T} - M^{-1}XBC_c, \tag{8.27}$$

where M *and* Ψ *are any nonsingular matrices solving*

$$M\Psi^T = I - XY \tag{8.28}$$

with M *being the free parameter determining the state space realization of the controller.*

Proof First, we apply the S-Procedure [11] and standard technique of change of variables [10] to (8.7) to obtain (8.22). Then, observe from (8.24) that for the given ρ, η and γ_2, minimizing N_0 with respect to the controller and the matrix P is equivalent to minimizing $\lambda_1\lambda_2$, where $\bar{B}^T P\bar{B} < \lambda_1$ and $\bar{C}P^{-1}\bar{C}^T < \lambda_2$, which is equivalent to (8.23), following Schur complement and the above change of variables. The detail is omitted. □

We now explain how to solve the optimization problem in Theorem 8.2. We assume that $\gamma_2 < 1$ is pre-specified (say, e.g., $\gamma_2 = 0.9$). Now, given η and δ (or ρ), (8.21) is bilinear in λ_1 and λ_2. Note that when $\lambda_1 = 1$, the first inequality in (8.23) is a part of (8.22). This means that the minimum of $\lambda_1\lambda_2$ is achieved at some $0 < \lambda_1 < 1$. Then, the minimum of N_0 can be found by numerically searching over a three-dimensional set Ω defined by

$$\Omega = \{(\delta, \eta, \lambda_1) : 0 < \delta < \delta_{\sup},\ 1 - \gamma_2^2 < \eta < 1,\ 0 < \lambda_1 < 1\},$$

and solving (8.21) for each chosen candidate in Ω. A simple brute force method, which also works well, is to discretize Ω uniformly in each dimension and solve (8.21) at each grid point. Since we only need an integer solution for the number of bits $N_b = \log_2(2N_0)$, the discretization can typically be done coarsely.

8.1.3 Robustness Against Additive Noises

Next, we consider the scenario where the system (7.1)–(7.2) is subject to some bounded additive noise, i.e., we consider the following system instead:

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{8.29}$$
$$y_k = Cx_k, \tag{8.30}$$

where $||w_k|| \le \bar{w}$ for some constant $\bar{w} > 0$. The corresponding closed-loop system becomes

$$\bar{x}_{k+1} = \bar{A}\bar{x}_k + \bar{B}g_k^{-1}Q(g_k\bar{C}_k\bar{x}_k) + \bar{I}w_k, \tag{8.31}$$

where $\bar{I} = [I \ \ 0]^T$. Using the scaled state $z_k = g_k\bar{x}_k$, (8.31) becomes

$$z_{k+1} = g_{k+1}^{-1}g_k\left(\bar{A}z_k + \bar{B}Q(\bar{C}_k z_k) + g_k\bar{I}w_k\right). \tag{8.32}$$

In this case, we want to drive the state to a bounded region. To do so, we first generalize Lemma 8.1 as follows:

Lemma 8.2 *Consider the system* (8.29)–(8.30) *and the dynamically scaled logarithmic controller as given before. Then, the scaled state* $z_k = g_k\bar{x}_k$ *is bounded as follows:*

$$V(z_{k+1}) \le \begin{cases} (1+\alpha)(1-\eta))V(z_k) & \text{if } \varepsilon < |Q(\bar{C}z_k)| \le 1, \\ +(1+\alpha^{-1})\|P_{11}\|g_{k+1}^2\bar{w}^2, & \\ (1+\alpha)\gamma_2^{-1}\sqrt{1-\eta}V(z_k) & \text{if } |Q(\bar{C}z_k)| = \varepsilon, \\ +(1+\alpha)\eta_2\varepsilon^2 & \\ +(1+\alpha^{-1})\|P_{11}\|g_{k+1}^2\bar{w}^2, & \end{cases} \tag{8.33}$$

for any $\alpha > 0$, *where* $P_{11} = \bar{I}^T P\bar{I}$.

Proof The proof is simply extended from the proof of Lemma 8.1. The detail is omitted. □

From Lemma 8.2, we see that by choosing α sufficiently small, the scaled state converges to a bounded set when g_k has an upper bound $\bar{g}$. There are two steady state bounding sets for $V(z_k)$ from (8.33), associated with the three cases of $|Q(\bar{C}z_k)|$, and they are given by

$$Z_{\infty,1} = \left\{z : V(z) \le \frac{1+\alpha^{-1}}{1-(1+\alpha)(1-\eta)}\|P\|\bar{g}^2\bar{w}^2\right\}, \tag{8.34}$$

$$Z_{\infty,2} = \left\{z : V(z) \le \frac{(1+\alpha)\eta_2}{1-(1+\alpha)\gamma_2^{-1}\sqrt{1-\eta}}\varepsilon^2 + \frac{1+\alpha^{-1}}{1-(1+\alpha)\gamma_2^{-1}\sqrt{1-\eta}}\|P\|\bar{g}^2\bar{w}^2\right\}. \tag{8.35}$$

It is straightforward to minimize $Z_{\infty,1}$ with respect to α and the result is given by

$$Z_{\infty,1} = \left\{ z : V(z) \leq \frac{1}{(1-\sqrt{1-\eta})^2} \|P_{11}\| \bar{g}^2 \bar{w}^2 \right\}. \tag{8.36}$$

Noting that the overall minimization of $Z_{\infty,2}$ is difficult, we choose to minimize the term associated with $\bar{w}$. It is easy to verify that the result is given by

$$\begin{aligned} Z_{\infty,2} = \Bigg\{ z : V(z) \leq & \frac{\eta_2}{\sqrt{1-\eta_1}(1-\sqrt{1-\eta_1})} \varepsilon^2 \\ & + \frac{1}{(1-\sqrt{1-\eta_1})^2} \|P_{11}\| \bar{g}^2 \bar{w}^2 \Bigg\}, \end{aligned} \tag{8.37}$$

where $1-\eta_1 = \gamma_2^{-1}\sqrt{1-\eta}$.

Theorem 8.3 *Consider the system* (8.29)–(8.30) *and the dynamically scaled logarithmic controller as given before. We require $N > N_0$ with N_0 given by* (8.20) *and modify the scaling factor g_k by saturating it at some $\bar{g}$. Then, both the closed-loop system state $\bar{x}_k$ and the scaled state $z_k = g_k \bar{x}_k$ are bounded when $k \to \infty$.*

Proof The asymptotic boundedness of z_k follows easily from (8.36)–(8.37). To show the asymptotic boundedness of $\bar{x}_k$, it suffices to show that g_k has some lower bound $\underline{g}$ asymptotically. To do so, we define

$$Z_{\infty,3} = \{ z : z^T P z \leq (\bar{C} P^{-1} \bar{C}^T)^{-1} \}.$$

It is easy to check that $|\bar{C}z| \leq 1$ for all $z \in Z_{\infty,3}$. Since $N > N_0$, we know that $Z_\infty = \lambda Z_{\infty,3}$ for some $0 < \lambda < 1$, where Z_∞ is defined in (8.15) and

$$\lambda Z_{\infty,3} = \{ z : \lambda^{-1} z \in Z_{\infty,3} \}.$$

Next, we note that if $g_{k+1} \to 0$, we can take $\alpha = g_{k+1}$ so that (8.33) becomes (8.11). Using $Z_\infty = \lambda Z_{\infty,3}$, the above means that there exist $\check{g}$ and η_3, both positive and sufficiently small, such that, if $g_{k+1} \leq \check{g}$, then

$$V(z_{k+1}) \leq (1-\eta_3) V(z_k), \quad \forall\, z_k \notin Z_{\infty,3} \text{ or } |\bar{C} z_k| > \varepsilon. \tag{8.38}$$

The exponential convergence rate above implies that there exists some integer $\kappa > 0$ such that for any initial $z_k \in Z_{\infty,2}$ in (8.37), it takes at most κ steps for z_k to reach $Z_{\infty,3}$, provided that g_{k+1} can be kept below $\check{g}$ all the way. Once $z_k \in Z_{\infty,3}$, it will stay there until $g_k > \check{g}$ again. Since the exponential decay in (8.38) continues to happen as long as $|\bar{C} z_k| > \varepsilon$, z_k will decay sufficiently to allow g_k to grow back until $g_k > \check{g}$.

Now we define $\underline{g} = \gamma_1^{\kappa+1} \check{g}$ and proceed to prove that $g_k \geq \underline{g}$ asymptotically. We assume, on the contrary, that there exists an increasing sequence of k_i, $i = 1, 2, \ldots$, such that $k_i \to \infty$ as $i \to \infty$ and $g_{k_i} < \underline{g}$ for all i. Since $z_k \to Z_{\infty,2}$ in (8.37) as

$k \to \infty$, we may assume that k_1 is so large that $z_{k_1-\kappa} \in Z_{\infty,2}$. From the definition of $\underline{g}$, we know that $g_k \le \check{g}$ for all $k_1 - \kappa < k < k_1$. Hence, from our earlier discussion, we know that $z_{k_1} \in Z_{\infty,3}$ and that g_k will stop decaying when $k > k_1$ and will eventually grow back to $g_k > \check{g}$ while keeping $z_k \in Z_{\infty,2}$. Once this happens, g_k can not decay down to $\underline{g}$ again because as soon as $g_k < \check{g}$ (but with $g_k > \gamma_1 \check{g}$), it takes at most κ steps for z_k to reach $Z_{\infty,3}$ again while keeping $g_k < \check{g}$, and g_k can not go below $\underline{g}$ in κ steps. This conclusion contradicts the assumption made on the sequence $\{k_i\}$. Hence, $g_k \ge \underline{g}$ asymptotically. □

8.1.4 Illustrative Examples

In this section, we use two examples to illustrate the proposed dynamic scaling method.

Example 8.1 We consider a first order system:

$$x_{k+1} = ax_k + u_k, \quad y_k = x_k, \tag{8.39}$$

where $a > 1$. It turns out that we can have a relatively simple expression for N_0. Indeed, to stabilize the system using a logarithmic quantizer (8.1) with density ρ, the controller $H(z) = h$, where h is a constant, because of full state feedback. The closed-loop system is given by

$$x_{k+1} = (a + h(1 + \Delta_k))x_k, \quad |\Delta_k| \le \delta,$$

where δ relates to ρ as in (8.1). Since it is a first order system, we take $V(x_k) = x_k^2$, which gives

$$V(x_{k+1}) = (a + h(1 + \Delta_k))^2 x_k^2 \le (|a + h| + \delta|h|)^2 x_k^2$$

with the right-hand side being the worst-case value. Minimizing it gives $h = -a$ and

$$V(x_{k+1}) \le \delta^2 a^2 V(x_k).$$

This gives the upper bound for δ to be a^{-1}.

Now, for any $\delta < a^{-1}$, η in (8.7) is given by $\eta = 1 - \delta^2 a^2$. Applying it to (8.20), we obtain

$$N_0 = 1 + \frac{\log(\gamma_2 a^{-1} - \delta)}{\log(1 - \delta) - \log(1 + \delta)}, \ \delta < a^{-1}, \tag{8.40}$$

which can be minimized numerically. The result is shown in Fig. 8.1, where two curves for the required bit rate, one for $\gamma_2 = 1$ and another for $\gamma_2 = 0.9$, are compared with the minimum bit rate $\lceil \log_2(a) \rceil$ given in [2]. We see that the difference is only a few bits even when a is taken up to 100.

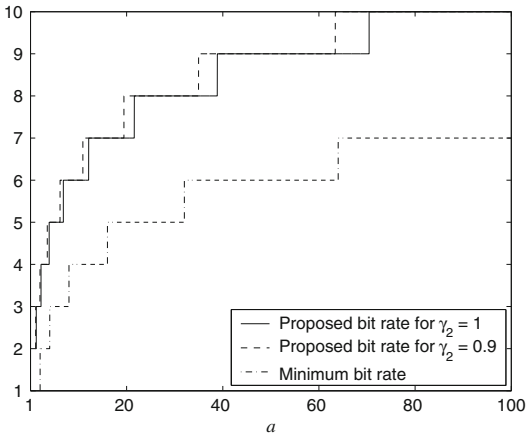

Fig. 8.1 Bit rate comparison for a first order system

Example 8.2 The second example we consider aims at demonstrating the convergence rate and robustness of the dynamic scaling method. Consider the system (7.1)–(7.2) with

$$A = \begin{bmatrix} 2.7 & -2.41 & 0.507 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$C = [1 \quad -0.5 \quad 0.04].$$

The system is unstable with two unstable open-loop poles at $1.2 \pm i0.5$ but without unstable zero and the relative degree is 1. It follows from [1] that

$$\delta_{\text{sup}} = |1.2 \pm i0.5|^{-2} = 0.5917, \quad \rho_{\text{inf}} = 0.2565.$$

By applying the search mentioned in Sect. 8.1, we obtain the optimal values:

$$\delta = 0.201, \ \eta = 0.561, \ \lambda_1 = 0.7223, \ \lambda_2 = 8.5046$$

The optimal controller is given by

$$A_c = \begin{bmatrix} -255.6834 & 46.7502 & 217.854 \\ 616.3274 & -111.8387 & -523.9270 \\ -431.7862 & 79.0425 & 368.0348 \end{bmatrix},$$

$$B_c = \begin{bmatrix} 5.8122 \\ -14.0003 \\ 9.8161 \end{bmatrix},$$

$$C_c = \begin{bmatrix} 81.6699 & -15.0325 & -69.6715 \end{bmatrix},$$

$$D_c = -1.8594.$$

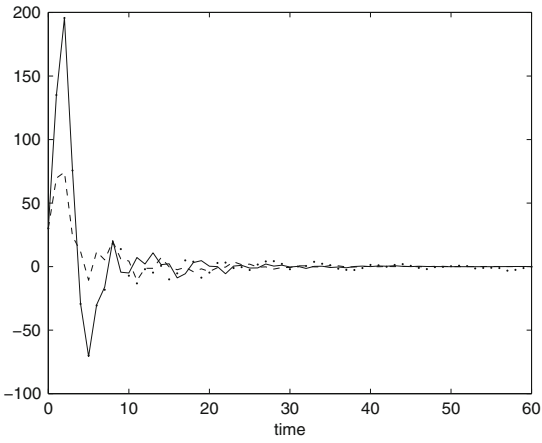

Fig. 8.2 The responses of x_1 of the closed-loop system with $N = 8$

Since γ_2 is lower bounded by $\sqrt{1-\eta} = 0.6626$, we choose $\gamma_2 = 0.8$. This gives $N_0 \approx 8$. We try $N = 8$ (4 bits). Note that the minimal bit rate required for stabilizing this system is one bit [2].

Next, it can be easily verified that (8.8) is satisfied if $\gamma_1 \leq 0.25$. Thus, we take $\gamma_1 = 0.2$. Let the initial state of the controller be $\hat{x}_0 = [0 \;\; 0 \;\; 0]^T$ and $\mu_0 = 1$. The response of the first state variable of the closed-loop system with the initial state $x_0 = [30 \;\; -30 \;\; 0]^T$, $g_0 = 0.1$ and $N = 8$ is shown in Fig. 8.2 (solid line). Other state variables are not shown since they are similar. If x_0 is known, we may set $g_0 = 1/|Cx_0|$. The response of the first state variable under this situation is also given in Fig. 8.2 (dash line) which as expected, shows a much reduced overshoot. We also examine the robustness of the closed-loop system. Let w_k in (8.29) be a saturated Gaussian white noise with zero mean, covariance matrix $Q_w = 3I$ and $\bar{w} = 100$. For $N = 8$, $g_0 = 0.1$, $\mu_0 = 1$, and $\bar{g} = 0.3$, the response of the first state variable of the closed-loop system with $x_0 = [30 \;\; -30 \;\; 0]^T$ is also shown in Fig. 8.2 (dot line) for comparison. The corresponding scaling gains g_k for the above three cases are compared in Fig. 8.3.

8.2 Attainability of the Minimum Data Rate for Stabilization

In the previous section, it was shown that a finite-level logarithmic quantizer with dynamical scaling can be applied to quadratically stabilize an unstable system. We also demonstrated in early chapter that a uniform quantizer can be applied to achieve the minimum data rate for stabilization. Observe that the data rate for achieving *quadratic* stabilization under logarithmic quantization is usually greater than the lower bound of (2.10), it would be of interest to know if a logarithmic quantizer can approach the minimum data rate if only stabilization is concerned. In this section, we shall study this issue.

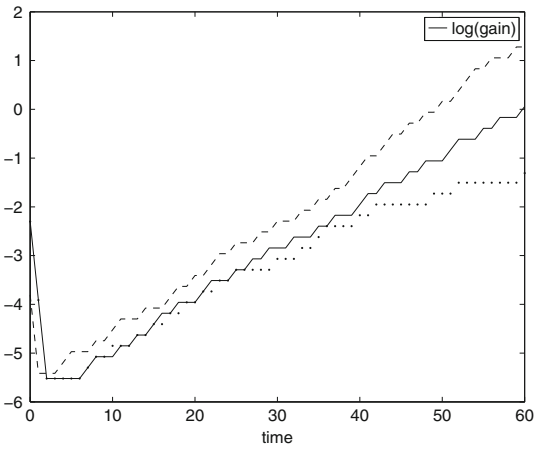

Fig. 8.3 The scaling factors g_k with $N = 8$

To this purpose, consider a discrete-time system as follows:

$$\begin{cases} x_{k+1} = Ax_k + Bu_k + w_k, \\ y_k \quad\;\; = Cx_k + v_k, \end{cases} \quad \forall k \in \mathbb{N}, \tag{8.41}$$

where $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^m$ is the control input, $y_k \in \mathbb{R}^\ell$ is the output, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^\ell$ are bounded additive disturbances. (A, B) and (C, A) are stabilizable and detectable pairs, respectively and $\text{rank}(B) = m \leq n$.

Our objective is to show that given an average data rate $R > H_T(A)$ of the feedback channel, we are able to design a finite-level quantizer to stabilize the networked linear systems (Fig. 8.4).

8.2.1 Problem Simplification

We perform the following techniques to simplify the presentation. Since we are concerned with the stabilization of linear systems, we can assume that all the eigenvalues of A lie outside or on the unit circle. Otherwise, the matrix A can be transformed to a block diagonal form $\text{diag}\{A_s, A_u\}$ by a coordinate transformation, where A_s and A_u respectively correspond to the stable and unstable (including marginally unstable) subspaces. State variables associated with the stable block A_s will converge to a bounded region for any bounded control sequence. Thus, without loss of generality, we assume that A has all eigenvalues lie outside or on the unit circle and (A, B, C) are controllable and observable.

Then, a deadbeat observer [3] can be constructed to estimate the state of the system. The estimation error will be uniformly bounded after n steps and independent of the initial state. Hence, it is sensible to focus on the state feedback case.

By applying the Wonham decomposition to (8.41) [3], one can convert the multiple inputs system to l single input ones. More specifically, there is a nonsingular real matrix $T \in \mathbb{R}^{n\times n}$ such that $\bar{A} = T^{-1}AT$ and $\bar{B} = T^{-1}B$ take the form:

$$\bar{A} = \begin{bmatrix} A_1 & A_{12} & \dots & A_{1m} \\ 0 & A_2 & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_m \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B_1 & B_{12} & \dots & B_{1m} \\ 0 & B_2 & \dots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_m \end{bmatrix},$$

where (A_i, B_i) with $A_i \in \mathbb{R}^{n_i\times n_i}$ and $B_i \in \mathbb{R}^{n_i}$, $i \in \{1, \dots, m\}$, is a controllable pair and $\sum_{i=1}^{\tau} n_i = n$. For illustration and brevity, let $l = 2$ and assume that the state feedback system is already given by

$$x_{k+1} = \bar{A}x_k + \bar{B}u_k + w_k.$$

By partitioning the state $x_k \triangleq [(x_k^1)', (x_k^2)']'$ in conformity with the upper triangular form of $\bar{A}$, two single input subsystems are written as

$$x_{k+1}^1 = A_1x_k^1 + B_1u_k^1 + A_{12}x_k^2 + B_{12}u_k^2 + w_k^1; \tag{8.42}$$

$$x_{k+1}^2 = A_2x_k^2 + B_2u_k^2 + w_k^2. \tag{8.43}$$

If x_k^2 is stabilized with a communication data rate greater than $H_T(A_2)$, then $\|x_k^2\|_\infty$ will be uniformly bounded and can be treated as a bounded disturbance input to the subsystem (4), which can be stabililized similarly with a data rate greater than $H_T(A_1)$. As in [2, Sect. 3], it is convenient to put A_2 into real Jordan canonical form so as to decouple its unstable dynamical modes.

Consequently, it is sufficient to focus on the following discrete linear time-invariant unstable system

$$x_{k+1} = Ax_k + Bu_k + w_k, \forall k \in \mathbb{N}, \tag{8.44}$$

where $x_k \in \mathbb{R}^n$ is the measurable state, $u_k \in \mathbb{R}$ is the control input, and $w_k \in \mathbb{R}^n$ is a uniformly bounded disturbance input, i.e., $\|w_k\|_\infty \le d$, $\forall k \in \mathbb{N}$, where $\|\cdot\|_\infty$ is the l^∞ norm for vectors or the induced matrix norm for matrices.

Moreover, $A \in \mathbb{R}^{n\times n}$ has two distinct real Jordan blocks, i.e.,

$$A = \mathrm{diag}(J_1, J_2),$$

where $J_i \in \mathbb{R}^{n_i\times n_i}$ corresponds to one unstable real eigenvalue $\lambda_i \in \mathbb{R}$ or a pair of unstable complex conjugate eigenvalues $\lambda_i, \lambda_i^* \in \mathbb{C}$ and $|\lambda_1| \ne |\lambda_2|$. Moreover, (A, B) is a controllable pair.

Differently from the finite-level quantizer in (8.1), we now use a $(2N+2)$-level logarithmic quantizer with density $\rho \in (0,\ 1)$ as follows:

$$\mathrm{Q}_N(v) = \begin{cases} \rho^i(1-\delta), & \text{if } \rho^{i+1} < v \le \rho^i,\ 0 \le i \le N-1; \\ 0, & \text{if } 0 \le v \le \rho^{N-1}; \\ -\mathrm{Q}_N(-v), & \text{if } -1 \le v < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{8.45}$$

In the above, we have chosen

$$u_0 = \frac{2\rho}{1+\rho}$$

in (8.1) and for any $v \notin [-1, 1]$, the alarm level 1 is introduced to indicate the saturation of the quantizer. Thus, the number of bits to represent each quantizer output is $\lceil \log_2(2N+2) \rceil$, where $\lceil \cdot \rceil$ is the standard ceiling function, i.e.,

$$\lceil x \rceil = \min\{l \in \mathbb{Z} | l \ge x\}.$$

8.2.2 Network Configuration

Two basic network configurations shown in Fig. 8.4 are to be studied. **Configuration I** refers to the scenario where the downlink channel has limited bandwidth while in **configuration II**, the uplink channel has limited bandwidth. Thus, the output of the controller in **Configuration I**, which is a scalar for the system (8.44), is to be quantized. In **configuration II**, the vector state measurement is quantized. The encoder/decoder pair for the limited data rate communication is described in Fig. 8.5. The first stage of the encoding process consists of designing a scaling factor g^{-1} such

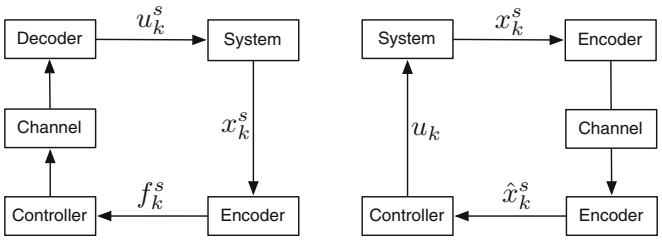

Fig. 8.4 **Configuration** I (*left*) versus **Configuration** II (*right*)

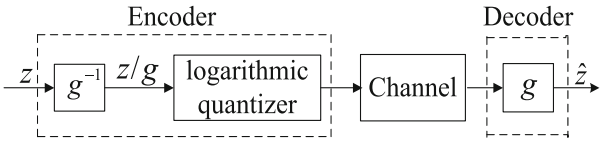

Fig. 8.5 Encoder/decoder pair for a digital channel: $g > 0$ is a scaling factor

that the quantizer input z/g lies in the quantization range. The output of the finite-level logarithmic quantizer $Q_N(z/g)$, which takes values from the set

$$\{\pm\rho^i(1-\delta) : i = 0, \ldots, N-1\} \cup \{0, 1\},$$

is encoded into binary sequence and transmitted via a communication channel with limited data rate. The decoder receives the binary sequence and correctly decodes it as $Q_N(z/g)$ since we neglect transmission errors of the channel. The quantizer output is then scaled back by g, i.e., $\hat{z} = gQ_N(z/g)$ to recover z if $Q_N(z/g) \neq 1$. In **Configuration I**, z and $\hat{z}$ respectively correspond to f_k^s and u_k^s. While in **Configuration II**, z is a vector and corresponds to x_k^s, which is the state of the system after down sampling. Thus, the quantizer in Fig. 8.5 is a product quantizer and consists of n finite-level logarithmic quantizers. The above notations will be defined in the sequel. Note that there is no separate channel to communicate the gain value g. The main task is to jointly design the scaling factor g, the finite-level logarithmic quantizer and the corresponding control law to approach the minimum average data rate of the channel for stabilizing the unstable system in (8.44). We mention that an earlier attempt has been made on scalar systems under **Configuration I** in [4].

Remark 8.1 The two configurations differ in the way that **Configuration II** quantizes the state first and use the quantized state to construct the control signal whereas in **Configuration I**, the control signal is constructed using the un-quantized state and then quantized by a finite-level logarithmic quantizer. From the information preservation point of view, **Configuration II** appears to generate worse control actions because quantization (or information loss) happens earlier. However, what we show in the chapter is that for the purpose of stabilization, the two configurations require the same minimum average data rate, if variable rate logarithmic quantization is used.

Remark 8.2 The two configurations have been widely adopted in literature. For example, [1, 5, 6] focus on **Configuration I** while [2, 7, 8] are restricted to **Configuration II**. The differences in the present chapter are that the quantizer in the encoder of Fig. 8.5 is limited to a finite-level logarithmic quantizer and we aim to approach the minimum average data rate of the channel for stabilizing the system (8.44).

In this section, we shall design finite-level logarithmic quantizers and the corresponding control laws to approach the minimum average data rate for stabilizing the unstable system in (8.44) under **configuration I** and **configuration II**, respectively.

8.2.3 Quantized Control Feedback

Theorem 8.4 *Consider the system in* (8.44) *and network* **configuration I** *of Fig.* 8.4*, stabilization can be achieved based on quantized control feedback with a finite-level logarithmic quantizer if and only if the average data rate R of the channel exceeds* $R_{\min}$*, i.e.,*

$$R > n_1 \log_2 |\lambda_1| + n_2 \log_2 |\lambda_2|.$$

Before giving the proof, the controller and quantizer are first proposed. Note that $|\lambda_1| \neq |\lambda_2|$, define the subset $\mathscr{L}(A) \subset \mathbb{N}$ by

$$\mathscr{L}(A) = \begin{cases} \mathbb{N}, & \text{if } \lambda_1, \lambda_2 \in \mathbb{R}, \\ \{i \in \mathbb{N} | \lambda_1^i \neq (\lambda_1^*)^i\}, & \text{if } \lambda_1 \in \mathbb{C}, \lambda_2 \in \mathbb{R}; \\ \{i \in \mathbb{N} | \lambda_2^i \neq (\lambda_2^*)^i\}, & \text{if } \lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{C}; \\ \{i \in \mathbb{N} | \lambda_j^i \neq (\lambda_j^*)^i, j = 1, 2\}, & \text{otherwise.} \end{cases}$$

Obviously, $\mathscr{L}(A)$ has infinitely many elements. Since (A, B) is a controllable pair, it is readily verified that $(A^\tau, A^{\tau-1}B)$ is a controllable pair if $\tau \in \mathscr{L}(A)$. By applying the control input $u_{\tau k+t} = 0$, if $1 \leq t \leq \tau - 1$, the down-sampled system of (8.44) with a down-sampling factor τ is expressed as

$$x_{\tau(k+1)} = A^\tau x_{\tau k} + A^{\tau-1} B u_{\tau k} + d_k, \tag{8.46}$$

where

$$d_k = \sum_{t=0}^{\tau-1} A^{\tau-1-t} w_{\tau k+t}.$$

Due to the controllability of $(A^\tau, A^{\tau-1}B)$, (8.46) can be transformed into a controllable canonical form, i.e., there exists a nonsingular real matrix $P \in \mathbb{R}^{n \times n}$ that transforms (8.46) into the controllable canonical form:

$$x_{k+1}^s = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \dots & -\alpha_1 \end{bmatrix} x_k^s + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u_k^s + w_k^s. \tag{8.47}$$

Here we denote $x_k^s \triangleq P x_{\tau k}$, $u_k^s \triangleq u_{\tau k}$ and $w_k^s \triangleq P d_k$.

It is clear from (8.47) that if we can stabilize the last element of the vector state x_k^s, denoted by $x_k^s(n)$, then x_k^s is stabilized, which further implies the stabilization of (8.44) due to $\tau < \infty$. Thus, a deadbeat controller is proposed whose output is then quantized by a finite-level logarithmic quantizer and applied to the down-sampled system.

Specifically, the quantized control input to the down-sampled system is given by

$$\begin{cases} u_k^s = gQ_N(f_k^s/g); \\ f_k^s = \begin{cases} 0, & \text{if } k < n; \\ \sum_{j=0}^{n-1} \alpha_{j+1} x_{k-j}^s(n), & \text{if } k \geq n, \end{cases} \end{cases} \tag{8.48}$$

where the quantization level parameter N and scaling factor $g > 0$ are to be designed.

Denote $|A| = |\lambda_1|^{n_1}|\lambda_2|^{n_2}$, it follows that there exists an $\alpha_0 > 0$ such that

$$|\alpha_k| \leq \alpha_0 |A|^\tau, \forall k \in \{1, \ldots, n\}$$

since $|\lambda_j| \geq 1, \forall j \in \{1, 2\}$. Given any $n \geq 1$, $\beta_1, \beta_2 \geq 0$ and $\beta_1 + \beta_2 > 0$, define

$$\kappa(\tau, \lambda) = (\beta_1 \tau^{n-1} + \beta_2)|\lambda|^\tau,$$

we have the following result.

Lemma 8.3 *$\forall \alpha > 1$, $\forall \varepsilon > 0$ and $|\lambda| \geq 1$, there exist positive integers τ and N such that*

$$\begin{aligned} \log_2 \left[1 + \frac{2 \log_2 \kappa(\tau, \lambda)}{\log_2 \frac{\kappa(\tau,\lambda)+\varepsilon+1}{\kappa(\tau,\lambda)+\varepsilon-1}} \right] &< \log_2(2N+2) \\ &\leq \tau \log_2 \alpha + \log_2 |\lambda|^m - 1. \end{aligned} \tag{8.49}$$

Proof It is trivial if $|\lambda| = 1$ and $\beta_1 = 0$. Assume $|\lambda| > 1$ or $\beta_1 > 0$, then $\kappa(\tau, \lambda) \to \infty$ as $\tau \to \infty$. Jointly with the fact that

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = e$$

yields

$$\log_2 \frac{\kappa(\tau, \lambda) + \varepsilon + 1}{\kappa(\tau, \lambda) + \varepsilon - 1} \cong 2\kappa^{-1}(\tau, \lambda) \log_2 e, \tag{8.50}$$

where m is sufficiently large. Next, two cases are discussed.

Case 1: $\beta_1 > 0$, $\beta_2 \geq 0$ and $|\lambda| \geq 1$.

Selecting a large $m \geq 1$ such that $\ln \kappa(\tau, \lambda) \geq 1$, we have

$$\begin{aligned} \left(1 + \kappa(\tau, \lambda) \ln \kappa(\tau, \lambda)\right)^{1/\tau} &= \left(1 + (\beta_1 \tau^{n-1} + \beta_2)|\lambda|^\tau \ln \kappa(\tau, \lambda)\right)^{1/\tau} \\ &\geq |\lambda| \beta_1^{1/\tau} (\tau^{1/\tau})^{n-1} (\ln \kappa(\tau, \lambda))^{1/\tau} \\ &\geq |\lambda| \beta_1^{1/\tau} \to |\lambda| \text{ as } \tau \to \infty \end{aligned}$$

due to that $\lim_{\tau\to\infty} x^{1/\tau} = 1, \forall x > 0$. On the other hand, choosing a large τ such that $\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda) \geq 1$, $\beta_1\tau^n \geq \beta_2$ and $2\beta_1\tau^n \geq \ln(2\beta_1\tau^n)$ gives the following inequalities:

$$\begin{aligned}\big(1+\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda)\big)^{1/\tau} &\leq \big(2\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda)\big)^{1/\tau}\\ &\leq |\lambda|(4\beta_1)^{1/\tau}(\tau^{1/\tau})^n[\tau\ln|\lambda| + \ln(2\beta_1\tau^n)]^{1/\tau}\\ &\leq |\lambda|(4\beta_1)^{1/\tau}(\tau^{1/\tau})^n[\tau^n\ln|\lambda| + 2\beta_1\tau^n]^{1/\tau}\\ &= |\lambda|[4\beta_1(\ln|\lambda| + 2\beta_1)]^{1/\tau}(\tau^{1/\tau})^{2n}\\ &\to |\lambda| \text{ as } \tau\to\infty,\end{aligned}$$

due to that $\lim_{\tau\to\infty}\tau^{1/\tau} = 1$.

Case 2: $\beta_1 = 0$, $\beta_2 > 0$ and $|\lambda| > 1$.

Let $\tau \geq 1$, it immediately follows that

$$\begin{aligned}\big(1+\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda)\big)^{1/\tau} &\geq |\lambda|\beta_2^{1/\tau}(\ln\beta_2 + \tau\ln|\lambda|)^{1/\tau}\\ &\geq |\lambda|\beta_2^{1/\tau}(\ln\beta_2 + \ln|\lambda|)^{1/\tau}\\ &\to |\lambda| \text{ as } \tau\to\infty.\end{aligned}$$

Also, for a sufficiently large τ, e.g., $\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda) \geq 1$ and $m \geq \ln\beta_2/\ln|\lambda|$, one can establish that

$$\begin{aligned}\big(1+\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda)\big)^{1/\tau} &\leq (2\beta_2|\lambda|^\tau\ln(\beta_2|\lambda|^\tau))^{1/\tau}\\ &\leq |\lambda|(4\beta_2\ln|\lambda|)^{1/\tau}\tau^{1/\tau}\\ &\to |\lambda| \text{ as } \tau\to\infty.\end{aligned}$$

Consequently, under any situation, we derive the limit

$$\lim_{\tau\to\infty}\big(1+\kappa(\tau,\lambda)\ln\kappa(\tau,\lambda)\big)^{1/\tau} = |\lambda|.$$

In the light of (8.50), it is clear that

$$\lim_{\tau\to\infty}\left[1+\frac{2\log_2\kappa(\tau,\lambda)}{\log_2\frac{\kappa(\tau,\lambda)+\varepsilon+1}{\kappa(\tau,\lambda)+\varepsilon-1}}\right]^{\frac{1}{\tau}} = |\lambda|, \tag{8.51}$$

which further implies that for a sufficiently large τ,

$$\log_2\left[1+\frac{2\log_2\kappa(\tau,\lambda)}{\log_2\frac{\kappa(\tau,\lambda)+\varepsilon+1}{\kappa(\tau,\lambda)+\varepsilon-1}}\right] \cong \log_2|\lambda|^\tau.$$

Since $\alpha > 1$, $\tau \log_2 \alpha \to \infty$ as $\tau \to \infty$, the difference between the left hand side and the right hand side of (8.49) tends to infinity if $\tau \to \infty$. Thus, it is always possible to select τ and N to satisfy (8.49). □

Proof of Theorem 8.4 The necessity part has been well established in [2, 7]. Only the sufficiency needs to be elaborated.

First, note that given any $R > \log_2 |A|$, there exists an $\alpha > 1$ satisfying $R \geq \log_2(\alpha|A|)$. Based on Lemma 8.3 and by choosing $\beta_1 = 0$, $\beta_2 = n\alpha_0$, it is possible to select a pair of $\tau \in \mathscr{L}(A)$ and $N > 0$ such that $\forall \varepsilon > 0$,

$$\log_2 \left[1 + \frac{2\log_2(n\alpha_0|A|^\tau)}{\log_2 \frac{n\alpha_0|A|^\tau+\varepsilon+1}{n\alpha_0|A|^\tau+\varepsilon-1}} \right] < \log_2(2N+2) \leq \tau \log_2 \alpha + \log_2 |A|^\tau - 1. \tag{8.52}$$

The quantizer level parameter N is determined by (8.52) and the number of bits required to represent each quantizer output is $\lceil \log_2(2N+2) \rceil$. The quantizer works as follows. At time k, the quantizer first detects the overflow of $x_k^s(n)$ and then proceeds to detect the overflow of f_k^s/g. Precisely, if $|x_k^s(n)| > \triangle$ is detected, it generates the alarm level 1 and in this case there is no need to further check f_k^s/g. Here the parameters g and $\triangle$ are to be determined later. Otherwise, it continues to check f_k^s/g. If $|f_k^s/g| > 1$ is detected, the quantizer generates the alarm level 1. Thus, the alarm level 1 will be generated if either $|x_k^s(n)| > \triangle$ or $|f_k^s/g| > 1$.

It is verified from (8.52) that the average data rate of this protocol satisfies

$$\frac{\lceil \log_2(2N+2) \rceil}{\tau} \leq \log_2(\alpha|A|) \leq R.$$

Since R is any given number greater than $\log_2 |A|$, the average data rate of the proposed quantizer can be made arbitrarily close to $\log_2 |A|$. Thus, what remains to be proved is the stability.

Mathematical induction arguments are adopted to show that

$$\lim_{k\to\infty} \sup |x_k^s(n)| < \infty$$

for any given initial condition.

First, assume $|x_k^s(n)| \leq \triangle, \forall k \in \{0, \ldots, n-1\}$, which will be relaxed later. Then, for any $k \geq n$, assume that $|x_j^s(n)| \leq \triangle, \forall j \leq k$, it is obvious that $\forall j \in \{n, n+1, \ldots, k\}$,

$$|f_j^s| = |\sum_{t=0}^{n-1} \alpha_{t+1} x_{j-t}^{(n)}| \leq n\alpha_0|A|^\tau \triangle \triangleq g.$$

Since $|f_j^s/g| \leq 1$ and $|x_j^s(n)| \leq \triangle, \forall j \leq k$, no alarm level 1 occurs before time k. From (8.47), there exist vectors $c_j \in \mathbb{R}^n, j \in \{0, \ldots, n-1\}$ such that

$$s_k^s = \sum_{j=0}^{n-1} c_j^T w_{k-j}^s$$

and the down-sampled system is expressed by

$$x_{k+1}^s(n) = -\sum_{j=0}^{n-1} \alpha_{j+1} x_{k-j}^s(n) + u_k^s + s_k^s, k \geq n. \tag{8.53}$$

Moreover,

$$|s_k^s| \leq \sum_{j=0}^{n-1} \|c_j^T\|_\infty \|w_{k-j}^s\|_\infty \triangleq \tilde{s}, \forall k \in \mathbb{N}.$$

Choose the quantizer density parameters

$$\delta = \frac{1}{n\alpha_0|A|^\tau + \varepsilon}, \rho = \frac{1-\delta}{1+\delta}$$

and $\triangle > 0$ to satisfy that

$$\triangle > \max\{\frac{\tilde{s}}{1-(n\alpha_0|A|^\tau)^2\rho^{2N+1}}, \frac{\tilde{s}}{1-n\alpha_0|A|^\tau\delta}\}. \tag{8.54}$$

In light of (8.52), it is easy to verify that

$$\begin{cases} (n\alpha_0|A|^\tau)^2\rho^{2N+1} < 1 \\ n\alpha_0|A|^\tau\delta < 1. \end{cases} \tag{8.55}$$

Inserting the quantized control in (8.48) into the system in (8.53) results in that

$$\begin{aligned} |x_{k+1}^s(n)| &\leq \begin{cases} |f_k^s| + \tilde{s}, & \text{if } |f_k^s/g| \leq \rho^N \\ \delta|f_k^s| + \tilde{s}, & \text{if } \rho^N < |f_k^s/g| \leq 1 \end{cases} \\ &\leq \begin{cases} n\alpha_0|A|^\tau\rho^N\triangle + \tilde{s}, & \text{if } |f_k^s/g| \leq \rho^N \\ n\alpha_0|A|^\tau\delta\triangle + \tilde{s}, & \text{if } \rho^N < |f_k^s/g| \leq 1 \end{cases} \\ &\leq \triangle \text{ due to the selection of } \triangle \text{ in (8.54) and (8.55).} \end{aligned}$$

Inductively, $|x_k^s(n)| \leq \triangle$ for all $k \in \mathbb{N}$.

Next, suppose that

$$|x_j^s(n)| \leq \triangle, \forall j \in \{0, \ldots, n-1\}$$

is violated, which can be detected by the decoder via the alarm level. Denote the *first* time of receiving the alarm level 1 by k. Choose a scaling factor

$$\gamma = \sum_{j=1}^{n} \alpha_j + 1 \tag{8.56}$$

to dynamically update the scaling factor. Specifically, set $\triangle_k = \triangle$ and update the scaling factor as follows:

$$\triangle_{k+j+1} = \begin{cases} \gamma \triangle_{k+j}, & \text{alarm level 1 occurs,} \\ \triangle_{k+j}, & \text{otherwise,} \end{cases}$$

which is simultaneously processed on the both sides of the channel. Set $u^s_{k+j} = 0$, the increasing speed of $\triangle_{k+j}$ is thus faster than that of $x^s_{k+j}(n)$ by (8.47) and (8.56). $\triangle_{k+j}$ will eventually capture $x^s_{k+j}(n)$ or

$$\lim_{j\to\infty} \frac{x^s_{k+j}(n)}{\triangle_{k+j}} = 0.$$

Let the scaling factor be

$$g_{k+j} = n\alpha_0 |A|^\tau \triangle_{k+j},$$

it follows that

$$\lim_{j\to\infty} f^s_{k+j}/g_{k+j} = 0,$$

implying that there exists a finite $k_0 \geq n-1$ such that the signals received within the time period $\{k+k_0-n+1, \ldots, k+k_0\}$ do not give rise to the alarm level 1, which suggests that

$$|x^s_{k+j}(n)| \leq \triangle_{k+k_0}, \forall j \in \{k_0-n+1, \ldots, k_0\}.$$

Then, repeating the above proof as the bounded case at time $k+k_0+n-1$ yields that

$$|x^s_{k+j}(n)| \leq \triangle_{k+k_0}, \forall j \geq k_0+n-1.$$

Finally, it follows that $\limsup_{k\to\infty} |x^s_k(n)| < \infty$, which eventually leads to that $\limsup_{k\to\infty} \|x_k\|_\infty < \infty$. □

The following corollary gives the corresponding result for asymptotic stabilization, i.e., $\lim_{k\to\infty} \|x_k\|_\infty = 0$.

Corollary 8.2 *Consider the system in* (8.44) *with* $w_k = 0$ *and network* **configuration I** *of Fig.* 8.4*, asymptotic stabilization can be achieved via a quantized control feedback with a finite-level logarithmic quantizer if and only if the average data rate* R *of the channel is strictly greater than* $R_{\min}$*, i.e.,*

$$R > n_1 \log_2 |\lambda_1| + n_2 \log_2 |\lambda_2|.$$

Proof Similarly, only the sufficiency part needs to be established. Define a scaling factor

$$\eta \triangleq \max\{(n\alpha_0|A|^\tau)^2 \rho^{2N+1}, n\alpha_0|A|^\tau \tau\delta\}, \tag{8.57}$$

which is strictly less than one by (8.55), i.e., $\eta < 1$. Let $\triangle_{k+1} = \eta\triangle_k$ with an *arbitrary* $\triangle_0 > 0$, which is assumed to be agreed by both the quantizer and the decoder.

Assume that

$$|x_k^s(n)| \leq \triangle_0, \forall k \in \{0, \ldots, n-1\},$$

the control input and quantizer are given in Theorem 8.4 with $\triangle$ replaced by $\triangle_k$. Then, it is straightforward that

$$|x_{k+1}^s(n)| \leq \eta\triangle_k = \triangle_{k+1}.$$

Thus, $x_k^s(n)$ can be driven exponentially to zero since

$$\lim_{k\to\infty} |x_k^s(n)| \leq \triangle_0 \lim_{j\to\infty} \eta^j = 0.$$

Due to $m < \infty$, it follows that $\lim_{k\to\infty} \|x_k\|_\infty = 0$. The removal of the boundedness assumption for the initial state is similar to what we have done in Theorem 8.4. □

8.2.4 Quantized State Feedback

We proceed to validate the attainability of the minimum average data rate under **configuration II** via logarithmic quantization where the control design solely relies on the quantized state. Intuitively, this might require a larger average data rate since the quantized state contains less information than its unquantized version. However, the result of this section shows that the logarithmic quantizer can still approach the minimum average data rate.

Theorem 8.5 *Consider the system in* (8.44) *and network* **configuration II** *of Fig.* 8.4 *stabilization can be achieved based on the quantized state feedback with a finite-level logarithmic quantizer if and only if the average data rate R of the channel exceeds* $R_{\min}$*, i.e.,*

$$R > n_1 \log_2 |\lambda_1| + n_2 \log_2 |\lambda_2|.$$

In this case, two scalar logarithmic quantizers with appropriately chosen parameters are designed and applied to the down-sampled state x_k^s of (8.44), where the down-sampling factor $m \geq 2n$ is to be determined later. More precisely, index the scalar components of the state of (8.44) by an additional superscript $h \in \{1, \ldots, n\}$.

At time $t = \tau k + h - 1$, the hth element of x_k^s will be quantized by

$$Q_{N_1}(x_k^s(h)/\triangle)$$

if $h \leq n_1$ and $Q_{N_2}(x_k^s(h)/\triangle)$ otherwise, where the quantization level parameter N_i and $\triangle$ are determined by the available data rate. Neglecting the transmission time implies that the quantized x_k^s can reach the controller before time $\tau k + n$. Since $\tau \geq 2n$, the

control law within one cycle $\{\tau k, \ldots, \tau(k+1)-1\}$ can be proposed as follows:

$$\begin{cases} \begin{bmatrix} u_{\tau k+\tau-1} \\ \vdots \\ u_{\tau k+\tau-n} \end{bmatrix} = -\triangle \mathscr{C}^T (\mathscr{C}\mathscr{C}^T)^{-1} A^\tau Q(\frac{x_k^s}{\triangle}), \\ u_{\tau k+t} = 0, \forall t \in \{0, \ldots, \tau-n-1\}, \end{cases} \tag{8.58}$$

where the controllability matrix $\mathscr{C}$ is defined as

$$\mathscr{C} \triangleq [B, AB, \ldots, A^{n-1}B]$$

and the product quantizer $Q(\cdot)$ is composed by

$$Q(\cdot) = [\underbrace{\mathrm{Q}_{N_1}(\cdot), \ldots, \mathrm{Q}_{N_1}(\cdot)}_{n_1}, \underbrace{\mathrm{Q}_{N_2}(\cdot), \ldots \mathrm{Q}_{N_2}(\cdot)}_{n_2}]^T.$$

Proof of Theorem 8.5 As in the case of **configuration I**, only the sufficiency part requires to be proved.

Given any

$$R > n_1 \log_2 |\lambda_1| + n_2 \log_2 |\lambda_2|,$$

there exists an $\alpha > 1$ satisfying $R = R_1 + R_2$ and

$$R_i \geq n_i \log_2(\alpha|\lambda_i|), \forall i \in \{1, 2\}.$$

In view of Lemma 8.3 for any $\varepsilon > 0$, we can choose a pair of integers $\tau \geq 2n$ and N_i satisfying:

$$\log_2 \left[1 + \frac{2\log_2 \zeta \sqrt{n_i} \tau^{n_i-1} |\lambda_i|^\tau}{\log_2 \frac{\zeta \sqrt{n_i} \tau^{n_i-1} |\lambda_i|^\tau + \varepsilon + 1}{\zeta \sqrt{n_i} \tau^{n_i-1} |\lambda_i|^\tau + \varepsilon - 1}} \right] < \log_2(2N_i + 2)$$
$$\leq \tau \log_2 \alpha + \log_2 |\lambda_i|^\tau - 1, \forall i \in \{1, 2\}. \tag{8.59}$$

The quantization level parameter N_i is selected based on (8.59). The average data rate of this protocol is computed by

$$\frac{n_1 \lceil \log_2(2N_1 + 2) \rceil + n_2 \lceil \log_2(2N_2 + 2) \rceil}{\tau} \leq \log_2(\alpha|A|) \leq R,$$

which implies that the minimum average data rate can be approached by the above protocol. Also, the quantizer density parameters for $Q_{N_i}(\cdot)$ are chosen by

$$\delta_i = \frac{1}{\zeta \sqrt{n_i} \tau^{n_i-1} |\lambda_i|^\tau + \varepsilon}$$

and

$$\rho_i = \frac{1-\delta_i}{1+\delta_i} = \frac{\zeta\sqrt{n_i}\tau^{n_i-1}|\lambda_i|^\tau + \varepsilon - 1}{\zeta\sqrt{n_i}\tau^{n_i-1}|\lambda_i|^\tau + \varepsilon + 1},$$

which gives that

$$\begin{cases} (\zeta\sqrt{n_i}\tau^{n_i-1})^2\rho_i^{2N_i+1} < 1, \\ \delta_i\zeta\sqrt{n_i}\tau^{n_i-1} < 1. \end{cases} \tag{8.60}$$

Define the uniform upper bound of d_k in (8.46) by

$$D \triangleq d\sum_{t=0}^{\tau-1}\|A\|_\infty^{\tau-1-t},$$

then $\|d_k\|_\infty \le D, \forall k \in \mathbb{N}$.

Similarly, the initial state x_0 is assumed to be bounded by $\triangle > 0$, where $\triangle$ is selected to satisfy

$$\triangle \ge \max_{i\in\{1,2\}}\{\frac{D}{1-(\zeta\sqrt{n_i}\tau^{n_i-1})^2\rho_i^{2N_i+1}}, \frac{D}{1-\delta_i\zeta\sqrt{n_i}\tau^{n_i-1}}\}. \tag{8.61}$$

Inserting the control law in (8.58) into (8.44) yields that

$$\begin{aligned} x_{k+1}^s &= A^\tau x_k^s + \sum_{t=0}^{\tau-1} A^{\tau-1-t}(Bu_{\tau k+t} + w_{\tau k+t}) \\ &= A^\tau x_k^s + \sum_{t=\tau-n}^{\tau-1} A^{\tau-1-t}Bu_{\tau k+t} + d_k \\ &= A^\tau[x_k^s - \triangle Q(\frac{x_k^s}{\triangle})] + d_k. \end{aligned} \tag{8.62}$$

Assuming that $\|x_k^s\|_\infty \le \triangle$, there is no alarm level 1 for the scaled state $x_k^s(h)$ by the scaling factor $\triangle$. Denote $(x_k^s)^{(1)}$ the state vector consisting of the first n_1 elements of x_k^s while $(x_k^s)^{(2)}$ is the state vector by collecting the remaining elements of x_k^s. Similar notations will be made for $Q^{(i)}$ and $d_k^{(i)}$, $i \in \{1, 2\}$. Consider the system of (8.62), it follows from Lemma 3.1 that:

$$\begin{aligned} \|(x_{k+1}^s)^{(i)}\|_\infty &= \|J_i^\tau[(x_k^s)^{(i)} - \triangle Q^{(i)}(\frac{(x_k^s)^{(i)}}{\triangle})] + d_k^{(i)}\|_\infty \\ &\le \|J_i^\tau\|_\infty\|(x_k^s)^{(i)} - \triangle Q^{(i)}(\frac{(x_k^s)^{(i)}}{\triangle}))\|_\infty + D \\ &\le \begin{cases} \|J_i^\tau\|_\infty\|(x_k^s)^{(i)}\|_\infty + D, & \text{if } \|(x_k^s)^{(i)}/\triangle\|_\infty \le \rho_i^{N_i-1} \\ \|J_i^\tau\|_\infty\delta_i\|(x_k^s)^{(i)}\|_\infty + D, & \text{if } \rho_i^{N_i-1} < \|(x_k^s)^{(i)}/\triangle\|_\infty \le 1 \end{cases} \end{aligned}$$

$$\leq \begin{cases} \zeta\sqrt{n_i}\tau^{n_i-1}\rho_i^{N_i-1}\triangle + D, & \text{if } \|(x_k^s)^{(i)}/\triangle\|_\infty \leq \rho_i^{N_i-1} \\ \zeta\sqrt{n_i}\tau^{n_i-1}\delta_i\triangle + D, & \text{if } \rho_i^{N_i-1} < \|(x_k^s)^{(i)}/\triangle\|_\infty \leq 1. \end{cases}$$

$\leq \triangle$ by (8.60) and (8.61).

Inductively, $\|x_k^s\|_\infty \leq \triangle, \forall k \in \mathbb{N}$. Since $\tau < \infty$, it follows that $\limsup_{k\to\infty} \|x_k\|_\infty < \infty$.

The removal of the boundedness assumption of $\|x_0^s\|_\infty \leq \triangle$ is similar to the case of **configuration I** and is omitted. □

Remark 8.3 It is worth mentioning that the parallel result in Corollary 8.2 can be given under **configuration II**. Furthermore, the attainability of the logarithmic quantization can also be established for the case where quantization appears in both the state measurement and the control signal.

8.3 Summary

In this chapter, we have studied two problems concerning logarithmic quantization. The first is the design of a dynamic finite level logarithmic quantizer for quadratically stabilizing an unstable plant. The other is the attainability of the minimum average data rate via logarithmic quantization for stabilizing an unstable discrete-time linear system. For any average data rate greater than the minimum rate given by the data rate theorem, a finite-level logarithmic quantizer and a controller were constructed to stabilize the system under two different network configurations with different schemes of quantizer bits assignment. It should be noted that since our main concern is the attainability of the minimum average data rate by logarithmic quantization, the proposed control law and quantizer may produce a poor transient response.

References

1. M. Fu, L. Xie, The sector bound approach to quantized feedback control. IEEE Trans. Autom. Control **50**(11), 1698–1711 (2005)
2. G. Nair, R. Evans, Stabilizability of stochastic linear systems with finite feedback data rates. SIAM J. Control Optim. **43**(2), 413–436 (2004)
3. C. Chen, *Linear System: Theory and Design* (Saunders College Publishing, Philadelphia, 1984)
4. M. Fu, L. Xie, W. Su, Connections between quantized feedback control and quantized estimation, in *10th International Conference Control, Automation, Robotics and Vision* (2008)
5. N. Elia, S. Mitter, Stabilization of linear systems with limited information. IEEE Trans. Autom. Control **46**(9), 1384–1400 (2001)
6. K. Tsumura, H. Ishii, H. Hoshina, Tradeoffs between quantization and packet loss in networked control of linear systems. Automatica **45**(12), 2963–2970 (2009)
7. S. Tatikonda, S. Mitter, Control under communication constraints. IEEE Trans. Autom. Control **49**(7), 1056–1068 (2004)
8. R. Brockett, D. Liberzon, Quantized feedback stabilization of linear systems. IEEE Trans. Autom. Control **45**(7), 1279–1289 (2000)

Chapter 9
Stabilization of Markov Jump Linear Systems via Logarithmic Quantization

This chapter aims at stabilizing an unstable plant across a lossy channel via quantized feedback. Since a large class of NCSs with random packet dropouts can be modeled as Markov jump linear systems (MJLSs), we first consider the quantized stabilization problem for a single-input MJLS in Sect. 9.1. Given a measure of quantization coarseness, a mode-dependent logarithmic quantizer and a mode-dependent linear state feedback law can achieve optimal coarseness for mean square quadratic (MSQ) stabilization of an MJLS. The sector bound approach is shown to be non-conservative in investigating the corresponding quantized state feedback problem, and then a method of optimal quantizer and controller design is presented in terms of LMIs. Moreover, when the mode process is not directly observed by the controller, we give a mode estimation algorithm by maximizing a probability criterion. In Sect. 9.2, the results presented in Sect. 9.1 are applied to the quantized stabilization of NCSs over lossy channels with binary or bounded packet losses. Both TCP-like and UDP-like protocols are discussed, and an extension to output feedback under the TCP-like protocol is also included. Section 9.3 summarizes the chapter.

9.1 State Feedback Case

As we can see from Fig. 9.1, a quantized feedback control system generally comprises three parts: a system to be controlled, a controller and a quantizer.

Assume that the system is described by a discrete-time single-input MJLS as

$$x_{k+1} = A_{\theta_k} x_k + B_{\theta_k} u'_k + w_k, \tag{9.1}$$

where $x_k \in \mathbb{R}^n$ is the state with x_0 being a second-order random variable, $u'_k \in \mathbb{R}$ is the quantized control input, $w_k \in \mathbb{R}^n$ is a second-order process noise with zero mean and covariance matrix $\Sigma_{\theta_k} > 0$, and $\theta_k \in \Theta \triangleq \{0, 1, \ldots, N\}$ is the system mode governed by a time-homogeneous Markov chain with initial distribution

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_9

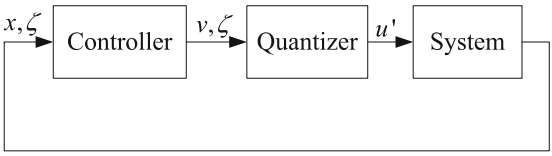

Fig. 9.1 Typical quantized feedback control system

$\pi = [\pi_0 \ \pi_1 \dots \pi_N]$ and transition probability matrix $\Pi = [\pi_{ij}]_{i,j\in\Theta}$, where

$$\pi_i \triangleq \Pr(\theta_0 = i), \quad \pi_{ij} \triangleq \Pr(\theta_{k+1} = j|\theta_k = i). \tag{9.2}$$

For all $k \geq 0$, x_0 is assumed to be independent of θ_0^k and w_0^k, where θ_0^k and w_0^k denote respectively the set $\{\theta_0, \theta_1, \dots, \theta_k\}$ and the set $\{w_0, w_1, \dots, w_k\}$. Suppose that x_k is available at both the controller and the quantizer, and the static quantized state feedback is denoted by

$$v_k = g(x_k, \zeta_k), \tag{9.3}$$
$$u_k' = f(v_k, \zeta_k), \tag{9.4}$$

where $\zeta_k \in \Theta$ is a direct observation or an estimate of system mode θ_{k-d} at the controller/quantizer side at time step k with $d \in \mathbb{Z}_+$ the constant mode observation/estimation delay. For every $k \in [0, d-1]$, the initial ζ_k is chosen arbitrarily from the mode set Θ.

The closed-loop system of (9.1), (9.3) and (9.4) is given by

$$x_{k+1} = A_{\theta_k} x_k + B_{\theta_k} f(g(x_k, \zeta_k), \zeta_k) + w_k. \tag{9.5}$$

It is worth mentioning that (9.5) is generally nonlinear, since the control signal u_k' can be a nonlinear function of v_k in (9.4) due to quantization. We adopt the following definitions of mean square stability and MSQ stability.

Definition 9.1 For $w_k \equiv 0$ and any initial condition of x_0, θ_0, ζ_0, the equilibrium point at the origin of (9.5) is mean square stable if

$$\lim_{k\to\infty} \mathscr{E}[\|x_k\|^2 | x_0, \theta_0, \zeta_0] = 0; \tag{9.6}$$

it is mean square quadratically (MSQ) stable, if, for every $\zeta_k \in \Theta$, there exist a function

$$V(x_k, \zeta_k) \triangleq x_k^T P_{\zeta_k} x_k, \quad P_{\zeta_k} > 0, \tag{9.7}$$

and a positive-definite matrix Q_{ζ_k} such that, for all $k \geq d$,

$$\begin{aligned}\nabla V(x_k, \zeta_k) &\triangleq \mathscr{E}[V(x_{k+1}, \zeta_{k+1}) - V(x_k, \zeta_k)|x_0^k, \zeta_0^k] \\ &= \mathscr{E}[V(x_{k+1}, \zeta_{k+1})|x_0^k, \zeta_0^k] - V(x_k, \zeta_k) \\ &< -x_k^T Q_{\zeta_k} x_k, \quad \forall x_k \in \mathbb{R}^n, \ x_k \neq 0. \end{aligned} \tag{9.8}$$

Remark 9.1 Following a similar line of arguments as in the proof of Theorem 1 in [1], we can prove that the MSQ stability of the equilibrium point at the origin of (9.5) always implies the mean square stability. When $d \geq 1$, after taking $u'_k \equiv 0$ for all $k < d$, it is easy to see that x_d is still a second-order random variable, and thus we can consider $k = d$ as the starting point in (9.8) without loss of generality.

Remark 9.2 Imposing the condition (9.8) in every system mode introduces some degree of conservativeness but has the following advantages: (1) under the notion of MSQ stability, we can prove the optimality of the logarithmic quantizer defined in the next section; (2) it makes existing well-established results in robust control of MJLSs applicable in quantized feedback control.

Assume that $f(\cdot, \cdot)$ in (9.4) is an odd function of v_k, i.e.,

$$f(-v_k, \zeta_k) = -f(v_k, \zeta_k). \tag{9.9}$$

We define the mode quantization density with respect to mode i, $i \in \Theta$, similarly to the quantization density in the LTI case (7.5), as

$$\eta_f(i) \triangleq \limsup_{\varepsilon \to 0} \frac{\#l[\varepsilon, i]}{-\ln \varepsilon}, \tag{9.10}$$

where $\#l[\varepsilon, i]$ is the number of quantization levels in the interval $[\varepsilon, 1/\varepsilon]$ for the quantizer $f(\cdot, i)$. Evidently, the mode quantization density is reduced to the quantization density defined in [2] when $N = 0$. For $N \neq 0$, there is a set of mode quantization densities $\eta_f(i)$, $i = 0, 1, \ldots, N$. Motivated by the Pareto optimality of vector-valued criterion [3], we introduce the overall coarseness for a mode-dependent quantizer as follows.

Definition 9.2 The overall coarseness of a mode-dependent quantizer (9.4) is defined as

$$\mathrm{C}_f \triangleq e(\eta_f(0), \eta_f(1), \ldots, \eta_f(N)), \tag{9.11}$$

where e is a scalar-valued function of $\eta_f(i)$, $i = 0, 1, \ldots, N$, satisfying the following property: if $\eta_{f1}(i) \leq \eta_{f2}(i)$ for all $i \in \Theta$, then

$$e(\eta_{f1}(0), \eta_{f1}(1), \ldots, \eta_{f1}(N)) \leq e(\eta_{f2}(0), \eta_{f2}(1), \ldots, \eta_{f2}(N)). \tag{9.12}$$

The property (9.12) reveals that the overall coarseness should always be nondecreasing when any one of the mode quantization densities is increasing and all the others are fixed. Note that the smaller the value of C_f, the coarser the quantizer.

The form of e in (9.11) can be chosen according to physical constraints or performance requirements of the quantizer. It is easy to see that the set of $\eta_f(i)$, $i = 0, 1, \ldots, N$, corresponding to the globally optimal C_f may not be unique. The main purpose of this section is to find one possible combination of the controller (9.3) and the quantizer (9.4) with the optimal C_f such that the closed-loop system is MSQ stable.

9.1.1 Feedback Stabilization

A mode-dependent quantizer is said to be logarithmic, if, for any $\zeta_k \in \Theta$, the corresponding set of quantization levels $\mathscr{Z}_{\zeta_k}$ has the following form:

$$\mathscr{Z}_{\zeta_k} = \{\pm u_l'(\zeta_k) : u_l'(\zeta_k) = \rho^l(\zeta_k)u_0', u_0' > 0, \quad \text{for } l \in \pm 1, \pm 2, \ldots\} \cup \{\pm u_0'\} \cup \{0\}, \tag{9.13}$$

where

$$\rho(\zeta_k) = \frac{1 - \delta(\zeta_k)}{1 + \delta(\zeta_k)}, \tag{9.14}$$

and $\delta(\zeta_k)$ denotes the sector bound as shown in Fig. 9.2. Specifically, the associated logarithmic quantizer is defined as follows.

For the given ζ_k:

- if $\delta(\zeta_k) = 0$, then

$$f(v_k, \zeta_k) = v_k; \tag{9.15}$$

- if $0 < \delta(\zeta_k) < 1$, then

$$f(v_k, \zeta_k) = \begin{cases} u_l'(\zeta_k), & \text{if } \frac{1}{1+\delta(\zeta_k)}u_l'(\zeta_k) < v_k \le \frac{1}{1-\delta(\zeta_k)}u_l'(\zeta_k), \\ 0, & \text{if } v_k = 0, \\ -f(-v_k, \zeta_k), & \text{if } v_k < 0; \end{cases} \tag{9.16}$$

- if $\delta(\zeta_k) = 1$, then

$$f(v_k, \zeta_k) = \begin{cases} u_0', & \text{if } v_k > \frac{1}{2}u_0', \\ 0, & \text{if } 0 \le v_k \le \frac{1}{2}u_0', \\ -f(-v_k, \zeta_k), & \text{if } v_k < 0. \end{cases} \tag{9.17}$$

There is no loss of generality by choosing the same u_0' for every $\zeta_k \in \Theta$; see Lemma 2.1 in [2]. For a logarithmic quantizer, it is easy to verify that

$$\eta_f(i) = -2/\ln \rho(i) \tag{9.18}$$

for every $i \in \Theta$. Thus, the coarser the quantizer for mode i, the smaller the $\eta_f(i) \in \mathbb{R}_+ \cup \{\infty\}$ and $\rho(i) \in [0, 1]$, or equivalently the larger the $\delta(i) \in [0, 1]$.

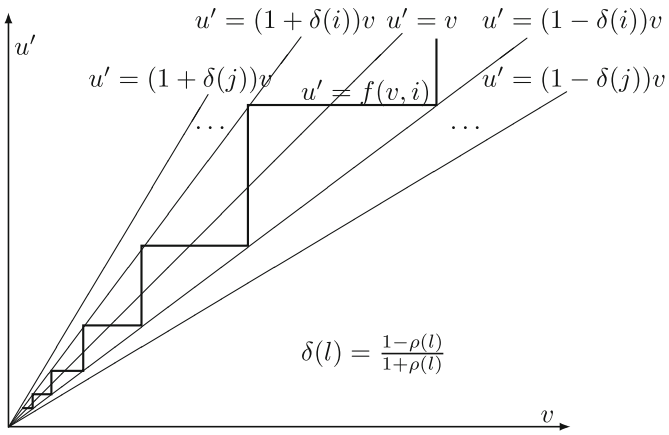

Fig. 9.2 Mode-dependent logarithmic quantizer

Since we mainly focus on global stabilization in this section, we let $w_k \equiv 0$ without loss of generality. The next assumption is essential to the existence of an optimal memoryless quantization strategy in the MSQ stability sense.

Assumption 9.1 (a) The system (9.1) is not mean square stable with $u'_k \equiv 0$ but can be mean square stabilized via a linear state feedback law:

$$u'_k = \bar{K}_{\zeta_k} x_k. \tag{9.19}$$

(b) For any $i_1, i_2, i_3 \in \Theta$ and $k \geq d$,

$$\Pr\{\theta_k = i_1, \zeta_{k+1} = i_2 | x_0^k, \zeta_0^{k-1}, \zeta_k = i_3\} = \Pr\{\theta_k = i_1, \zeta_{k+1} = i_2 | \zeta_k = i_3\}. \tag{9.20}$$

Moreover, the conditional probability on the right-hand side of (9.20), denoted by $q_{i_1 i_2 i_3}$, is constant over time and known to the controller/quantizer.

Remark 9.3 Assumption 9.1(a) clearly avoids triviality and imposes a necessary restriction for ensuring the solvability of the stabilization problem. A systematic way to find a stabilizing state feedback law for an MJLS can be found in [4]. Assumption 9.1(b) facilitates an explicit evaluation of (9.8) and covers several situations as collected in Sect. 9.1.2.

As the first result of this chapter, it will be shown that for a fixed set of

$$P_i > 0, Q_i > 0, \ i = 0, 1, \ldots, N,$$

the coarsest quantization in the sense of MSQ stability can be approached by a linear state feedback law and a logarithmic quantizer. To this end, for every $i \in \Theta$, let us define the row vector $\mathbf{a}_i$, the matrix F_i, and two scalars b_i, $\delta_m(i)$ as

$$\mathbf{a}_i \triangleq \sum_{i_1\in\Theta, i_2\in\Theta} \left[q_{i_1 i_2 i} {B_{i_1}}^T P_{i_2} A_{i_1} \right], \tag{9.21}$$

$$F_i \triangleq \sum_{i_1\in\Theta, i_2\in\Theta} \left[q_{i_1 i_2 i} {A_{i_1}}^T P_{i_2} A_{i_1} \right] \geq 0, \tag{9.22}$$

$$b_i \triangleq \sum_{i_1\in\Theta, i_2\in\Theta} \left[q_{i_1 i_2 i} {B_{i_1}}^T P_{i_2} B_{i_1} \right] \geq 0, \tag{9.23}$$

$$\delta_m(i) \triangleq \begin{cases} \infty, & \text{if } b_i = 0, \\ \frac{1}{\sqrt{K_{mi} M_i^{-1} K_{mi}^T}}, & \text{otherwise}, \end{cases} \tag{9.24}$$

where

$$K_{mi} \triangleq -\frac{\mathbf{a}_i}{b_i}, \tag{9.25}$$

$$M_i \triangleq \frac{\mathbf{a}_i^T \mathbf{a}_i}{b_i^2} - \frac{F_i - P_i + Q_i}{b_i}. \tag{9.26}$$

Theorem 9.1 *Consider the MSQ stabilization with a given set of*

$$P_i > 0, Q_i > 0,\ i = 0, 1, \ldots, N$$

in (9.8) *for the system* (9.1) *using quantized state feedback* (9.3) *and* (9.4)*. Then, under Assumption* 9.1*, the smallest* C_f *defined in* (9.11) *can be approached by a linear state feedback law* $v_k = K_{\zeta_k} x_k$ *and a logarithmic quantizer* (9.15)–(9.17) *with controller and quantizer parameters chosen as below:*

$$K_i = \begin{cases} 0, & \text{if } \delta_m(i) > 1, \\ K_{mi}, & \text{otherwise}, \end{cases} \qquad \delta(i) = \begin{cases} 1, & \text{if } \delta_m(i) > 1, \\ \delta_m(i), & \text{otherwise}. \end{cases}$$

Proof Suppose that $\zeta_k = i$ for any $i \in \Theta$, and drop the time index k when no confusion is caused. Then, for $k \geq d$ and the system (9.1) with $w_k \equiv 0$, we have

$$\begin{aligned} &\nabla V(x, i) \\ &= \sum_{i_1\in\Theta, i_2\in\Theta} \left[q_{i_1 i_2 i} (A_{i_1} x + B_{i_1} u')^T P_{i_2} (A_{i_1} x + B_{i_1} u') \right] - x^T P_i x \\ &= b_i u'^2 + 2\mathbf{a}_i x u' + x^T (F_i - P_i) x. \end{aligned} \tag{9.27}$$

For Case 1: $b_i = 0$. Based on the definition of b_i, it is direct to get $q_{i_1 i_2 i} B_{i_1}^T = 0$ for any $i_1, i_2 \in \Theta$ since $P_{i_2} > 0$, which further implies that $\mathbf{a}_i = 0$. The MSQ stabilization guarantees that $F_i - P_i + Q_i < 0$, and thus $K_i = 0$, i.e., $u' = 0$ can be adopted, which renders $\eta_f(i) = 0$. In this situation, we can set $\delta_m(i) = \infty$ without loss of generality.

For Case 2: $b_i \neq 0$. From (9.27), it holds that

$$\nabla V(x,i) + x^T Q_i x = \{-x^T M_i x + (u' - K_{mi}x)^2\} b_i,$$

and therefore the MSQ stabilization ensures that $M_i > 0$.

Then, $\nabla V(x,i) < -x^T Q_i x,\ \forall x \neq 0$, if and only if $u' = f(v,i) \in (u'_1(i), u'_2(i))$, where

$$u'_1(i) = K_{mi}x - \sqrt{x^T M_i x},\ \ u'_2(i) = K_{mi}x + \sqrt{x^T M_i x}.$$

By applying the orthogonal decomposition method, $M_i^{1/2}x$ can be decomposed into

$$M_i^{1/2}x = \alpha(i) M_i^{-1/2} K_{mi}^T + \beta(i), \tag{9.28}$$

where $\alpha(i)$ is a scalar, and the vector $\beta(i)$ is orthogonal to $M_i^{-1/2}K_{mi}^T$. Therefore, $u'_1(i),\ u'_2(i)$ can be rewritten with respect to the new coordinate system (9.28) as

$$u'_1(i) = \frac{\alpha(i)}{\delta_m(i)^2} - \sqrt{\frac{\alpha(i)^2}{\delta_m(i)^2} + \beta(i)^T\beta(i)},$$

$$u'_2(i) = \frac{\alpha(i)}{\delta_m(i)^2} + \sqrt{\frac{\alpha(i)^2}{\delta_m(i)^2} + \beta(i)^T\beta(i)}.$$

Moreover, if $\delta_m(i) > 1$, then we can again choose $u' = 0$ similarly to Case 1, since $u' = 0$ belongs to the interval $(u'_1(i), u'_2(i))$; if $\delta_m(i) \leq 1$, then it can be proved that the optimal quantization strategy with the smallest $\eta_f(i)$ for mode i is logarithmic, as shown in (9.16) and (9.17), with $\delta(i) = \delta_m(i)$ [2].

By combining the above two cases and taking note of the property (9.12), we can conclude that the logarithmic quantizer stated in this theorem can achieve the smallest C_f for a given set of

$$P_i > 0, Q_i > 0,\ i = 0, 1, \ldots, N.$$

The technique in the proof of Lemma 2.1 in [5] can still be used to prove that a linear state feedback law $v_k = K_{\zeta_k}x_k$ is sufficient to obtain the coarsest quantization for Case 2 with $\delta_m(i) \leq 1$, while, for Case 2 with $\delta_m(i) > 1$ and Case 1, the argument is trivial, since $K_i = 0$ is adopted. This completes the proof. □

The quantization error of a logarithmic quantizer is given by

$$e_k = u'_k - v_k = f(v_k, \zeta_k) - v_k = \Delta(v_k, \zeta_k)v_k, \tag{9.29}$$

where $\Delta(v_k, \zeta_k) \in [-\delta(\zeta_k), \delta(\zeta_k)]$. The closed-loop quantized feedback system with $v_k = K_{\zeta_k} x_k$ becomes the following uncertain MJLS:

$$x_{k+1} = A_{\theta_k} x_k + B_{\theta_k}(1 + \Delta(K_{\zeta_k} x_k, \zeta_k))K_{\zeta_k} x_k. \tag{9.30}$$

Before optimizing the overall coarseness with respect to all possible $P_i > 0$, $Q_i > 0$, $i \in \Theta$ such that (9.30) is MSQ stable in part (b) of the theorem that follows, we note that the uncertainty in (9.30) is a nonlinear function of $K_{\zeta_k} x_k$, which cannot be handled directly. The validity of the sector bound approach proved in part (a) of the next theorem shows that quantized stabilization is equivalent to the robust MSQ stabilization of an uncertain system with time-varying uncertainties.

Theorem 9.2 (a) *Given a logarithmic quantizer* (9.15)–(9.17) *with a set of fixed* $\delta(i) \in [0, 1]$, $i = 0, 1, \ldots, N$, *the system* (9.1) *under Assumption* 9.1 *is MSQ stabilizable via quantized linear state feedback if and only if the following uncertain system*

$$x_{k+1} = A_{\theta_k} x_k + B_{\theta_k}(1 + \Delta(\zeta_k))v_k \tag{9.31}$$

is robustly MSQ stabilizable for uncertainty $\Delta(\zeta_k) \in [-\delta(\zeta_k), \delta(\zeta_k)]$ *via a linear state feedback law* $v_k = K_{\zeta_k} x_k$.

(b) *Under Assumption* 9.1, *the optimal overall coarseness for the system* (9.1) *to be MSQ stabilizable via quantized linear state feedback can be obtained by the following optimization:*

$$\underline{\mathrm{C}}_f \triangleq \min_{S_i > 0, W_i > 0, Y_i, \tau(i) > 0, \forall i \in \Theta} \mathrm{C}_f$$

subject to the constraint

$$\begin{bmatrix} -S_i & S_i & Y_i^T & \Phi_{0i} & \Phi_{1i} & \cdots & \Phi_{Ni} \\ * & -W_i & 0 & 0 & 0 & \cdots & 0 \\ * & * & -\tau(i) & 0 & 0 & \cdots & 0 \\ * & * & * & \Xi_{0i} & 0 & \cdots & 0 \\ * & * & * & * & \Xi_{1i} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & * & * & \Xi_{Ni} \end{bmatrix} < 0, \tag{9.32}$$

where

$$\Phi_{ji} = \left[\sqrt{q_{j0i}}(S_i A_j^T + Y_i^T B_j^T) \;\; \sqrt{q_{j1i}}(S_i A_j^T + Y_i^T B_j^T) \;\; \cdots \;\; \sqrt{q_{jNi}}(S_i A_j^T + Y_i^T B_j^T)\right],$$

$$\Xi_{ji} = \begin{bmatrix} -S_0 + \tau(i)\delta(i)^2 q_{j0i} B_j B_j^T & \tau(i)\delta(i)^2 \sqrt{q_{j0i} q_{j1i}} B_j B_j^T & \cdots \\ * & -S_1 + \tau(i)\delta(i)^2 q_{j1i} B_j B_j^T & \cdots \\ \vdots & \vdots & \ddots \\ * & * & \cdots \end{bmatrix}$$
$$\begin{bmatrix} \tau(i)\delta(i)^2 \sqrt{q_{j0i} q_{jNi}} B_j B_j^T \\ \tau(i)\delta(i)^2 \sqrt{q_{j1i} q_{jNi}} B_j B_j^T \\ \vdots \\ -S_N + \tau(i)\delta(i)^2 q_{jNi} B_j B_j^T \end{bmatrix},$$

for all $i, j = 0, 1, \ldots, N$. *Moreover, a logarithmic quantizer* (9.15)–(9.17) *and a linear state feedback law* $v_k = K_{\zeta_k} x_k$ *are sufficient to achieve the* $\underline{C}_f$, *and a set of suitable state feedback gains is given by* $K_i = Y_i S_i^{-1},\ i = 0, 1, \ldots, N$.

Proof (a) Again, suppose that $\zeta_k = i,\ i \in \Theta$. For the system (9.30), we have

$$\begin{aligned} &\nabla V(x, i) \\ &= \sum_{i_1 \in \Theta, i_2 \in \Theta} \Big[q_{i_1 i_2 i} (A_{i_1} x + B_{i_1}(1 + \Delta(K_i x, i)) K_i x)^T \\ &\qquad \times P_{i_2} (A_{i_1} x + B_{i_1}(1 + \Delta(K_i x, i)) K_i x) \Big] \\ &\quad - x^T P_i x. \end{aligned} \tag{9.33}$$

Following a similar proof of Lemma 2.2 in [5], it can be shown that, for all $x \neq 0$, the inequality $\nabla V(x, i) < -x^T Q_i x$ is equivalent to

$$\begin{aligned} &x^T P_i x - x^T Q_i x > \\ &\sum_{i_1 \in \Theta, i_2 \in \Theta} \Big[q_{i_1 i_2 i} (A_{i_1} x + B_{i_1}(1 + \Delta(i)) K_i x)^T P_{i_2} (A_{i_1} x + B_{i_1}(1 + \Delta(i)) K_i x) \Big], \end{aligned} \tag{9.34}$$

where $\Delta(i)$ is defined as in (9.31) with $\zeta_k = i$. This kind of equivalence is true for any $i \in \Theta$, and thus, by Definition 9.1, (9.34) is the condition for the robust MSQ stabilization of (9.31).

(b) The constraint (9.32) is obtained by using Schur's complement documented in Lemma A.1 on the inequality (9.34) and taking

$$S_i = P_i^{-1},\ W_i = Q_i^{-1},\ Y_i = K_i S_i, \tag{9.35}$$

where $\tau(i) > 0$ is the scaling variable. From the proof in part (a), we see that the quantized stabilization for (9.30) and the robust stabilization for (9.31) can share the same set of $P_i,\ Q_i,\ i = 0, 1, \ldots, N$, as well as the same set of feedback gains. The result then follows directly from Theorem 9.1. □

As we can see from (9.14) and (9.18), the overall coarseness C_f of a logarithmic quantizer can also be defined in terms of the set of $\delta(i)$ or $\rho(i),\ i = 0, 1, \ldots, N$. For example, one possible choice is

$$\mathrm{C}_{f1} \triangleq -\min_{i\in\Theta}\{\delta(i)\}, \tag{9.36}$$

which captures the worst-case mode with the smallest sector bound (equivalently the largest mode quantization density) among all system modes. In this case, the optimization in part (b) of Theorem 9.2 becomes $\max_{S_i,W_i,Y_i,\tau(i)}\delta$ over (9.32) with $\delta(i) = \delta$ for every $i \in \Theta$. Moreover, suppose that θ_k is driven by an ergodic Markov chain which admits a limiting probability distribution $\{\bar{\pi}_i; \bar{\pi}_i > 0, i \in \Theta\}$. Then, another choice for the overall coarseness can be

$$\mathrm{C}_{f2} \triangleq -\sqrt{\sum_{i=0}^{N}\bar{\pi}_i\delta(i)^2}, \tag{9.37}$$

which characterizes the weighted average quantization performance. Since, for any fixed set of $\delta(i)$, (9.32) is convex in S_i, W_i, Y_i and $\tau(i)$, $\underline{\mathrm{C}}_f$ can be obtained by searching the space of $\delta(i),\ i = 0, 1, \ldots, N$. Note that such a method may be time-consuming especially when the number of system modes is large.

9.1.2 Special Schemes

The results presented in Sect. 9.1.1 can be applied to different scenarios as follows.

- Scheme I [Current mode observation (CMO)]: $\zeta_k = \theta_k$. In this situation,

$$q_{i_1i_2i_3} = \begin{cases} \pi_{i_3i_2}, & \text{if } i_1 = i_3, \\ 0, & \text{otherwise.} \end{cases}$$

- Scheme II [One-step-delayed mode observation (OSDMO)]: $\zeta_k = \theta_{k-1}$. In this situation,

$$q_{i_1i_2i_3} = \begin{cases} \pi_{i_3i_2}, & \text{if } i_1 = i_2, \\ 0, & \text{otherwise.} \end{cases}$$

- Scheme III [Mode-independent manner]: $\zeta_k \equiv \phi$ with ϕ representing a void signal. Assumption 9.1(b) is reduced to that

$$\Pr\{\theta_k = i_1|x_0^k\} = \Pr\{\theta_k = i_1\}, \quad \forall i_1 \in \Theta,$$

  is constant over time and known to the controller/quantizer, which is true if the underlying Markov chain is an i.i.d. process, i.e., $\pi_{ij} = \bar{\pi}_j$, for every $i, j \in \Theta$ [6],

or the Markov chain is ergodic and the initial distribution π is equal to its limiting distribution.

9.1.3 Mode Estimation

When the system mode is not directly observed at the controller/quantizer, one can adopt the mode-independent strategy (Scheme III) in Sect. 9.1.2. However, Scheme III may be conservative since no mode information is taken into consideration. Alternatively, we can try to estimate the mode process at the controller. First of all, a special case of mode estimation is given.

- Scheme IV [Mode estimation without process noise]: $w_k \equiv 0$ and for any $x \neq 0, i_1, i_2, i_3 \in \Theta, i_1 \neq i_2$,

$$A_{i_1}x + B_{i_1}f(g(x, i_3), i_3) \neq A_{i_2}x + B_{i_2}f(g(x, i_3), i_3). \tag{9.38}$$

  In this situation, the algorithm:

$$\zeta_k = \hat{\theta}_{k-1} = \arg\min_{i\in\Theta} \|x_k - A_i x_{k-1} - B_i f(g(x_{k-1}, \zeta_{k-1}), \zeta_{k-1})\|^2$$

  with arbitrary $\zeta_0 \in \Theta$ can ensure $\zeta_k = \theta_{k-1}$ for all $k \geq 1$. Thus, the result on OSDMO (Scheme II) can be applied directly.

With nonzero process noise w_k, one can still estimate the previous mode θ_{k-1} at time k based on x_0^k, ζ_0^{k-1} and the closed-loop system model (9.5). Assume that x_0 is white Gaussian and w_k is zero-mean white Gaussian. Denote $\Omega(x, \mu, \Sigma)$ as the vector-valued Gaussian probability density function with mean vector μ and covariance matrix Σ. Suppose that the initial distribution π and the transition probability matrix Π of the underlying Markov process as well as the set of covariance matrices $\Sigma_i,\ i = 0, 1, \ldots, N$, of the process noise w_k are exactly known to the controller/quantizer. The next algorithm gives an estimate of θ_{k-1} by maximizing the probability

$$L(\theta_{k-1}) \triangleq \Pr\{\theta_{k-1} | x_0^k, \zeta_0^{k-1}\} \tag{9.39}$$

with respect to $\theta_{k-1} \in \Theta$.

Algorithm 9.1.1 A recursive procedure to compute $\zeta_k = \hat{\theta}(k-1)$ at time $k \geq 1$ for the quantized system (9.5), such that L defined in (9.39) is maximized, is stated as follows.

(a) Choose ζ_0 as an arbitrary element in Θ and set $u_0' = 0$.
(b) For $k = 1$, $\zeta_1 = \arg\max_{i\in\Theta}[a(i, 1)]$ with

$$a(i, 1) = \pi_i \Omega(x_1, A_i x_0, \Sigma_i).$$

(c) For $k \geq 2$, $\zeta_k = \arg\max_{i\in\Theta}[a(i,k)]$, where $a(i,k)$ can be computed iteratively as

$$a(i,k) = \sum_{j\in\Theta} a(j,k-1)\pi_{ji}\Omega(x_k, A_i x_{k-1} + B_i f(g(x_{k-1},\zeta_{k-1}),\zeta_{k-1}), \Sigma_i).$$

The above algorithm is modified from the well-known Viterbi algorithm [7, 8]. The optimality criterion of the standard Viterbi algorithm [7], different from (9.39), is to find the single best mode sequence. Moreover, the maximum likelihood estimation can be used to iteratively update the parameters such as π, Π, Σ_i, if some or all of them are unknown to the controller/quantizer. For more complicated cases, e.g., partial state observation with corrupted noise, algorithms for mode estimation may be constructed based on a more sophisticated hidden Markov model; see, e.g., [9, 10].

Remark 9.4 Note that, for direct mode observation $\zeta_k = \theta_{k-d}$ with $d \geq 2$, and general cases of Algorithm 9.1.1, the probability on the left-hand side of equation (9.20) becomes a function of state x and thus is dynamic, which renders an optimal memoryless quantization strategy impossible. In this situation, dynamic or state-dependent quantization strategy would be required for an optimal design.

The next numerical example demonstrates the usefulness of Algorithm 9.1.1.

Example 9.1 Consider an MJLS (9.1) with $A_0 = 1.2,\ A_1 = -1.2,\ B_0 = B_1 = 1$, and transition probability matrix

$$\Pi_1 = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}.$$

First, suppose that direct mode observation is available at the controller/quantizer. Then, for CMO (Scheme I), the smallest allowable C_{f1} defined in (9.36) is -0.8333 with $K_0 = -1.2,\ K_1 = 1.2$; for OSDMO (Scheme II), the smallest achievable C_{f1} is -0.7229 with $K_0 = 0.9600,\ K_1 = -0.9600$.

Second, if the system mode is not observed at the controller/quantizer, then we can easily verify that the mode-independent strategy (Scheme III) cannot stabilize the system. Furthermore, assume that the covariance of w_k is given by $W_0 = W_1 = 1$ and the initial state x_0 is Gaussian distributed with mean 20 and variance 10. Then, the first 30 mode estimates for one sample of simulation using Algorithm 9.1.1 are shown in Fig. 9.3. The parameters of the controller and quantizer are chosen as in OSDMO: $K_0 = 0.9600,\ K_1 = -0.9600,\ \delta(0) = \delta(1) = 0.7229$. Figure 9.4 further gives the empirical norm of state by averaging 10,000 Monte Carlo simulations. As we can see from Figs. 9.3 and 9.4, there exist some mode estimation errors, but the error rate is low, and the empirical norm of state by applying Algorithm 9.1.1 is convergent.

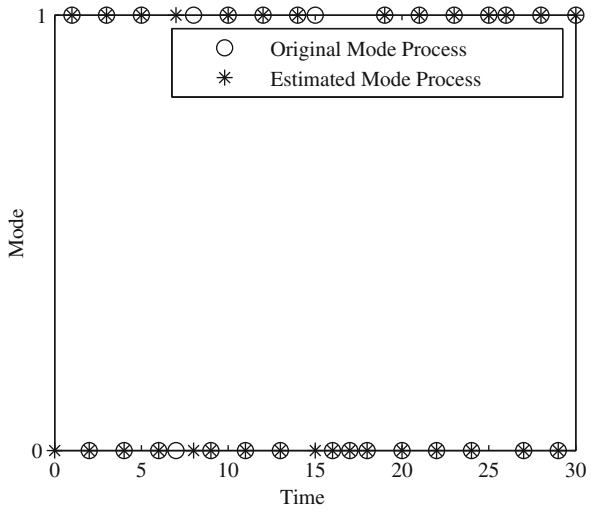

Fig. 9.3 Original mode process $\{\theta_{k-1}\}$ and estimated mode process $\{\zeta_k\}$ for one sample of simulation using Algorithm 9.1.1

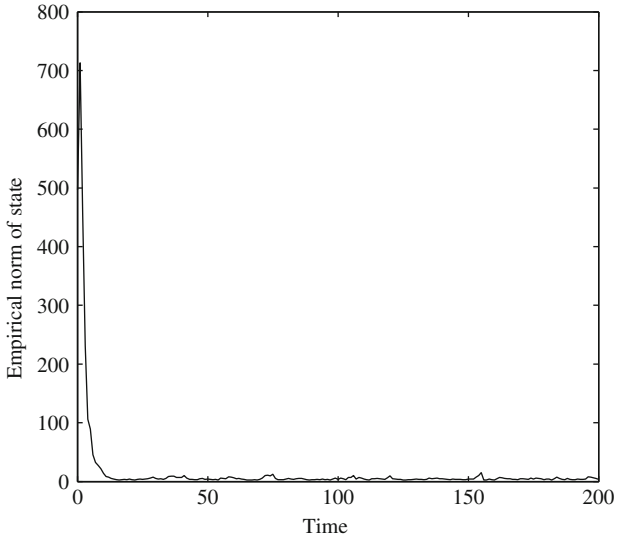

Fig. 9.4 The empirical norm of state $|x_k|^2$ by averaging 10,000 Monte Carlo simulations using Algorithm 9.1.1

9.2 Stabilization Over Lossy Channels

We refer to TCP-like channel if there exist acknowledgments that testify successful transmissions or not at the transmitter side, and to UDP-like channel otherwise. The packet acknowledgment is assumed to be available with one step delay. See [11, 12] for more details on TCP-like and UDP-like protocols.

Consider a quantized feedback NCS in Fig. 9.5, where the LTI plant is described in a discrete-time form as

$$x_{k+1} = Ax_k + Bu_k, \tag{9.40}$$

which may be obtained through discretization of a continuous-time system. The pair (A, B) is assumed to be stabilizable. We note that the TCP-like protocol falls into the OSDMO scheme, whereas the UDP-like protocol falls into the mode-independent pattern. Next, two stochastic models for the lossy channel will be discussed.

9.2.1 Binary Dropouts Model

Suppose that a zero-control strategy is adopted in dealing with the binary packet losses over the network. The network is modeled by

$$u_k = \gamma_k u'_k, \tag{9.41}$$

where $\gamma_k \in \Theta = \{0, 1\}$ represents the loss (with $\gamma_k = 0$) or arrival (with $\gamma_k = 1$) of the packet at time k. The system as a combination of the network and the plant can be modeled as an MJLS (9.1) with

$$A_0 = A_1 = A, \quad B_0 = 0, \quad B_1 = B. \tag{9.42}$$

For TCP-like channel with $\zeta_k = \gamma_{k-1}$, we assume that γ_k is driven by a Markov chain with transition probability matrix

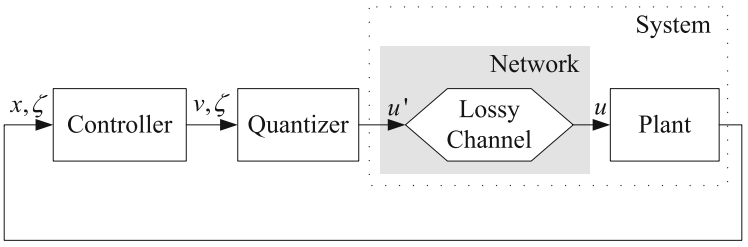

Fig. 9.5 Quantized control over an input lossy channel

$$\Pi = \begin{bmatrix} 1-q & q \\ p & 1-p \end{bmatrix}. \tag{9.43}$$

In this situation, the OSDMO result (Scheme II) is applicable. Note that the CMO result (Scheme I) is of theoretical importance in the quantization of MJLSs but may not be practical in the NCS depicted in Fig. 9.5, since it is unrealistic for the quantizer to know whether the current packet will be lost or not before the packet is sent over the network.

For the UDP-like protocol, we can easily verify that the inequality (9.38) is true, and thus Scheme IV can be used directly when $w_k \equiv 0$. If γ_k is assumed to be an i.i.d. random variable with

$$\Pr(\gamma_k = 0) = \alpha, \quad \Pr(\gamma_k = 1) = 1 - \alpha, \tag{9.44}$$

i.e., the NCS adopts an unreliable network with packet-loss rate α, then Scheme III is applicable, and the inequality (9.32) is reduced to the following modified Riccati inequality:

$$A^T PA - P + Q - (1-\alpha)(1-\delta^2)A^T PB(B^T PB)^{-1}B^T PA < 0. \tag{9.45}$$

Based on Lemma 5.4 in [12], the condition

$$(1-\alpha)(1-\delta^2) > 1 - \frac{1}{\mathscr{M}(A)^2}$$

can ensure the existence of $P > 0$ to (9.45), where $\mathscr{M}(A)$ denotes the Mahler measure of A. It is easy to check that the above result is consistent with Theorem 2.1 of [13], which can be seen as a special case of Theorem 9.2 (b).

Example 9.2 Consider the NCS in Fig. 9.5, where the plant (9.40) is borrowed from [13] with

$$A = \begin{bmatrix} 0 & 1 \\ 1.8 & -0.3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{9.46}$$

Suppose that the transition probability matrix (9.43) is given by

$$\Pi_2 = \begin{bmatrix} 0.1 & 0.9 \\ 0.3 & 0.7 \end{bmatrix}. \tag{9.47}$$

For TCP-like channel, the smallest allowable C_{f1} defined in (9.36) is -0.2474 with

$$K_0 = K_1 = [-1.8000 \;\; 0.5991]$$

according to Scheme II. The limiting transition probability of (9.47) is given by

$$\Pi_3 = \lim_{k\to\infty} \Pi_2^k = \begin{bmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{bmatrix}. \tag{9.48}$$

Therefore, for UDP-like channel with initial distribution $\pi_1 = [0.25 \quad 0.75]$, the smallest achievable C_{f1} is -0.2796 with $K_\phi = [-1.8000 \quad 0.6750]$ based on Scheme III.

9.2.2 Bounded Dropouts Model

Suppose that the lossy channel undergoes bounded Markovian dropouts, and a zero-order hold is used in dealing with the bounded packet losses. The model of the network is given by

$$u_k = \gamma_k u'_k + (1 - \gamma_k)u'_{k-1}, \quad \gamma_k \in \{0, 1\}.$$

Then, from Theorem 9 of [14], the system as a combination of the network and the plant in Fig. 9.5 can be modeled as an MJLS (9.1) with

$$A_i = A^{i+1}, \quad B_i = \sum_{r=0}^{i} A^r B, \quad i \in \Theta = \{0, 1, \ldots, N\}, \tag{9.49}$$

where N is the number of maximum successive packet losses.

Example 9.3 Let the plant in Fig. 9.5 be given by (9.40) with [14]

$$A = \begin{bmatrix} 0.6065 & 0 & -0.2258 \\ 0.3445 & 0.7788 & -0.0536 \\ 0 & 0 & 1.2840 \end{bmatrix}, B = \begin{bmatrix} -0.0582 \\ -0.0093 \\ 0.5681 \end{bmatrix}. \tag{9.50}$$

The number of maximum consecutive packet dropouts of the underlying network is 4 and the transition probability matrix is given by:

$$\Pi_4 = \begin{bmatrix} 0.5 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.5 & 0.3 & 0 & 0 \\ 0 & 0.2 & 0.5 & 0.3 & 0 \\ 0 & 0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.1 & 0.2 & 0.5 \end{bmatrix}. \tag{9.51}$$

For TCP-like channel, the smallest allowable C_{f1} is -0.3470 with

$$K_0 = [0 \quad 0 \quad -0.7783], \quad K_1 = [0 \quad 0 \quad -1.0614], \quad K_2 = [0 \quad 0 \quad -0.8321],$$
$$K_3 = [0 \quad 0 \quad -0.7452], \quad K_4 = [0 \quad 0 \quad -0.7201].$$

For UDP-like channel with initial distribution

$$\pi_2 = [0.1101 \;\; 0.1878 \;\; 0.2718 \;\; 0.2552 \;\; 0.1751],$$

the smallest achievable C_{f1} is -0.3700 with $K_\phi = [0 \;\; 0 \;\; -0.7936]$.

9.2.3 Extension to Output Feedback

If the plant state is not fully observed at the controller in Fig. 9.5, then output feedback should be adopted. Assume that the plant is described as

$$\begin{aligned} x_{k+1} &= Ax_k + Bu'_k, \\ y_k &= Cx_k, \end{aligned} \tag{9.52}$$

and the measured output $y_k \in \mathbb{R}^\ell$ is available at the controller. The proposition that follows can be proved similarly to Theorem 7.2.

Proposition 9.2.1 *Consider the plant (9.52) to be controlled over a lossy channel via quantized output feedback. Assume that the pair (A, C) is observable. Under the TCP-like protocol, the optimal overall coarseness for MSQ stabilization by state feedback can be achieved by output feedback. Moreover, one possible deadbeat-observer-based controller can be constructed as*

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu'_k + \hat{L}(y_k - C\hat{x}_k), \\ v_k &= K_{\zeta_k} x_k, \\ u'_k &= f(v_k, \zeta_k), \end{aligned} \tag{9.53}$$

where K_{ζ_k} and $f(\cdot, \cdot)$ are chosen according to the OSDMO scheme, and $\hat{L}$ is any deadbeat observer gain.

9.3 Summary

Motivated by quantized feedback control over lossy channels, the quantized stabilization problem for MJLSs has been investigated in this chapter. It has been shown that, for a single-input linear system with Markovian jump parameters, a mode-dependent logarithmic quantizer is still optimal in the MSQ stability sense, and the sector bound approach again provides a non-conservative way for studying the corresponding quantized state feedback stabilization problem. In addition, a recursive algorithm has been presented to estimate the unknown mode process at the controller side. The above results have been applied in the quantized stabilization of NCSs over lossy channels under either TCP-like or UDP-like protocol.

The results in this chapter are based mainly on [6, 15]. As a special case of this chapter, the stabilization of a single-input system over a channel subject to both quantization and binary i.i.d. packet losses is addressed in [13, 16].

References

1. E. Boukas, Z. Liu, Robust H_∞ control of discrete-time Markovian jump linear systems with mode-dependent time-delays. IEEE Trans. Autom. Control **46**(12), 1918–1924 (2001)
2. N. Elia, S. Mitter, Stabilization of linear systems with limited information. IEEE Trans. Autom. Control **46**(9), 1384–1400 (2001)
3. P. Khargonekar, M. Rotea, Multiple objective optimal control of linear systems: the quadratic norm case. IEEE Trans. Autom. Control **36**(1), 14–24 (2002)
4. O. Costa, M. Fragoso, R. Marques, *Discrete-Time Markov Jump Linear Systems* (Springer, New York, 2005)
5. M. Fu, L. Xie, The sector bound approach to quantized feedback control. IEEE Trans. Autom. Control **50**(11), 1698–1711 (2005)
6. N. Xiao, L. Xie, M. Fu, Quantized stabilization of markov jump linear systems via state feedback, in *Proceedings of American Control Conference*, pp. 4020–4025 (2009)
7. A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory **13**(2), 260–269 (1967)
8. L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
9. R. Elliott, L. Aggoun, J. Moore, *Hidden Markov Models Estimation and Control* (Springer, Berlin, 1995)
10. R. Elliott, F. Dufour, W. Malcolm, State and mode estimation for discrete-time jump Markov systems. SIAM J. Control Optim. **44**, 1081–1104 (2005)
11. O. Imer, S. Yüksel, T. Başar, Optimal control of LTI systems over unreliable communication links. Automatica **42**(9), 1429–1439 (2006)
12. L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks. Proc. IEEE **95**(1), 163–187 (2007)
13. H. Hoshina, K. Tsumura, H. Ishii, The coarsest logarithmic quantizers for stabilization of linear systems with packet losses, in *Proceedings of the 46th IEEE Conference on Decision and Control* (2007)
14. J. Xiong, J. Lam, Stabilization of linear systems over networks with bounded packet loss. Automatica **43**(1), 80–87 (2007)
15. N. Xiao, L. Xie, M. Fu, Stabilization of Markov jump linear systems using quantized state feedback. Automatica **46**(10), 1696–1702 (2010)
16. K. Tsumura, H. Ishii, H. Hoshina, Tradeoffs between quantization and packet loss in networked control of linear systems. Automatica **45**(12), 2963–2970 (2009)

Chapter 10
Kalman Filtering with Quantized Innovations

This chapter presents a multi-level quantized innovations Kalman filter (MLQ-KF) of linear stochastic systems. For a given multi-level quantization and under the Gaussian assumption on the predicted density, a quantized innovations filter that achieves the MMSE is derived. The filter is given in terms of quantization thresholds and a simple modified Riccati difference equation (MRDE). By optimizing the filtering error covariance w.r.t. quantization thresholds, the associated optimal thresholds and the corresponding filter are obtained. Furthermore, the convergence of the filter to the standard Kalman filter is established. We also discuss the design of a robust minimax quantized filter when the innovation covariance is not exactly known. Simulation results illustrate the effectiveness and advantages of the proposed quantized filter.

The chapter is organized as follows. The problem is formulated in Sect. 10.1. The multi-level quantized innovations Kalman filter is derived in Sect. 10.2, where we first derive the filter based on the general quantization scheme and then proceed to seek the optimal quantization. In Sect. 10.3, a max-min quantization optimization is described to address the robust quantization problem. Simulation and experiments are carried out in Sect. 10.4 to illustrate the performance of the quantized filter. Some concluding remarks are drawn in Sect. 10.5.

10.1 Problem Formulation

Consider the following discrete-time linear stochastic system:

$$x_{k+1} = Ax_k + w_k, \tag{10.1}$$

$$y_k = Cx_k + v_k, \tag{10.2}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}$ are vector state and scalar measurement. $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}$ are white Gaussian noises with zero means and covariance matrices $Q > 0$ and $R > 0$, respectively. The initial state x_0 is a random Gaussian vector of

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_10

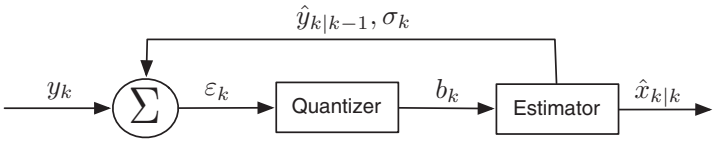

Fig. 10.1 Network configuration

mean $\hat{x}_0$ and the covariance matrix $P_0 > 0$. Moreover, w_k, v_k and x_0 are mutually independent.

Let $\hat{y}_{k|k-1}$ be the one-step ahead prediction of the output at the time instant k and

$$\sigma_k^2 = CP_{k|k-1}C^T + R$$

be the prediction error (innovation) covariance, where $P_{k|k-1}$ is the prediction error covariance matrix of the state at time instant k. We assume that the sensor can access $\hat{y}_{k|k-1}$ and σ_k which are either broadcasted by the estimation center (EC) (see Fig. 10.1) or computed by the sensor. Due to the communication constraint, the sensor information has to be quantized before being transmitted to the EC. Once the sensor receives an observation y_k, it computes the innovation $\eta_k := y_k - \hat{y}_{k|k-1}$ and normalizes it by $\varepsilon_k = \eta_k/\sigma_k$. Due to the symmetric property of the standard Gaussian distribution, we consider the associated symmetric quantizer (denoted by 'Q' in Fig. 10.1) with *quantization thresholds* $\{\pm z_i\}_{i=1}^N$ and

$$0 = z_0 < z_1 < \cdots < z_N < \infty = z_{N+1},$$

namely, the output b_k of the quantizer $Q(\cdot)$ is given by

$$b_k \triangleq Q(\varepsilon_k) = \begin{cases} z_N, & z_N < \varepsilon_k \\ \vdots & \vdots \\ z_0, & z_0 < \varepsilon_k \le z_1 \\ -q(-\varepsilon_k), & \varepsilon_k \le 0. \end{cases} \tag{10.3}$$

Note that when $b_k = z_0$, it will not be transmitted to the EC. That is, for a 1-bit budget, the quantizer has 3 quantization levels unlike that in [1] where innovations are quantized to 1 or -1, depending on whether they are positive or negative. Intuitively, our approach should perform better since one more quantization level is added. This will be confirmed in theory and simulation later. Assuming that there is no transmission error, our goal is to find and analyze the MMSE state estimate based on the quantized innovations.

Denote the quantized sequence by

$$\mathbf{b}_{0:k} \triangleq \{b_0, b_1, \ldots, b_k\}$$

and $\hat{x}_{k|k}$ represents the MMSE estimate of x_k given $\mathbf{b}_{0:k}$, i.e.,

$$\hat{x}_{k|k} \triangleq \mathbb{E}[x_k|\mathbf{b}_{0:k}] = \int_{\mathbb{R}^n} x_k\, p[x_k|\mathbf{b}_{0:k}]dx_k. \tag{10.4}$$

The following equalities can be easily obtained [1]:

$$\hat{x}_{k|k-1} \triangleq \mathbb{E}[x_k|\mathbf{b}_{0:k-1}] = A\hat{x}_{k-1|k-1}, \tag{10.5}$$
$$\hat{y}_{k|k-1} \triangleq \mathbb{E}[y_k|\mathbf{b}_{0:k-1}] = C\hat{x}_{k|k-1}. \tag{10.6}$$

Their filtering error covariance matrices are respectively defined by

$$P_{k|k} \triangleq \mathbb{E}[(\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)^T], \tag{10.7}$$
$$P_{k|k-1} \triangleq \mathbb{E}[(\hat{x}_{k|k-1} - x_k)(\hat{x}_{k|k-1} - x_k)^T] = AP_{k-1|k-1}A^T + Q. \tag{10.8}$$

10.2 Quantized Innovations Kalman Filter

In this section, given a multi-level quantization for normalized innovations, we will first derive the MMSE estimate of state under the assumption that the predicted density is Gaussian. The filtering error covariance matrix is given in terms of quantization thresholds. Minimizing the filtering error covariance w.r.t. the quantization thresholds leads to the corresponding optimal quantizer and filter. For the simplicity of notation, denote $z_1^N = (z_1, z_2, \ldots, z_N)$.

10.2.1 Multi-level Quantized Filtering

Theorem 10.1 *Consider the systems* (10.1) *and* (10.2)*, given a multi-level quantization of normalized innovations in* (10.3)*, if*

$$p[x_k|\mathbf{b}_{0:k-1}] = \mathcal{N}[x_k; \hat{x}_{k|k-1}, P_{k|k-1}],$$

the MMSE estimate of the state can be computed by

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \frac{f(z_1^N, b_k)P_{k|k-1}C^T}{\sqrt{CP_{k|k-1}C^T + R}}, \tag{10.9}$$

$$P_{k|k} = P_{k|k-1} - F(z_1^N)\frac{P_{k|k-1}C^T CP_{k|k-1}}{CP_{k|k-1}C^T + R}, \tag{10.10}$$

with

$$f(z_1^N, b_k) = \sum_{j=-N}^{N} 1_{\{z_j\}}(b_k) \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})}, \tag{10.11}$$

$$F(z_1^N) = 2 \sum_{j=0}^{N} \frac{[\phi(z_j) - \phi(z_{j+1})]^2}{T(z_j) - T(z_{j+1})}, \tag{10.12}$$

where $z_{-j} = -z_j$. The functions $\phi(\cdot)$ and $T(\cdot)$ respectively denote the density and tail probability function of a standard Gaussian random variable, i.e.,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

and

$$T(x) = \int_x^{\infty} \phi(s) ds.$$

Proof By iterated conditioning, we obtain that

$$\hat{x}_{k|k} = \mathbb{E}[x_k | \mathbf{b}_{1:k}] = \mathbb{E}[\mathbb{E}[x_k | \mathbf{b}_{1:k-1}, y_k] | \mathbf{b}_{1:k}]. \tag{10.13}$$

And the posterior density $p[x_k | \mathbf{b}_{1:k-1}, y_k]$ is obtained as

$$p[x_k | \mathbf{b}_{1:k-1}, y_k] = \frac{p(y_k | x_k) p(x_k | \mathbf{b}_{1:k-1})}{\int_{\mathbb{R}^n} p(y_k | x_k) p(x_k | \mathbf{b}_{1:k-1}) dx_k}. \tag{10.14}$$

By the Gaussian assumption and following the technique of the Kalman filter, the inner conditional expectation in (10.13) can be easily obtained as follows:

$$\hat{x}_{k|k}^* \triangleq \mathbb{E}[x_k | \mathbf{b}_{1:k-1}, y_k] = \hat{x}_{k|k-1} + K_k (y_k - C \hat{x}_{k|k-1}), \tag{10.15}$$

where

$$K_k = (C^T P_{k|k-1} C + R)^{-1} P_{k|k-1} C^T.$$

By (10.3) and Gaussian assumption, it follows that

$$\mathbb{E}[\varepsilon_k | b_k = z_j, \mathbf{b}_{1:k-1}] = \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})}. \tag{10.16}$$

Note that $\hat{x}_{k|k-1}$ is measurable with respect to the σ-algebra generated by $\mathbf{b}_{1:k}$. Inserting (10.15) and (10.16) into (10.13) yields (10.9).

Applying the same arguments, we have that

$$P_{k|k} = \mathbb{E}[\mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T | \mathbf{b}_{1:k-1}, y_k]], \tag{10.17}$$

where the inner expectation is computed as follows:

$$\begin{aligned}
&\mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T | \mathbf{b}_{1:k-1}, y_k] \\
&\quad = \mathbb{E}[(x_k - \hat{x}^*_{k|k})(x_k - \hat{x}^*_{k|k})^T | \mathbf{b}_{1:k-1}, y_k] \\
&\qquad + \mathbb{E}[(\hat{x}_{k|k} - \hat{x}^*_{k|k})(\hat{x}_{k|k} - \hat{x}^*_{k|k})^T | \mathbf{b}_{1:k-1}, y_k] \\
&\quad = P^*_{k|k} + \mathbb{E}[(\hat{x}_{k|k} - \hat{x}^*_{k|k})(\hat{x}_{k|k} - \hat{x}^*_{k|k})^T | \mathbf{b}_{1:k-1}, y_k],
\end{aligned} \tag{10.18}$$

where

$$P^*_{k|k} = P_{k|k-1} - P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}CP_{k|k-1}.$$

Under the Gaussian assumption, we have that

$$\begin{aligned}
&\mathbb{E}[(\hat{x}_{k|k} - \hat{x}^*_{k|k})(\hat{x}_{k|k} - \hat{x}^*_{k|k})^T] \\
&\quad = \mathbb{E}[\mathbb{E}[(\hat{x}_{k|k} - \hat{x}^*_{k|k})(\hat{x}_{k|k} - \hat{x}^*_{k|k})^T | \mathbf{b}_{1:k-1}]] \\
&\quad = (CP_{k|k-1}C^T + R)K_k\mathbb{E}[\mathbb{E}[(\varepsilon_k - f(z_1^N, b_k))^2 | \mathbf{b}_{1:k-1}]]K_k^T \\
&\quad = \frac{P_{k|k-1}C^TCP_{k|k-1}}{CP_{k|k-1}C^T + R}\left(1 - \mathbb{E}[\mathbb{E}[f^2(z_1^N, b_k) | \mathbf{b}_{1:k-1}]]\right) \\
&\quad = \frac{P_{k|k-1}C^TCP_{k|k-1}}{CP_{k|k-1}C^T + R}(1 - F(z_1^N)).
\end{aligned} \tag{10.19}$$

Together with (10.17) and (10.18), one can easily derive (10.10).

Remark 10.1 Strictly speaking, the system state conditioned on the quantized innovations can not remain Gaussian due to the nonlinear operator of quantization. To enable the development of a simple and practically useful recursive filter, we have made the Gaussian assumption as in [1]. A similar hypothesis can also be found in [2]. It will be demonstrated later that the performance of the quantized filter for a moderate number of bits (say 2 or more bits) is close to the standard Kalman filter, suggesting that the approximation is reasonable.

We observe that the approximate filtering error covariance matrix $P_{k|k}$ computed from (10.12) depends on $F(z_1^N)$ which is a function of quantization thresholds. Without quantization, $F(z_1^N) = 1$, which gives rise to the standard Kalman filter. We call $F(z_1^N)$ *performance recovery factor*.

10.2.2 Optimal Quantization Thresholds

The performance of the MLQ-KF can be approximately measured by the quantity of $tr[P_{k|k}]$ or $tr[P_{k|k-1}]$. According to Theorem 10.1, the optimal thresholds $(z^*)_1^N$ of the quantizer in (10.3) can be obtained by maximizing $F(z_1^N)$, i.e.,

$$(z^*)_1^N = \arg\max_{z_1^N} F(z_1^N).$$

The numerical solutions to the above optimization is obtained by evoking the Matlab command,

$$[x, y] = \text{fmincon}(fun, x_0, A, b, Aeq, beq, lb, ub). \tag{10.20}$$

Thus, the optimal quantizer can be numerically obtained, see Table 10.1. As mentioned above, $b_k = 0$ will not be transmitted to the EC as it will not improve the predicted estimate. The comparison of performance recovery factor for the cases of with and without dead zone is showed in Fig. 10.2. Apparently, for single bit quantization, the quantized filter with dead zone gives a much improved performance recovery factor (0.8098) than the SOI-KF [1] ($2/\pi \approx 0.6366$). Figure 10.2 also indicates that the higher the bit rate, the higher the performance recovery factor. The

Table 10.1 Solutions to (10.20) for optimal quantization thresholds

N = 1	N = 2	N = 3	N = 4
$z_1^* = 0.612$	$z_1^* = 0.382$ $z_2^* = 1.244$	$z_1^* = 0.280$ $z_2^* = 0.874$ $z_3^* = 1.611$	$z_1^* = 0.221$ $z_2^* = 0.681$ $z_3^* = 1.198$ $z_4^* = 1.866$
$F(z_1^*) = 0.810$	$F((z^*)_1^2) = 0.920$	$F((z^*)_1^3) = 0.956$	$F((z^*)_1^4) = 0.972$

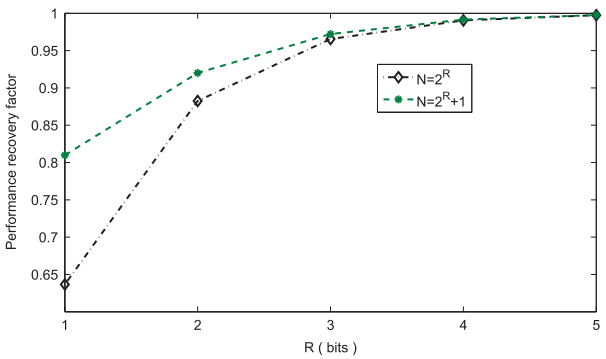

Fig. 10.2 $F((z^*)_1^N)$ versus R bits

relationship given in Fig. 10.2 will be useful in estimating the required bits number for desirable filtering performance and the stability of filter. We now investigate the stability of the MLQ-KF. Let $P_{k+1} = P_{k+1|k}$, the MRDE can be written as follows:

$$P_{k+1} = AP_kA^T + Q - \beta^* AP_kC^T(CP_kC^T + R)^{-1}CP_kA^T \tag{10.21}$$

with $\beta^* = F((z^*)_1^N)$. The corresponding algebraic Riccati equation (ARE) is

$$P = APA^T + Q - \beta^* APC^T(CPC^T + R)^{-1}CPA^T. \tag{10.22}$$

Corollary 10.1 *Consider MRDE* (10.21) *and assume that* $(A, Q^{1/2})$ *and* (C, A) *are respectively controllable and detectable. Then for an unstable* A*, the ARE* (10.22) *has a positive definite solution* P *and the MRDE* (10.21) *admits a unique positive definite solution* P_k *satisfying* $P_k \to P$ *for any* $P_0 \geq 0$ *if and only if*

$$\beta^* > 1 - \prod_i |\lambda_i^u(A)|^{-2},$$

where $\lambda_i^u(A)$ *are unstable eigenvalues of* A.

Corollary 10.1 reveals that for systems with faster growth rate, a higher bits number is required for the quantizer to ensure the convergence of MRDE.

10.2.3 Convergence Analysis

The following result establishes the convergence of the MLQ-KF to the standard Kalman filter when $N \to \infty$.

Theorem 10.2 *Let* $\triangle = \sup_{j\in\mathbb{N}} \triangle_j$*, where* $\triangle_j = |z_j - z_{j+1}|$ *and assume that the quantization thresholds in* (10.3) *satisfy*

(1) $\triangle_j \leq \triangle \to 0$;

(2) $S(N) = \sum_{k=1}^{N-1} \triangle_j \to \infty$ *as* $N \to \infty$.

Then, it follows that

$$\frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})} \to z_{j+1}, \tag{10.23}$$

and

$$F(z_1^N) \to 1 \text{ as } N \to \infty. \tag{10.24}$$

Proof Using some basic results in mathematical analysis, we obtain that

$$\lim_{z_j \to z_{j+1}} \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})} = \lim_{\triangle_j \to 0} \frac{\phi(z_{j+1} - \triangle_j) - \phi(z_{j+1})}{\int_{z_{j+1}-\triangle_j}^{z_{j+1}} \phi(t)dt}$$

$$= \lim_{\triangle_j \to 0} \frac{\phi^{'}(z_{j+1} - \triangle_j)}{\phi(z_{j+1} - \triangle_j)} = z_{j+1}.$$

$$S \triangleq \sum_{k=1}^{\infty} \frac{[\phi(z_j) - \phi(z_{j+1})]^2}{T(z_j) - T(z_{j+1})} = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{[e^{(-\frac{z_j^2}{2})} - e^{(-\frac{z_{j+1}^2}{2})}]^2}{\int_{z_j}^{z_{j+1}} e^{(-\frac{t^2}{2})} dt}$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{[e^{(-\frac{z_j^2}{2})} - e^{[-\frac{(z_j+\triangle_j)^2}{2}]}]^2}{e^{-\frac{(z_j+\theta_j\triangle_j)^2}{2}} \triangle_j}$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} e^{-\frac{z_j^2}{2}} \frac{[1 - e^{-(\frac{\triangle_j^2}{2} - z_j\triangle_j)}]^2}{\triangle_j e^{-\frac{(\theta_j\triangle_j)^2}{2} - \theta_j z_j \triangle_j}},$$

where $0 \leq \theta_j \leq 1$. Using the Taylor expansion, we have that $S \triangleq S_0 + S_1$, where

$$S_1 = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{\infty} z_j^2 e^{-\frac{z_j^2}{2}} \triangle_j \to \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^2 e^{-t^2/2} dt = \frac{1}{2}$$

as $\triangle \to 0$, and there exists $|c_{i,j,v,u}| < \infty$ such that

$$S_0 = \frac{1}{\sqrt{2\pi}} \sum_{i,j,u,v} \sum_{k=1}^{\infty} [c_{i,j,v,u} e^{(-\frac{z_k^2}{2})} \triangle_k^{2+i} z_k^j \theta_j^u o(\triangle_k^v)]$$

$$< \triangle \sum_{i,j,u,v} c_{i,j,v,u} \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} z_k^j e^{(-\frac{z_k^2}{2})} \triangle_k \triangleq C\triangle \to 0$$

as $\triangle \to 0$, where C is a finite constant since for $\triangle \to 0$,

$$\frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} z_k^j e^{-\frac{z_k^2}{2}} \triangle z_k \to \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^j e^{-t^2/2} dt < \infty$$

and the nonnegative integers i, j, u, v can only take a finite number of elements.

Remark 10.2 As $z_j \to z_{j+1}$, then $\varepsilon_k \to z_{j+1}$ and $f(z_1^N, b_k) \to \varepsilon_k$ by (10.23). In light of (10.9), (10.10) and (10.24), the MLQ-KF converges to the standard Kalman filter since by Theorem 10.2, we obtain that

$$\hat{x}_{k|k} \to \hat{x}_{k|k-1} + \frac{P_{k|k-1}C^T\eta_k}{C\,P_{k|k-1}C^T + R} \quad \text{and} \quad P_{k|k} \to P_{k|k-1} - \frac{P_{k|k-1}C^T C\,P_{k|k-1}}{C\,P_{k|k-1}C^T + R}.$$

10.3 Robust Quantization

This section will examine the design of a robust quantizer when the prediction error covariance is not known exactly. In practice, the error covariance may not be accurate due to quantization error or we may not be able to transmit the covariance at every time instant due to the limited communication capacity.

Let

$$\delta_k = \sigma_k/\sigma_{k,e} \in [\underline{\delta}, \bar{\delta}],$$

where σ_k is the actual prediction error covariance and $\sigma_{k,e}$ is an estimated one. $\underline{\delta}$, and $\bar{\delta}$ are known lower and upper bounds of δ_k.

Denote $\hat{\varepsilon}_k = \eta_k/\sigma_{k,e}$, the quantization scheme (10.3) is modified as:

$$b_k \triangleq Q(\hat{\varepsilon}_k) = \begin{cases} z_N, & \delta z_N < \hat{\varepsilon}_k \\ \vdots & \vdots \\ z_0, & 0 < \hat{\varepsilon}_k \le \delta z_1 \\ -q(\hat{\varepsilon}_k), & \hat{\varepsilon}_k \le 0. \end{cases} \tag{10.25}$$

Note that the performance recovery factor becomes $F(\delta z_1^N)$. Since δ is not exactly known, we design a robust min-max quantizer by

$$\bar{z}_*^N = \arg\max_{z_1^N} \min_{\delta \in [\underline{\delta},\, \bar{\delta}]} F(\delta z_1^N).$$

Examining $F(z_1^N)$ results in that given z_1^N, the minimum is attained at the extreme point $\underline{\delta}$ or $\bar{\delta}$. Hence, it follows that

$$(\bar{z}^*)_1^N = \arg\max_{z_1^N} \min\{F(\underline{\delta} z_1^N),\, F(\bar{\delta} z_1^N)\}.$$

It is easy to check that

$$\frac{1}{\max\limits_{z_1^N} \min\{F(\underline{\delta} z_1^N),\, F(\bar{\delta} z_1^N)\}} = \min_{z_1^N} \max\left\{\frac{1}{F(\underline{\delta} z_1^N)}, \frac{1}{F(\bar{\delta} z_1^N)}\right\}.$$

We recall the command $x = \text{fminimax}(fun, x_0)$ in Matlab to get the numerical solutions. For instance, let $\underline{\delta} = 1$ and $\bar{\delta} = 2$. For $N = 2$, $\bar{z}_1^* = 0.2498$, $\bar{z}_2^* = 0.8638$, and $F(\delta^* \bar{z}_*^2) = 0.8994$ with $\delta^* = 1$ or 2. For $N = 1$ and $\bar{\delta} \to \infty$, the robust quantized filter will reduce to the SOI-KF.

10.4 A Numerical Example

Consider the following discrete time equations of motion for target tracking [3]:

$$x_{k+1} = \begin{bmatrix} 1 & \tau & \tau^2/2 \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} x_k + w_k, \tag{10.26}$$

$$y_k = [1\ 0\ 0] x_k + v_k, \tag{10.27}$$

where x_k is the state vector with its elements respectively denoting the target position, speed and acceleration at time k. w_k is a white noise sequence and is independent of the additive white noise v_k with variance R. When the sampling interval τ is sufficiently small, the covariance matrix of w_k is given by

$$Q = 2\alpha\sigma_m^2 \begin{bmatrix} \tau^5/20 & \tau^4/8 & \tau^3/6 \\ \tau^4/8 & \tau^3/3 & \tau^2/2 \\ \tau^3/6 & \tau^2/2 & \tau \end{bmatrix},$$

where σ_m^2 is the variance of the target acceleration and α is the reciprocal of the maneuver time constant. Let

$$\tau = 0.1s, \alpha = 0.1, \sigma_m^2 = 1, \sigma_R^2 = 10^{-2}.$$

It can be easily verified that $(A, Q^{1/2})$ and (C, A) are controllable and observable pairs. Corollary 10.1 implies the stability of the 1-bit and 2-bit quantized innovations Kalman filter, which are denoted by 1-LQ-KF and 2-LQ-KF, respectively. The initial state x_0 is a random vector with zero mean and covariance matrix [3]:

$$P_{0|0} = \begin{bmatrix} \sigma_R^2 & \sigma_R^2/\tau & 0 \\ \sigma_R^2/\tau & 2\sigma_R^2/\tau^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The filtering error variances of the position for 1-LQ-KF and SOI-KF are compared in Fig. 10.3, which shows that 1-LQ-KF outperforms SOI-KF. Figure 10.4 illustrates that for the 2-bit quantization, the computed variance by (10.21) is close to the one obtained by Monte Carlo simulations, indicating that the computed variance gives a good approximation to the true variance. The results for speed and acceleration estimates are similar and omitted. Next, we evaluate the robust mini-max quantizer where the prediction error variance is randomly perturbed by δ, which is uniformly

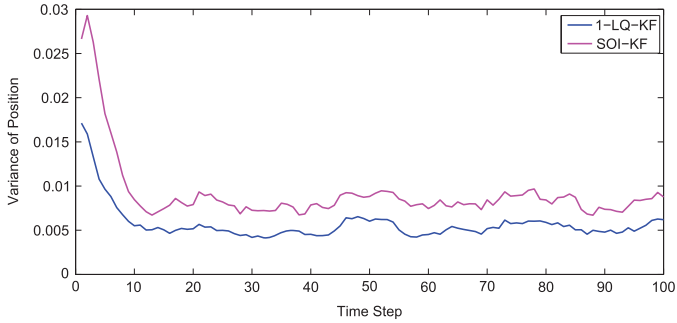

Fig. 10.3 Comparison of the error variance of the position estimate given by 1-LQ-KF and SOI-KF based on 500 samples

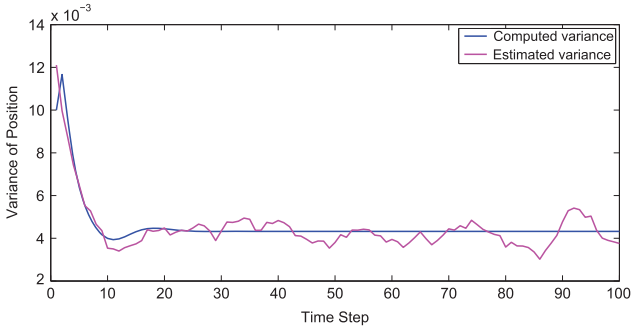

Fig. 10.4 Comparison of the position error variance computed by (10.21) and Monte Carlo method based on Monte Carlo methods with 500 samples

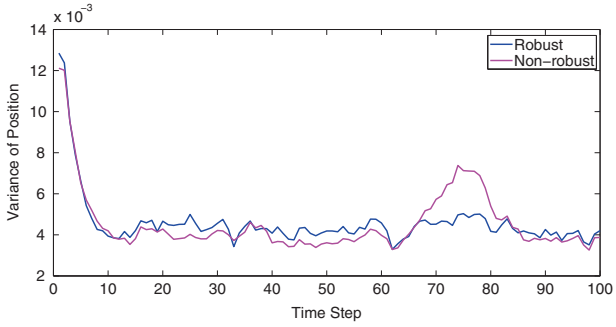

Fig. 10.5 Comparison of the position error variances by the robust and non-robust 2-bit quantized filters

distributed over [0.8, 1.2]. We design a 2-bit robust mini-max quantizer (2-LQ-RKF) according to (10.25) with $(\bar{z}^*)_1^2 = [0.3035\ 1.0094]$. For comparison, a non-robust quantizer (nominal) quantizer given in (10.3) is implemented as well. Figure 10.5 shows that 2-LQ-RKF has a better performance than the non-robust one.

10.5 Summary

Extending the existing work on SOI-KF, we have developed a general multi-level quantized innovations filter with a dead zone. It is established under the assumption that the system state conditioned on the quantized innovations is Gaussian, which is reasonable for quantizer with a moderate number of bits. The distinct feature of the quantized filter lies in its simplicity and efficiency. The convergence of the filter to the Kalman filter when the number of quantization levels goes to ∞ has also been established. The result is useful in applications such as WSNs whose communication capacity is limited.

References

1. A. Ribeiro, G. Giannakis, S. Roumeliotis, SOI-KF: distributed Kalman filtering with low-cost communications using the sign of innovations. IEEE Trans. Signal Process. **54**(12), 4782–4795 (2006)
2. J. Kotecha, P. Djuric, Gaussian particle filtering. IEEE Trans. Signal Process. **51**(10), 2592–2601 (2003)
3. R. Singer, Estimating optimal filter tracking performance for manned maneuvering targets. IEEE Trans. Aerosp. Electron. Syst. **6**(4), 473–483 (1970)

Chapter 11
LQG Control with Quantized Innovation Kalman Filter

In this chapter, we generalize the quantized innovation Kalman filter to a symmetric digital channel, and apply it to design the LQG control for discrete-time stochastic systems. The study of the quantized LQG control problem bears a long history and contains some misunderstandings in early works [1–3]. As pointed out in [4], it is claimed in [1] that the optimal quantized LQG control problem can be solved by optimizing the controller, the estimator and the quantizer separately which is found incorrect in [2, 3]. Instead of encoding the state directly, [5] introduces an "innovation process" coding scheme and shows that the separation principle remains valid for stable systems, which is extended to unstable systems in [6].

However, the aforementioned works exclusively focus on an error free channel, except the Gaussian channel in [6]. Noting that data packets transmitted over a real-time network may suffer not only from quantization distortion but also from transmission errors and packet dropouts. In this chapter, we look at a symmetric channel [7]. Motivated by the idea of quantizing innovations, whose control effects are removed before quantization, we consider the quantized LQG control problem subject to a symmetric channel and show that the separation principle remains valid.

The chapter is organized as follows. Problem formulation is delineated in Sect. 11.1. The investigation of the separation principle is made in Sect. 11.2. The quantized innovation Kalman filter over a symmetric channel is proposed in Sect. 11.3. An explicit suboptimal controller is given in Sect. 11.4. In Sect. 11.5, an illustrative example is included to demonstrate the performance of the controller. Conclusion remarks are drawn in Sect. 11.6.

11.1 Problem Formulation

Consider the following discrete linear time-invariant stochastic system:

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{11.1}$$

$$y_k = Cx_k + v_k, \tag{11.2}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_11

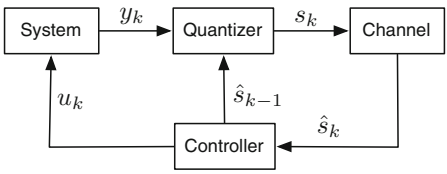

Fig. 11.1 Network configuration

where $x_k \in \mathbb{R}^n$ is the state, $y_k \in \mathbb{R}$ is the measured output, $u_k \in \mathbb{R}^m$ is the input, and $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}$ are two uncorrelated white additive Gaussian noises with variances Σ_w and σ_v^2, respectively. The random initial state x_0 is characterized by a Gaussian probability density function $p_{x_0}(\cdot)$ and is independent of the noises.

The network configuration under consideration is depicted in Fig. 11.1.

(a) **Quantizer**
The quantizer is geographically collocated with the system and can directly measure the output of the plant. Since our focus is on the communication limitation, it is assumed that the quantizer has sufficient computing power and memory required for state estimation and computation of a control signal.

(b) **Communication Channel**
The *forward* channel from the quantizer to the controller is a symmetric channel mapping an input symbol $i \in \S$ to an output symbol $j \in \S$ with a probability r_{ij}. Here the channel input and output take values from a finite set $\S \triangleq \{0, \ldots, c-1\}$. Such a symmetric channel is specified by a finite transmission matrix via

$$r_{ij} = \mathbb{P}\{\hat{s}_k = j | s_k = i\}$$

where $i \in \S$ is the transmitted symbol and $j \in \S$ is the actual symbol received by the controller. Moreover, the channel uncertainty is assumed to be independent of the channel input. The symmetric channel can also be modeled by

$$\hat{s}_k = s_k + z_k \pmod{c},$$

where z_k is an i.i.d. process taking values on the integers $\{0, \ldots, c-1\}$ and independent of the channel input s_k. The modulo operator "mod c" means that $\hat{s}_k \in \S$ is congruent to $s_k + z_k$ modulo "c". In particular, for an error free digital channel, the transmission matrix is expressed by

$$r_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases}$$

It is further assumed that there is a perfect *reverse* channel from the controller to quantizer. Thus, the quantizer has a full knowledge of the controller, except with one-time step delay.

(c) **Controller**
The first task of the controller is to estimate the state after receiving the quantized information. Then, the controller generates the control signal using the estimate of the state to achieve some performance objective.

Specifically, given a positive time horizon T, the aim is to find the optimal control policy $(u^*)^{T-1} = \{u_0^*, ..., u_{T-1}^*\}$ to minimize the linear quadratic cost functional

$$\begin{aligned} J(u^{T-1}) &= \mathbb{E}\left[x_T' K_T x_T + \sum_{k=0}^{T-1}(x_k' M_k x_k + u_k' R_k u_k)\right] \\ x_{k+1} &= Ax_k + Bu_k(\hat{s}^k, u^{k-1}) + w_k, \\ y_k &= Cx_k + v_k. \end{aligned} \tag{11.3}$$

Here the mathematical expectation is taken with respect to x_0, w^{T-1}, v^{T-1} and z^{T-1}. The specified weighting matrices $M_k \in \mathbb{R}^{p\times p}$ are positive semi-definite and $R_k \in \mathbb{R}$ are positive scalar numbers. The causal controller mapping $u_k(\cdot)$ is measurable with respect to $\sigma(\hat{s}^k, u^{k-1})$.

Remark 11.1 In the above, the encoder and decoder are lumped into the quantizer and controller respectively. The noisy *forward* channel with a noiseless feedback applies to wireless sensor networks, where the quantizer is built in a resource limited sensor while the controller is implemented in a resource abundant base station. Due to limited resources, the sensor may not correctly transmit the quantized signal while the base station has sufficient resources to perfectly feedback its received signal to the sensor via a *reverse* channel. The same communication scheme has been studied in [8]. It is worth mentioning that the method to be established can be extended to consider a lossy forward digital channel, in which the *reverse* channel plays the role of transmitting the packet reception acknowledgement.

11.2 Separation Principle

It is well-known that without the communication constraint, the above defined optimal LQG control problem is solved by the celebrated separation principle. In other words, the solution is obtained by first filtering the output to obtain a minimum mean square error variance estimate of the state and then the optimal control law is constructed by using the estimate and the optimal LQR control gain. This section will verify the validity of this classical separation principle under the network configuration in Fig. 11.1.

Before proceeding, we introduce the following notation to facilitate the presentation

$$\hat{x}_{k|k-1} \triangleq \mathbb{E}[x_k|\hat{s}^{k-1}, u^{k-1}], \tag{11.4}$$

$$\hat{y}_{k|k-1} \triangleq \mathbb{E}[y_k|\hat{s}^{k-1}, u^{k-1}] = C\hat{x}_{k|k-1} \tag{11.5}$$

and the corresponding estimation error covariance matrix

$$P_{k|k-1} \triangleq \mathbb{E}[(\hat{x}_{k|k-1} - x_k)(\hat{x}_{k|k-1} - x_k)'|\hat{s}^{k-1}, u^{k-1}]. \tag{11.6}$$

The updated state estimate is

$$\hat{x}_{k|k} \triangleq \mathbb{E}[x_k|\hat{s}^k, u^{k-1}] \triangleq \hat{x}_k \tag{11.7}$$

with the estimation error covariance matrix given by

$$P_{k|k} \triangleq \mathbb{E}[(\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)'|\hat{s}^k, u^{k-1}]. \tag{11.8}$$

It can be readily verified that

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k. \tag{11.9}$$

Similarly, the one-step state prediction error covariance is defined as

$$\begin{aligned} P_{k+1|k} &\triangleq \mathbb{E}[(\hat{x}_{k+1|k} - x_{k+1})(\hat{x}_{k+1|k} - x_{k+1})'|\hat{s}^k, u^{k-1}] \\ &= A\mathbb{E}[(\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)'|\hat{s}^k, u^{k-1}]A' + \mathbb{E}[w_k w_k'] \\ &= AP_{k|k}A' + \Sigma_w \end{aligned} \tag{11.10}$$

due to that $\mathbb{E}[(\hat{x}_{k|k} - x_k)w_k'|\hat{s}^k, u^{k-1}] = \mathbb{E}[w_k(\hat{x}_{k|k} - x_k)'|\hat{s}^k, u^{k-1}] = 0$ since w_k is independent of $\hat{x}_{k|k}$ and x_k.

Denote the innovation by $\varepsilon_k = y_k - \hat{y}_{k|k-1}$. We restrict ourselves to a class of time-varying quantizers, each of which is applied on the most recent innovation. Mathematically, the channel input at time k is computed by $s_k = \mathscr{E}_k(\varepsilon_k)$, where $\mathscr{E}_k(\cdot)$ is a time-varying quantizer. In the rest of the chapter, this class of quantizers will be denoted by $\mathscr{Q}$.

Given any control sequence $\{u_k\}_{t\geq 0}$, we know that

$$x_k = A^k x_0 + \sum_{i=0}^{k-1}(A^{k-1-i}Bu_i + w_i).$$

Define $\bar{x}_k = x_k - \sum_{i=0}^{k-1} A^{k-1-i}Bu_i$ as the state of the uncontrolled system and $\bar{y}_k = C\bar{x}_k + v_k$. The corresponding channel input and output are respectively defined as $\bar{s}_k$ and $\hat{\bar{s}}_k$, i.e., $\bar{s}_k = \mathscr{E}_k(\bar{y}_k - \hat{\bar{y}}_{k|k-1})$ and $\hat{\bar{s}}_k = \bar{s}_k + z_k$. Here $\hat{\bar{y}}_{k|k-1}$ is the prediction of $\bar{y}_k$ and is precisely given by $\hat{\bar{y}}_{k|k-1} = \mathbb{E}[\bar{y}_k|\hat{\bar{s}}^{k-1}]$. Let the state estimation error of the controlled and uncontrolled system be $\triangle_k = x_k - \hat{x}_k$ and $\bar{\triangle}_k = \bar{x}_k - \hat{\bar{x}}_k$, respectively.

Definition 11.1 ([9]) The control has *no dual effect* if for all $\{u_k\}_{t\geq 0}$ and $\forall t \geq 0$

$$\mathbb{E}[\triangle_k \triangle_k'|\hat{s}^k, u^{k-1}] = \mathbb{E}[\triangle_k \triangle_k'|\hat{\bar{s}}^k], \text{ P-almost sure.}$$

If the control has no dual effect, it states that the control only affects the evolution of the system and can not be used to reduce the state uncertainty.

Lemma 11.1 *For a quantizer acting on the most recent innovation, i.e.* $\{\mathscr{E}_k\}_{t\geq 0} \sqsubset \mathscr{Q}$, *the control law for the system in* (11.1) *and* (11.2) *is a certainty equivalent control law if and only if the control has no dual effect.*

Proof As in [6], the proof is completed by mimicking arguments in [9] and the detail is thus omitted.

The above lemma suggests that by using a quantizer acting on the most recent innovation, the estimation error covariance of the state is independent of control input. The intuitive explanation of this phenomenon lies in the fact that control effects have been already subtracted out from the innovation.

The following lemma [6] provides a necessary condition to guarantee a control without dual effects.

Lemma 11.2 *If $\sigma(\hat{\bar{s}}^k)$ is a sub σ-field of $\sigma(\hat{s}^k, u^{k-1})$, i.e., $\sigma(\hat{\bar{s}}^k) \sqsubset \sigma(\hat{s}^k, u^{k-1})$ and $\mathbb{E}[\bar{x}_k|\hat{\bar{s}}^k] = \mathbb{E}[\bar{x}_k|\hat{s}^k, u^{k-1}, \hat{\bar{s}}^k]$, there is no dual effect.*

Next, it is shown that under the network configuration in Fig. 11.1 with the quantizer acting on the innovation, conditions in Lemma 11.2 are satisfied. To elaborate it, random variables X, Y, Z are said to form a Markov chain in the order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. It is clear that $X \to Y \to Z$ implies $X \leftarrow Y \leftarrow Z$ [7]. Thus, we rewrite the Markov chain as $X \leftrightarrow Y \leftrightarrow Z$.

Proposition 11.1 *$\hat{\bar{s}}^k = \hat{s}^k$ and $\bar{x}_k \leftrightarrow \hat{s}^k \leftrightarrow u^{k-1}$ forms a Markov chain. Hence $\mathbb{E}[\bar{x}_k|\hat{s}^k, u^{k-1}] = \mathbb{E}[\bar{x}_k|\hat{s}^k] = \mathbb{E}[\bar{x}_k|\hat{\bar{s}}^k]$ and $\sigma(\hat{\bar{s}}^k) \sqsubset \sigma(\hat{\bar{s}}^k, u^{k-1}) = \sigma(\hat{s}^k, u^{k-1})$. That is, conditions in Lemma 11.2 are satisfied.*

Proof We extend the proof of [6] to our case. Due to that the controlled and uncontrolled systems have the same initial condition, it infers that $s_0 = \bar{s}_0$. Then $\hat{s}_0 = \hat{\bar{s}}_0$ since $\hat{s}_0 = s_0 + z_0 = \bar{s}_0 + z_0 = \hat{\bar{s}}_0$. Due to that u_0 is a function of $\hat{s}_0$, it implies $\bar{x}_1 \to \hat{s}_0 \to u_0$ or $\bar{x}_1 \leftrightarrow \hat{s}_0 \leftrightarrow u_0$ [7]. Thus, $s_1 = \mathscr{E}_1(y_1 - \hat{y}_{1|0}) = \mathscr{E}_1(C(x_1 - \hat{x}_{1|0}) + v_1) = \mathscr{E}_1(C(\bar{x}_1 - \mathbb{E}[\bar{x}_1|\hat{s}_0, u_0]) + v_1) = \mathscr{E}_1(C(\bar{x}_1 - \mathbb{E}[\bar{x}_1|\hat{\bar{s}}_0]) + v_1) = \mathscr{E}_1(C(\bar{x}_1 - \hat{\bar{x}}_{1|0}) + v_1) = \bar{s}_1$ and $\hat{s}_1 = \hat{\bar{s}}_1$. Furthermore, $\hat{s}_1 = \hat{\bar{s}}_1$ is independent of u_0, which together with the result $\bar{x}_1 \to \hat{s}_0 \to u_0$ leads to $\bar{x}_1 \to \hat{s}^1 \to u_0$.

Assume that $\hat{\bar{s}}^k = \hat{s}^k$ and $\bar{x}_k \to \hat{s}^k \to u^{k-1}, \forall 1 \le k \le t$. Since u_k is a function of $\hat{s}^k$ and u^{k-1}, u_k is independent of $(\bar{x}_k, w_k)$ conditioned on $\hat{s}^k$ and u^{k-1}, i.e., $(\bar{x}_k, w_k) \to (\hat{s}^k, u^{k-1}) \to u_k$. By induction hypothesis and the fact that w_k is independent of $\hat{s}^k, u^{k-1}$, it yields that $(\bar{x}_k, w_k) \to \hat{s}^k \to u^{k-1}$. Together with $(\bar{x}_k, w_k) \to (\hat{s}^k, u^{k-1}) \to u_k$, one can further derive $(\bar{x}_k, w_k) \to \hat{s}^k \to u^k$. This implies that $\bar{x}_{k+1} \to \hat{s}^k \to u_k$ as $\bar{x}_{k+1} = A\bar{x}_k + w_k$. Keeping in mind that $\hat{\bar{s}}^k = \hat{s}^k$, we similarly obtain $\hat{s}_{t+1} = \hat{\bar{s}}_{t+1}$. Finally, $\bar{x}_{k+1} \to \hat{s}^{k+1} \to u^k$ due to that $\hat{s}_{k+1} = \hat{\bar{s}}_{k+1}$ and $\hat{\bar{s}}_{k+1}$ is independent of v_{k+1}, u^k. Inductively, we obtain that $\hat{\bar{s}}^k = \hat{s}^k$ and $\bar{x}_k \to \hat{s}^k \to u^{k-1}, \forall k \in \mathbb{N}$.

Lemma 11.3 *Given a matrix $S \in \mathbb{R}^{p\times p}$, the following statement is true:*

$$\mathbb{E}[x_k' S x_k] = \mathbb{E}[\hat{x}_k' S \hat{x}_k] + tr(S\mathbb{E}[\triangle_k \triangle_k']), \tag{11.11}$$

where $tr(A)$ is defined as the summation of all diagonal elements of matrix A.

Proof The verification of the lemma is straightforward. First, we compute that

$$x_k' S x_k = (x_k - \hat{x}_k)' S (x_k - \hat{x}_k) + (x_k - \hat{x}_k)' S \hat{x}_k + \hat{x}_k' S (x_k - \hat{x}_k) + \hat{x}_k' S \hat{x}_k.$$

By the definition of $\hat{x}_k = x_k - \triangle_k$ and $\hat{x}_k$ is adapted to $\sigma(\hat{s}^k, u^{k-1})$, it follows that

$$\begin{aligned}
\mathbb{E}[x_k' S x_k | \hat{s}^k, u^{k-1}] &= \hat{x}_k' S \hat{x}_k + \mathbb{E}[\triangle_k' S \triangle_k | \hat{s}^k, u^{k-1}] \\
&= \hat{x}_k' S \hat{x}_k + \mathbb{E}[tr[S(\triangle_k \triangle_k')] | \hat{s}^k, u^{k-1}] \\
&= \hat{x}_k' S \hat{x}_k + tr(\mathbb{E}[S(\triangle_k \triangle_k') | \hat{s}^k, u^{k-1}]) \\
&= \hat{x}_k' S \hat{x}_k + tr(S \mathbb{E}[\triangle_k \triangle_k' | \hat{s}^k, u^{k-1}]). \qquad (11.12)
\end{aligned}$$

Jointly with the property of the conditional expectation [10] and linearity of the operators $\mathbb{E}[\cdot]$ and $tr(\cdot)$, (11.11) is obvious.

Proposition 11.2 *Given an arbitrary sequence of quantizers* $\{\mathscr{E}_k\}_{t\geq 0} \sqsubset \mathscr{Q}$, *the optimization of the cost functional in* (11.3) *with respect to* $u^{T-1} \triangleq \{u_0, \ldots, u_{T-1}\}$ *is expressed by*

$$\inf_{u^{T-1}} J(u^{T-1}) = \inf_{u^{T-1}} J'(u^{T-1}) + \sum_{k=0}^{T} tr(M_k \mathbb{E}[\bar{\triangle}_k \bar{\triangle}_k']), \qquad (11.13)$$

where the cost functional $J'(u^{T-1})$ *is given by*

$$J'(u^{T-1}) = \mathbb{E}[\hat{x}_T' M_k \hat{x}_T] + \sum_{k=0}^{T-1} \mathbb{E}[\hat{x}_k' M_k \hat{x}_k + u_k' R_k u_k]. \qquad (11.14)$$

Proof Substituting (11.11) into the cost functional (11.3), it is obvious that minimizing (11.3) is equivalent to the following optimization

$$\inf_{u^{T-1}} J(u^{T-1}) = \inf_{u^{T-1}} \left[J'(u^{T-1}) + \sum_{k=0}^{T} tr(M_k \mathbb{E}[\triangle_k \triangle_k']) \right] \qquad (11.15)$$

Since the control has no dual effect by Proposition 11.1, then

$$\mathbb{E}[\triangle_k \triangle_k'] = \mathbb{E}[\mathbb{E}[\triangle_k \triangle_k' | \hat{s}^k, u^k]] = \mathbb{E}[\mathbb{E}[\bar{\triangle}_k \bar{\triangle}_k' | \hat{\bar{s}}^k]] = \mathbb{E}[\bar{\triangle}_k \bar{\triangle}_k'].$$

In view of (11.15), the remaining part of the proof is trivial.

In the above, the optimization of the quadratic cost functional is decomposed into two independent parts: an optimal filtering problem to minimize the filtering error covariance $\mathbb{E}[\bar{\triangle}_k \bar{\triangle}_k']$ and an optimal control problem assuming full state observed.

Moreover, the dynamical equation of the fully observed state is described by

$$\begin{aligned}\hat{x}_{k+1} &= \mathbb{E}[x_{k+1}|\hat{s}^{k+1}, u^k] \\ &= \mathbb{E}[A(\hat{x}_k + \triangle_k)|\hat{s}^{k+1}, u^k] + Bu_k + \mathbb{E}[w_k|\hat{s}^{k+1}, u^k] \\ &= A\hat{x}_k + Bu_k + \beta_k, \end{aligned} \tag{11.16}$$

where $\beta_k \triangleq \mathbb{E}[A\triangle_k + w_k|\hat{s}^{k+1}, u^k]$ and the last equality is due to that $\hat{x}_k$ is adapted to $\sigma(\hat{s}^{k+1}, u^k)$. Note that a crucial gap between the above optimal control problem and the standard LQR problem in [11] is that the disturbance $\{\beta_k\}_{t\geq 0}$ in (11.16) does not satisfy the standard independence assumption.

Proposition 11.3 *The optimal control policy to minimize the cost functional in* (11.14) *subject to the dynamical equation of* (11.16) *is given by*

$$u_k^* = L_k\hat{x}_k, \quad \forall k \in \{0, \ldots, T-1\}$$

and the corresponding minimum cost is

$$J'((u^*)^{T-1}) = \mathbb{E}[\hat{x}_0'Q_0\hat{x}_0] + \sum_{k=0}^{T-1} tr(Q_{k+1}\mathbb{E}[\beta_k\beta_k'])$$

where the optimal control gain is given by

$$L_k = -(R_k + B'Q_{k+1}B)^{-1}B'Q_{k+1}A, \quad \forall k \in \{0, \ldots, T-1\} \tag{11.17}$$

and $\{Q_k\}_{k=0}^T$ *are obtained recursively by the backward RDE:*

$$Q_k = A'Q_{k+1}(I - B(B'Q_{k+1}B + R_k)^{-1}B'Q_{k+1})A + M_k, Q_T = M_T. \tag{11.18}$$

Proof Since $Q_T = M_T$, the cost functional $J'(u^{T-1})$ in (11.14) is rewritten as follows:

$$\begin{aligned} J'(u^{T-1}) = \mathbb{E}[\hat{x}_0'Q_0\hat{x}_0] &+ \sum_{k=0}^{T-1} \mathbb{E}[\hat{x}_k'M_k\hat{x}_k + u_k'R_ku_k] \\ &+ \sum_{k=0}^{T-1} \mathbb{E}[\hat{x}_{k+1}'Q_{k+1}\hat{x}_{k+1} - \hat{x}_k'Q_k\hat{x}_k]. \end{aligned}$$

By using the dynamical equation of (11.16), the above is further expressed by

$$\begin{aligned} J'(u^{T-1}) = \mathbb{E}[\hat{x}_0'Q_0\hat{x}_0] + \sum_{k=0}^{T-1} \mathbb{E}[&\hat{x}_k'(M_k - Q_k)\hat{x}_k + u_k'R_ku_k \\ &+ (A\hat{x}_k + Bu_k)'Q_{k+1}(A\hat{x}_k + Bu_k) \\ &+ \beta_k'Q_{k+1}\beta_k + 2\beta_k'Q_{k+1}(A\hat{x}_k + Bu_k)] \end{aligned}$$

$$= \mathbb{E}[\hat{x}_0' Q_0 \hat{x}_0] + \sum_{k=0}^{T-1} \mathbb{E}[\hat{x}_k'(M_k - Q_k)\hat{x}_k + u_k' R_k u_k + (A\hat{x}_k + Bu_k)' Q_{k+1}(A\hat{x}_k + Bu_k)] + \sum_{k=0}^{T-1} \mathbb{E}[\beta_k' Q_{k+1} \beta_k], \tag{11.19}$$

where the following fact is utilized in the second equality

$$\begin{aligned}\mathbb{E}[\beta_k' Q_{k+1}(A\hat{x}_k + Bu_k)|\hat{s}^k, u^k] &= \mathbb{E}[\beta_k|\hat{s}^k, u^k]' Q_{k+1}(A\hat{x}_k + Bu_k) \\ &= \mathbb{E}[\mathbb{E}[A\triangle_k + w_k|\hat{s}^{k+1}, u^k]|\hat{s}^k, u^k]' \\ &\quad \times Q_{k+1}(A\hat{x}_k + Bu_k) \\ &= \mathbb{E}[\triangle_k|\hat{s}^k, u^k]' A' Q_{k+1}(A\hat{x}_k + Bu_k) \\ &= 0\end{aligned}$$

since $\hat{x}_k$, u_k are adapted to $\sigma(\hat{s}^k, u^k)$ and $\mathbb{E}[\triangle_k|\hat{s}^k, u^k] = 0$.

Together with the backward RDE in (11.18), completing the square leads to that

$$\begin{aligned}J''(u^{T-1}) &\triangleq \sum_{k=0}^{T-1} \mathbb{E}[\hat{x}_k'(K_k - Q_k)\hat{x}_k + u_k' R_k u_k + (A\hat{x}_k + Bu_k)' Q_{k+1}(A\hat{x}_k + Bu_k)] \\ &= \sum_{k=0}^{T-1} \mathbb{E}[((R_k + B' Q_{k+1} B)u_k + B' Q_{k+1} A\hat{x}_k)'(R_k + B' Q_{k+1} B)^{-1} \\ &\quad \times ((R_k + B' Q_{k+1} B)u_k + B' Q_{k+1} A\hat{x}_k)] \\ &= \sum_{k=0}^{T-1} \mathbb{E}[(u_k - L_k\hat{x}_k)'(R_k + B' Q_{k+1} B)(u_k - L_k\hat{x}_k)].\end{aligned} \tag{11.20}$$

Hence, the optimal control policy is given by $u_k^* = L_k\hat{x}_k$, $\forall k \in \mathbb{N}$.

Observing that $\triangle_{k+1} = x_{k+1} - \hat{x}_{k+1} = A\triangle_k + w_k - \beta_k$, one can further derive that

$$\begin{aligned}\mathbb{E}[\beta_k \beta_k'] &= \mathbb{E}[(A\triangle_k + w_k - \triangle_{k+1})(A\triangle_k + w_k - \triangle_{k+1})'] \\ &= A\mathbb{E}[\triangle_k \triangle_k']A' + \Sigma_w + \mathbb{E}[\triangle_{k+1}\triangle_{k+1}'] - 2\mathbb{E}[(\triangle_{k+1} + \beta_k)\triangle_{k+1}'] \\ &= A\mathbb{E}[\triangle_k \triangle_k']A' + \Sigma_w - \mathbb{E}[\triangle_{k+1}\triangle_{k+1}'] \\ &= A\mathbb{E}[\bar{\triangle}_k \bar{\triangle}_k']A' + \Sigma_w - \mathbb{E}[\bar{\triangle}_{k+1}\bar{\triangle}_{k+1}'],\end{aligned}$$

where the second last equality is due to that

$$\mathbb{E}[\beta_k \triangle_{k+1}'] = \mathbb{E}[\beta_k(A\triangle_k + w_k - \beta_k)'] = 0.$$

Based on Propositions 11.2 and 11.3, we yield that given an arbitrary sequence of quantizers $\{\mathscr{E}_k\}_{t\geq 0} \sqsubset \mathscr{Q}$, the associated minimum cost functional is explicitly expressed as

$$J((u^*)^{T-1}) = \mathbb{E}[\hat{x}_0 Q_0 \hat{x}_0] + \sum_{k=0}^{T-1} tr(Q_{k+1}\Sigma_w) + D. \tag{11.21}$$

Here $D \triangleq \sum_{k=0}^{T} tr(M_k \mathbb{E}[\bar{\Delta}_k \bar{\Delta}_k']) + \sum_{k=0}^{T-1} tr(Q_{k+1}(A\mathbb{E}[\bar{\Delta}_k \bar{\Delta}_k']A' - \mathbb{E}[\bar{\Delta}_{k+1}\bar{\Delta}_{k+1}']))$ is independent of the control input. Finally, the quantized LQG control problem with the optimal quantizer in $\mathcal{Q}$ is converted into the following quantized estimation problem.

Proposition 11.4 *The optimal quantizers in $\mathcal{Q}$ to minimize the cost functional* (11.3) *are solved via the optimization*

$$(\mathcal{E}^*)^T = arg \inf_{\mathcal{E}^T \sqsubseteq \mathcal{Q}} D.$$

It should be noted that quantizers $(\mathcal{E}^*)^T$ are only optimal in the class of quantizers acting on the most recent innovation. In Sect. 11.4, the explicit form of $(\mathcal{E}^*)^T$ will be given under a Gaussian assumption on the predicted density.

11.3 State Estimator Design

As demonstrated in the previous section, the control problem for a certain class of quantizers is separated into two subproblems, one of which is to construct an estimator to produce the conditional expectation $\mathbb{E}[x_k|\hat{s}^k, u^k]$. If the estimator can be implemented as the classical Kalman filter, the linear form of the controller proposed in Proposition 11.3 would be particularly attractive from the application perspective.

In Chap. 10, under the assumption on the predicted density and an error free *forward* channel, we derived a recursive quantized innovations Kalman filter to minimize the mean square error. The main task of this section is to generalize the quantized filter to a symmetric digital channel.

Illustrated in [12], for a given estimation accuracy, a much lower number of bits is required to transmit the innovation than the measurement. Precisely, the quantizer receives an observation y_k at each time step, computes the normalized innovation

$$\bar{\varepsilon}_k \triangleq \varepsilon_k / \sqrt{CP_{k|k-1}C' + \sigma_v^2}$$

and quantizes it by a c-level($c \geq 2$) quantizer as follows

$$s_k \triangleq q(\bar{\varepsilon}_k) = \begin{cases} c-1, & \tau_{c-1} < \bar{\varepsilon}_k \\ \vdots & \vdots \\ 1, & \tau_1 < \bar{\varepsilon}_k \leq \tau_2 \\ 0, & \bar{\varepsilon}_k \leq \tau_1, \end{cases} \tag{11.22}$$

where $\tau_1 < \tau_2 < \cdots < \tau_{c-1}$ are quantizer thresholds. Note that there is a one-time step delay *reverse* channel from the controller to the quantizer, the quantizer can compute the one-time step ahead prediction estimate of the output and the prediction error variance for computing the normalized innovation. Similar to the case of extended Kalman filter or Gaussian filter [13], the following assumption is needed to derive a recursive and easily implementable quantized filter.

Assumption 11.1 The predicted density of the state is Gaussian, namely,

$$p(x_k|\hat{s}^{k-1}, u^{k-1}) = \mathcal{N}(x_k; \hat{x}_{k|k-1}, P_{k|k-1}),$$

where $\mathcal{N}(x_k; \hat{x}_{k|k-1}, P_{k|k-1})$ is the Gaussian probability density with mean $\hat{x}_{k|k-1}$ and covariance matrix $P_{k|k-1}$.

Using the conventions $\tau_0 = -\infty$ and $\tau_c = +\infty$, the quantized filter is explicitly given by the following proposition.

Proposition 11.5 *Consider the linear time-invariant stochastic system described in* (11.1) *and* (11.2) *and the network configuration in Fig.* 11.1 *with the quantizer taking the form of* (11.22)*. Under the Assumption* 11.1*, the quantized innovations Kalman filter to minimize the mean square error is updated by:*

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \frac{f(k, \hat{s}_k, \tau^c) P_{k|k-1} C'}{\sqrt{C P_{k|k-1} C' + \sigma_v^2}}, \tag{11.23}$$

$$P_{k|k} = P_{k|k-1} - F(\tau^c) \frac{P_{k|k-1} C' C P_{k|k-1}}{C P_{k|k-1} C' + \sigma_v^2}, \tag{11.24}$$

where the coefficient induced by quantization and channel uncertainty takes the form

$$f(k, \hat{s}_k, \tau^c) \triangleq \sum_{k=0}^{c-1} I_{\{k\}}(\hat{s}_k) \frac{\sum_{j=0}^{c-1} r_{jk}[g(\tau_j) - g(\tau_{j+1})]}{\sum_{j=0}^{c-1} r_{jk}[T(\tau_j) - T(\tau_{j+1})]}$$

and the performance recovery factor $F(\tau^c)$ *is computed by*

$$F(\tau^c) = \sum_{k=0}^{c-1} \frac{\left[\sum_{j=0}^{c-1} r_{jk}[g(\tau_j) - g(\tau_{j+1})]\right]^2}{\sum_{j=0}^{c-1} r_{jk}[T(\tau_j) - T(\tau_{j+1})]},$$

where $g(\cdot)$ and $T(\cdot)$ are respectively the density function and the tail probability function of a standard Gaussian random variable.

Proof By using the tower property of the conditional expectation [10], it follows that

$$\hat{x}_{k|k} = \mathbb{E}[x_k|\hat{s}^k, u^{k-1}] = \mathbb{E}[\mathbb{E}[x_k|\hat{s}^{k-1}, y_k, u^{k-1}]|\hat{s}^k, u^{k-1}]. \tag{11.25}$$

Under the Gaussian Assumption 11.1, the posterior density $p(x_k|\hat{s}^{k-1}, y_k)$ is obtained as

$$\begin{aligned} p(x_k|\hat{s}^{k-1}, y_k) &= \frac{p(x_k|\hat{s}^{k-1})p(y_k|x_k)}{\int_{\mathbb{R}^n} p(x_k|\hat{s}^{k-1})p(y_k|x_k)dx_k} \\ &= \mathcal{N}(x_k; \hat{x}_{k|k-1}, P_{k|k-1})\mathcal{N}(y_k; Cx_k, \sigma_v^2). \end{aligned} \tag{11.26}$$

By following the technique of the Kalman filter, the inner conditional expectation in (11.25) is computed by

$$\hat{x}^*_{k|k} \triangleq \mathbb{E}[x_k|\hat{s}^{k-1}, y_k, u^{k-1}] = \hat{x}_{k|k-1} + \frac{P_{k|k-1}C'}{CP_{k|k-1}C' + \sigma_v^2}(y_k - C\hat{x}_{k|k-1}). \tag{11.27}$$

On the other hand, if the controller receives a symbol $\hat{s}_k = k, \forall k \in \{0, \ldots, c-1\}$, then

$$\begin{aligned} \mathbb{E}[\bar{\varepsilon}_k|\hat{s}^{k-1}, \hat{s}_k = k, u^{k-1}] &= \int_{\mathbb{R}} \bar{\varepsilon}_k p(\bar{\varepsilon}_k|\hat{s}^{k-1}, \hat{s}_k = k)d\bar{\varepsilon}_k \\ &= \frac{1}{c_k}\int_{\mathbb{R}} \bar{\varepsilon}_k p(\bar{\varepsilon}_k|\hat{s}^{k-1})p(\hat{s}_k = k|\bar{\varepsilon}_k)d\bar{\varepsilon}_k \\ &= \frac{1}{c_k}\sum_{j=0}^{c-1}\int_{\mathbb{R}} \bar{\varepsilon}_k p(\bar{\varepsilon}_k|\hat{s}^{k-1})p(\hat{s}_k = k, s_k = j|\bar{\varepsilon}_k)d\bar{\varepsilon}_k \\ &= \sum_{j=0}^{c-1}\frac{r_{jk}}{c_k}\int_{\mathbb{R}} \bar{\varepsilon}_k p(\bar{\varepsilon}_k|\hat{s}^{k-1})p(s_k = j|\bar{\varepsilon}_k)d\bar{\varepsilon}_k \\ &= \sum_{j=0}^{c-1}\frac{r_{jk}}{c_k}\int_{\tau_j}^{\tau_{j+1}} \bar{\varepsilon}_k p(\bar{\varepsilon}_k|\hat{s}^{k-1})d\bar{\varepsilon}_k \\ &= \sum_{j=0}^{c-1}\frac{r_{jk}}{c_k}[g(\tau_j) - g(\tau_{j+1})], \end{aligned} \tag{11.28}$$

where the normalization factor c_k in the second equality is given by

$$\begin{aligned} c_k &\triangleq \int_{\mathbb{R}} p(\bar{\varepsilon}_k|\hat{s}^{k-1})p(\hat{s}_k = k|\bar{\varepsilon}_k)d\bar{\varepsilon}_k \\ &= \sum_{j=0}^{c-1} r_{jk}[T(\tau_j) - T(\tau_{j+1})]. \end{aligned} \tag{11.29}$$

Together with (11.25) and (11.27), the estimate is recursively updated by

$$\begin{aligned}\hat{x}_{k|k} &= \mathbb{E}[\hat{x}^*_{k|k}|\hat{s}^k, u^{k-1}] = \hat{x}_{k|k-1} + \frac{P_{k|k-1}C'}{\sqrt{CP_{k|k-1}C' + \sigma_v^2}}\mathbb{E}[\bar{\varepsilon}_k|\hat{s}^k, u^{k-1}] \\ &= \hat{x}_{k|k-1} + \frac{f(k, \hat{s}_k, \tau^c)P_{k|k-1}C'}{\sqrt{CP_{k|k-1}C' + \sigma_v^2}}. \end{aligned} \tag{11.30}$$

Similar to (11.28), one can show that $\mathbb{E}[f^2(k, \hat{s}_k, \tau^c)] = F(\tau^c)$.

Under the Gaussian assumption, we further obtain that

$$\mathbb{E}[(x_k - \hat{x}_{k|k-1})f(k, \hat{s}_k, \tau^c)] = F(\tau^c)\frac{P_{k|k-1}C'}{\sqrt{CP_{k|k-1}C' + \sigma_v^2}}. \tag{11.31}$$

Thus, the estimation error covariance matrix is derived via

$$\begin{aligned} P_{k|k} &= \mathbb{E}[(\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)'] \\ &= \mathbb{E}[(\hat{x}_{k|k-1} - x_k)(\hat{x}_{k|k-1} - x_k)'] + \frac{P_{k|k-1}C'CP_{k|k-1}}{CP_{k|k-1}C' + \sigma_v^2}\mathbb{E}[f^2(k, \hat{s}_k, \tau^c)] \\ &\quad - 2\mathbb{E}[(x_k - \hat{x}_{k|k-1})f(k, \hat{s}_k, \tau^c)]\frac{CP_{k|k-1}}{\sqrt{CP_{k|k-1}C' + \sigma_v^2}} \\ &= P_{k|k-1} - F(\tau^c)\frac{P_{k|k-1}C'CP_{k|k-1}}{CP_{k|k-1}C' + \sigma_v^2}. \end{aligned}$$

It is clear that the matrix $P_{k|k}$ is a decreasing function with respect to the *performance recovery factor* $F(\tau^c)$. This implies that the corresponding optimal quantized innovations Kalman filter can be derived by finding the optimal quantization thresholds to maximize the above performance recovery factor $F(\tau^c)$, namely,

$$(\tau_1^*, \ldots, \tau_{c-1}^*) = \arg\max_{\tau^c} F(\tau^c). \tag{11.32}$$

A numerical solution can be obtained by evoking the Matlab command, e.g.,

$$[x, y] = \text{fmincon}(f\ un, x_0, A, b, Aeq, beq, lb, ub).$$

It is worthy mentioning that the quantizer is designed in an off-line manner. Thus, a numerical solution is efficient for the implementation of the quantizer. In addition, if $r_{ij} = 0, i \neq j$, then $r_{jj} = 1$ since $\sum_{j=0}^{c-1} r_{ij} = 1$. This corresponds to a noiseless channel that the quantized signal is correctly received by the controller. For this special case, Table 11.1 lists numerical solutions to the optimization (11.32) and the corresponding *performance recovery factor* for various quantization levels. Moreover, it is

Table 11.1 Solutions to (11.32) for noiseless channels

c = 2	c = 3	c = 4	c = 5
$\tau_1^* = 0$	$\tau_1^* = -0.612$ $\tau_2^* = 0.612$	$\tau_1^* = -0.982$ $\tau_2^* = 0$ $\tau_3^* = 0.982$	$\tau_1^* = -1.244$ $\tau_2^* = -0.382$ $\tau_3^* = 0.382$ $\tau_4^* = 1.244$
$F((\tau^*)^2) = 0.637$	$F((\tau^*)^3) = 0.810$	$F((\tau^*)^4) = 0.883$	$F((\tau^*)^5) = 0.920$

interesting that if packet dropout process of the channel is an i.i.d. binary Bernoulli process $\{\gamma_k\}_{t\geq 0}$ with packet dropout rate r and there is no transmission error, the corresponding quantized filter is similarly derived by defining

$$f(k, \hat{s}_k, \tau^c) = \gamma_k \sum_{k=0}^{c-1} I_{\{k\}}(\hat{s}_k) \frac{g(\tau_k) - g(\tau_{k+1})}{T(\tau_k) - T(\tau_{k+1})}, \tag{11.33}$$

$$F(\tau^c) = (1-r) \sum_{k=0}^{c-1} \frac{[g(\tau_k) - g(\tau_{k+1})]^2}{T(\tau_k) - T(\tau_{k+1})}, \tag{11.34}$$

where $\gamma_k = 1$ indicates that the packet is successfully received by the controller while $\gamma_k = 0$ associates a packet dropout. This implies that for the i.i.d. packet dropout, the optimal quantization thresholds remain the same of [12], but the performance recovery factor further relies on the successful transmission probability $1 - r$.

11.4 Controller Design

In the previous section, we have designed a convenient quantized estimator for the state with the form of Kalman filtering algorithm. The computations are organized recursively so that only the most recent quantized symbol $\hat{s}_k$ and control u_{k-1} are required at time t, together with $\mathbb{E}[x_{k-1}|\hat{s}^{k-1}, u^{t-2}]$ in order to produce $\mathbb{E}[x_k|\hat{s}^k, u^{k-1}]$. This section will concentrate on finding a suboptimal control law based on the proposed optimal quantized innovations Kalman filter.

Proposition 11.6 *Given an arbitrary sequence of quantizers taking the form of* (11.22) *and under the Assumption* 11.1*, the optimal control policy is given by*

$$u_k^* = L_k \hat{x}_k$$

and the corresponding cost is

$$J((u^*)^{T-1}) = \mathbb{E}[\hat{x}_0' Q_0 \hat{x}_0] + \sum_{k=0}^{T} tr(M_k P_{k|k}) + \sum_{k=0}^{T-1} tr[Q_{k+1}(P_{k+1|k} - P_{k+1|k+1})]$$

$$= \mathbb{E}[\hat{x}_0' Q_0 \hat{x}_0] + tr(M_0 P_{0|0}) + \sigma_v^2 \sum_{k=0}^{T-1} tr(Q_{k+1})$$
$$+ \sum_{k=0}^{T-1} tr(L_k' B' Q_{k+1} A P_{k|k}) \tag{11.35}$$

where the control gain L_k is given in (11.17)*, the estimate $\hat{x}_k$ and $P_{k|k}$ are recursively computed by* (11.9)*,* (11.10)*,* (11.23) *and* (11.24)*.*

Proof The optimality of the control policy follows from the certainty equivalence property. Noting that

$$\mathbb{E}[\beta_k \beta_k'] = P_{k+1|k} - P_{k+1|k+1},$$

the minimum cost is calculated from Propositions 11.2 and 11.3 and (11.35) is derived by some simple algebraic manipulations.

Proposition 11.7 *Under the Assumption* 11.1*, the optimal quantizers in Proposition* 11.4 *are solved by* (11.22) *with quantizer thresholds computed in* (11.32)*.*

Proof In light of (11.35) in Proposition 11.6, the optimal quantized innovation Kalman filter will result in the minimum cost functional. Thus, the assertion is true.

It should be noted that Propositions 11.6 and 11.7 provide a suboptimal solution to the quantized LQG problem due to the restriction of the type of quantizers and the Assumption 11.1. However, the simulation in the sequel will illustrate the performance of the designed controller.

11.5 An Illustrative Example

Consider the discrete system describing a cart with an inverted pendulum hinged on the top of the cart [14]

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Bw_k, \\ y_k &= [1\ 0\ 0\ 0]x_k + v_k, \end{aligned} \tag{11.36}$$

where the matrices A and B are respectively given by

$$A = \begin{bmatrix} 1.0000 & 0.1000 & -0.0412 & -0.0014 \\ 0 & 1.0000 & -0.8323 & -0.0412 \\ 0 & 0 & 1.0577 & 0.1019 \\ 0 & 0 & 1.1652 & 1.0577 \end{bmatrix}, B = \begin{bmatrix} 0.0084 \\ 0.1678 \\ -0.0042 \\ -0.0849 \end{bmatrix}.$$

Different from [14], we assume that the horizontal external force exerted on the cart is corrupted by an additive white Gaussian noise w_k. Let weighting matrices be constant

$$M_k = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,0 \end{bmatrix}$$

and $R_k = 1$. For comparison, assume the *forward* channel is error free. The purpose is to solve the quantized LQG control problem with the quantized innovations Kalman filter and compare the performance of the designed controller with that of the standard optimal LQG controller.

To this aim, the initial value x_0 is taken from a 4-dimensional standard Gaussian random vector with zero mean and covariance matrix $P_0 = 9 \cdot I_4$. We use the optimal 1-bit quantized innovations Kalman filter with optimal quantization thresholds $-\tau_1^* = \tau_2^* = 0.612$ and $F((\tau^*)^3) = 0.8098$. Using the controller in Proposition 11.6, Fig. 11.2 shows that states of the closed-loop system respectively based on the Kalman filter and the optimal 1-bit quantized innovations Kalman filter come close to each other. Similar conclusion applies to the control input, see the Fig. 11.3. To compare the control performance, the minimum average cost functional, i.e., $J((u^*)^{k-1})/k$, based on the optimal 1-bit quantized innovations Kalman filter and that of Kalman filter are illustrated in Fig. 11.4. The computed minimum average cost based on the optimal 1-bit quantized innovations Kalman filter is derived from (11.35). The estimated minimum average cost for the optimal 1-bit quantized innovations Kalman filter is calculated from the Monte Carlo simulation using 500 samples, which shows the consistency of the derived minimum cost in (11.35). This example

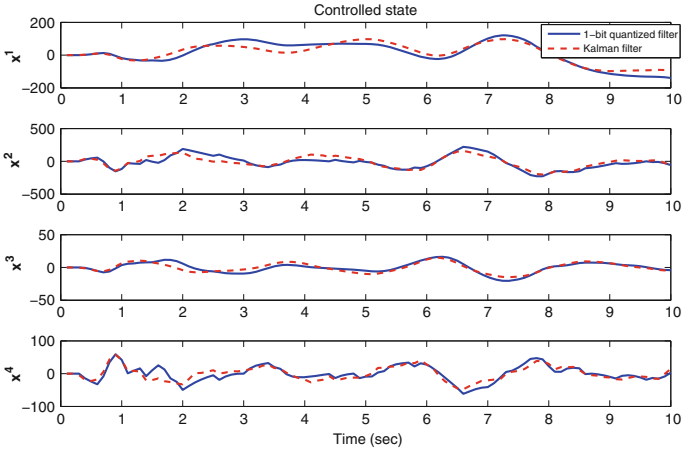

Fig. 11.2 The state of the closed-loop system based on the Kalman filter and the optimal 1-bit quantized innovations Kalman filter

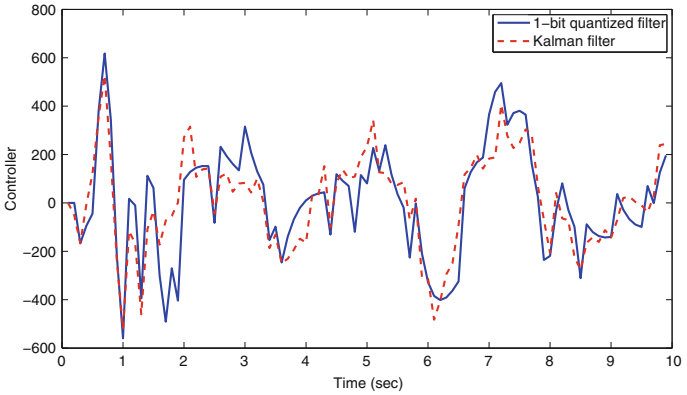

Fig. 11.3 Control input based on the Kalman filter and the optimal 1-bit quantized innovations Kalman filter

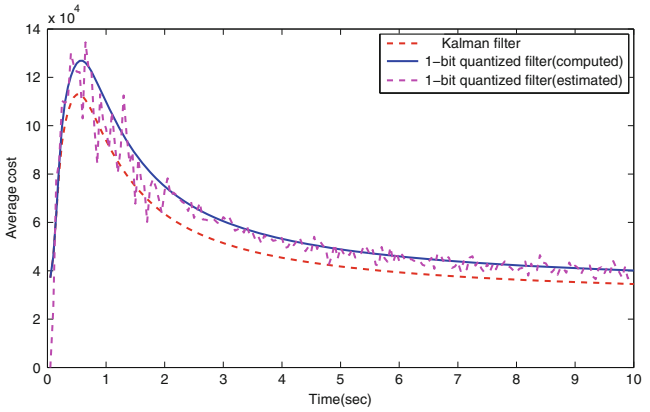

Fig. 11.4 The average cost based on the Kalman filter and the optimal 1-bit quantized innovations Kalman filter

demonstrates that adopting the optimal 1-bit quantized innovations Kalman filer, our proposed controller performance comes close to that of the standard optimal LQG controller.

11.6 Summary

We studied the quantized LQG control problem over a symmetric channel and established the separation principle for a certain class of quantizers. A suboptimal quantized innovations Kalman filter has been derived which is optimal under the Gaussian assumption for the predicted density to minimize the mean square error. Based on the quantized innovations Kalman filter, a suboptimal control law which has the same

complexity as the standard LQG controller and can be easily implemented in the real world was proposed. An example was provided to compare the performance of the designed LQG controller with that of the standard optimal LQG controller.

References

1. R. Larson, Optimum quantization in dynamic systems. IEEE Trans. Autom. Control **12**(2), 162–168 (1967)
2. R. Marleau, J. Negro, Comments on "Optimum quantization in dynamic systems". IEEE Trans. Autom. Control **17**(2), 273–274 (1972)
3. R. Larson, E. Tse, Reply. IEEE Trans. Autom. Control **17**(2), 274–276 (1972)
4. M. Fu, Linear quadratic Gaussian control with quantized feedback, in *American Control Conference*, pp. 2172–2177 (2009)
5. V. Borkar, S. Mitter, LQG control with communication constraints. *Communications, Computation, Control and Signal Processing: A Tribute to Thomas Kailath* (Kluwer, Norwell, 1997)
6. S. Tatikonda, A. Sahai, S. Mitter, Stochastic linear control over a communication channel. IEEE Trans. Autom. Control **49**(9), 1549–1561 (2004)
7. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
8. A. Sahai, S. Mitter, The necessity and sufficiency of anytime capacity for control over a noisy communication link. Part I: scalar systems. IEEE Trans. Inf. Theory **52**(8), 3369–3395 (2006)
9. Y. Bar-Shalom, E. Tse, Dual effect, certainty equivalence, and separation in stochastic control. IEEE Trans. Autom. Control **19**(5), 494–500 (1974)
10. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)
11. D. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, 2 (Athena Scientific, Belmont, 1995)
12. K. You, L. Xie, S. Sun, W. Xiao, Multiple-level quantized innovation Kalman filter, in *Proceedings of 17th IFAC World Congress*, pp. 1420–1425 (2008)
13. K. Ito, K. Xiong, Gaussian filters for nonlinear filtering problems. IEEE Trans. Autom. Control **45**(5), 910–927 (2000)
14. C. Chen, *Linear System: Theory and Design* (Saunders College Publishing, Philadelphia, 1984)

Chapter 12
Kalman Filtering with Faded Measurements

This chapter focuses on the network requirement for ensuring the stability of a remote Kalman filter with faded measurements, where the fading channels undergo transmission failure and signal fluctuation simultaneously.

The chapter is organized as follows. The networked estimation problem is formulated in Sect. 12.1. In Sect. 12.2, some preliminary results on modified algebraic Riccati operators (MAROs) and modified Lyapunov operator (MLOs) are derived for later developments. After that, necessary and sufficient conditions on the network for the stability of the mean error covariance matrix of the remote Kalman filter are presented in terms of the unstable poles of the plant. Lower and upper bounds for the mean error covariance matrix are provided in the form of a modified Lyapunov iteration and a modified Riccati iteration, respectively. A simulation is given in Sect. 12.3 to demonstrate the results, and Sect. 12.4 concludes the chapter.

12.1 Problem Formulation

Consider a discrete-time networked system as shown in Fig. 12.1, where an unreliable network with fading channel(s) is placed between the plant and the filter.

The plant is assumed to be linear with a state-space realization:

$$\begin{aligned} x_{k+1} &= Ax_k + w_k, \\ y_k &= Cx_k + v_k, \end{aligned} \tag{12.1}$$

where $x_k \in \mathbb{R}^n$ is the state, and $y_k \in \mathbb{R}^\ell$ is the measured output. The process noise $w_k \in \mathbb{R}^n$ and the measurement noise $v_k \in \mathbb{R}^\ell$ are white processes with zero means and covariance matrices $Q > 0$ and $R_v > 0$, respectively. Without loss of generality, assume that A is unstable, (C, A) is detectable, and $C = [C_1^T \; C_2^T \cdots C_\ell^T]^T$ has full-row rank.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_12

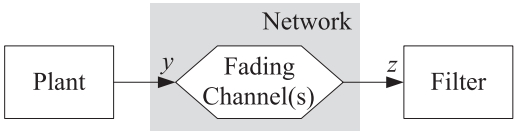

Fig. 12.1 Filtering over output fading channel(s)

The model of fading channel(s) is given by

$$z_k = \xi_k y_k + n_k, \tag{12.2}$$

where $n_k \in \mathbb{R}^\ell$ denotes the channel additive noise, which is white with zero mean and covariance matrix $R_n > 0$, and $\xi_k \in \mathbb{R}^{\ell\times\ell}$ represents the channel fading, which is assumed to have the diagonal structure:

$$\xi_k = \text{diag}\{\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{\ell,k}\}. \tag{12.3}$$

Remark 12.1 Different from a general MIMO communication system with multiple receiving and transmitting antennas [1], the input and the output of the network (12.2)–(12.3) are assumed to have the same dimension since our main focus is to investigate the effect of fading on the stability of a remote filter rather than recovery of the transmitted information.

It is assumed that the network experiences both transmission failure and signal fluctuation as

$$\xi_{i,k} = \gamma_{i,k}\Omega_{i,k}, \tag{12.4}$$

where $\gamma_{i,k}$ is $0/1$ valued (0 for "failure", 1 for "success"), white and identically distributed with probability distribution

$$\Pr\{\gamma_{i,k} = 0\} = \alpha_i,\ \Pr\{\gamma_{i,k} = 1\} = 1 - \alpha_i,\ 0 \le \alpha_i < 1, \tag{12.5}$$

and $\Omega_{i,k}$ is nonzero, white and identically distributed with a given continuous distribution function such as Nakagami. Furthermore, let

$$\begin{aligned}
\Sigma_\xi &\triangleq [\mathbb{E}\{\xi_{i,k}\xi_{j,k}\} - \mathscr{E}\{\xi_{i,k}\}\mathscr{E}\{\xi_{j,k}\}]_{i,j=1,2,\ldots,\ell},\\
\Pi_\xi &\triangleq \mathbb{E}\{\text{diag}\{\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{\ell,k}\}\},\\
\gamma_k &= \text{diag}\{\gamma_{1,k}, \gamma_{2,k}, \ldots, \gamma_{\ell,k}\},\\
\Omega_k &= \text{diag}\{\Omega_{1,k}, \Omega_{2,k}, \ldots, \Omega_{\ell,k}\},
\end{aligned}$$

and Σ_γ, Σ_Ω, Π_γ, Π_Ω are defined similarly to Σ_ξ, Π_ξ. Note that the fading experienced by different channels may or may not be correlated depending on whether a non-orthogonal or an orthogonal access scheme [2] is adopted. Further assume that the distribution of ξ_k is known and $\{w_k\}_{k\ge0}$, $\{v_k\}_{k\ge0}$, $\{n_k\}_{k\ge0}$, $\{\gamma_k\}_{k\ge0}$, $\{\Omega_k\}_{k\ge0}$ are uncorrelated with each other.

Remark 12.2 For Kalman filtering with faded measurements studied in [3], in order to apply the results on Kalman filtering with random parameters presented in [4, 5], one of the technical assumptions is that the channel fading $\xi_{i,k}$ is positive almost surely for all time $k \geq 0$ and every channel $i = 1, 2, \ldots, \ell$, which excludes the existence of possible transmission failure. The model (12.2)–(12.4) in this chapter considers transmission failure and signal fluctuation at the same time, for which the results in [4, 5] cannot be applied.

We assume that the exact value of ξ_k is known at the filter at time k. In this situation, it is easy to prove that the time-varying Kalman filter is optimal in the MMSE sense. The plant together with the network can be rewritten as

$$\begin{aligned} x_{k+1} &= Ax_k + w_k, \\ z_k &= \xi_k Cx_k + \xi_k v_k + n_k. \end{aligned} \tag{12.6}$$

Denote $\tilde{C}_k = \xi_k C$ and $\tilde{v}_k = \xi_k v_k + n_k$. The state prediction error covariance matrix P_k can be computed by the following random Riccati iteration:

$$P_{k+1} = AP_kA^T + Q - AP_k\tilde{C}_k^T\left[\tilde{C}_kP_k\tilde{C}_k^T + \tilde{R}_k\right]^{-1}\tilde{C}_kP_kA^T \tag{12.7}$$

with $\tilde{R}_k = \xi_k R_v \xi_k + R_n$ representing the covariance matrix of $\tilde{v}_k$. Since P_{k+1} is a function of ξ_k, it is random in nature and thus can only be computed for a given realization of ξ_0^k. Therefore, we turn to consider the stability of its expectation defined as follows.

Definition 12.1 The Kalman filter over the fading channel(s) (12.2) or the random Riccati iteration (12.7) is said to be mean covariance stable if $\sup_{k\geq 0}\mathbb{E}\{P_k\} < \infty$, where the expectation is taken over the channel fading sequence ξ_0^k.

12.2 Stability Analysis of Kalman Filter with Fading

In this section, we shall study the mean covariance stability of the Kalman filtering over fading channels.

12.2.1 Preliminaries

Define an MARO and an MLO as below:

$$g(X) \triangleq AXA^T + Q - AXC^T\Pi\left[\Pi CXC^T\Pi + \Sigma \odot (CXC^T) + R\right]^{-1}\Pi CXA^T, \tag{12.8}$$

$$h(X) \triangleq \alpha AXA^T + Q, \tag{12.9}$$

where $\Sigma \geq 0,\ R > 0,\ \Pi > 0$ is a diagonal matrix, and $\alpha \geq 0$ is a scalar. The next two lemmas present several properties of the MARO (12.8) and the MLO (12.9).

Lemma 12.1 *Let*

$$\begin{aligned} g_1(X) &\triangleq AXA^T - AXC^T\Pi\left[\Pi CXC^T\Pi + \Sigma \odot (CXC^T)\right]^{-1}\Pi CXA^T, \\ \phi_1(L,X) &\triangleq (A + L\Pi C)X(A + L\Pi C)^T + L[\Sigma \odot (CXC^T)]L^T, \\ \phi(L,X) &\triangleq \phi_1(L,X) + LRL^T + Q. \end{aligned}$$

Then, the following properties hold.

(i) *For any* $b \in \mathbb{R}$, $\phi_1(L, bX) = b\phi_1(L,X)$.
(ii) *For any* $X \geq 0$ *and* L, *we have* $h(X) \geq Q$, *and*

$$0 \leq g_1(X) = \phi_1(L_1^*(X), X) \leq \phi_1(L,X), \tag{12.10}$$
$$Q \leq g(X) = \phi(L^*(X), X) \leq \phi(L,X), \tag{12.11}$$

with

$$\begin{aligned} L_1^*(X) &= -AXC^T\Pi\left[\Sigma \odot (CXC^T)\right]^{-1}, \\ L^*(X) &= -AXC^T\Pi\left[\Sigma \odot (CXC^T) + R\right]^{-1}. \end{aligned}$$

(iii) *For any* $X_2 \geq X_1 \geq 0$, *we have*

$$h(X_2) \geq h(X_1),\ g_1(X_2) \geq g_1(X_1),\ g(X_2) \geq g(X_1).$$

Proof (i) Straightforward.

(ii) The inequality $h(X) \geq Q$ is obvious. It follows directly from the expressions of $L_1^*(X)$ and $L^*(X)$ that $g_1(X) = \phi(L_1^*(X), X)$ and $g(X) = \phi(L^*(X), X)$. Based on the property of Hadamard product in Lemma A.5, we have $\Sigma \odot (CXC^T) \geq 0$ since both Σ and CXC^T are positive semidefinite. Thus, $\phi_1(L,X) \geq 0$ and $\phi(L,X) \geq Q$ for any $X \geq 0$ and L. Considering the partial derivatives $\partial\phi_1(L,X)/\partial L$ and $\partial\phi(L,X)/\partial L$, we can conclude that $\phi_1(L_1^*(X), X) \leq \phi_1(L,X)$ and $\phi(L^*(X), X) \leq \phi(L,X)$.

(iii) Observe that $h(X)$, $\phi_1(L,X)$ and $\phi(L,X)$ are affine in X. It is direct to see that $h(X_2) \geq h(X_1)$. We also have

$$\begin{aligned} g_1(X_1) &= \phi_1(L_1^*(X_1), X_1) \leq \phi_1(L_1^*(X_2), X_1) \leq \phi_1(L_1^*(X_2), X_2) = g_1(X_2), \\ g(X_1) &= \phi(L^*(X_1), X_1) \leq \phi(L^*(X_2), X_1) \leq \phi(L^*(X_2), X_2) = g(X_2), \end{aligned}$$

which completes the proof. □

Lemma 12.2 *The following statements are equivalent.*

(i) *There exists* $X > 0$ *such that* $X > g_1(X)$.
(ii) *There exist* $X > 0$ *and* L *such that* $X > \phi_1(L, X)$.
(iii) *There exists* $X > 0$ *such that* $X > g(X)$.
(iv) *There exist* $X > 0$ *and* L *such that* $X > \phi(L, X)$.

Moreover, if any of the conditions (i)–(iv) *holds, then the claim as below is true.*

(v) *The sequence* $\{X_k\}_{k\geq 0}$ *computed by* $X_{k+1} = g(X_k)$ *with any initial condition* $X_0 \geq 0$ *is bounded from above and convergent as* k *approaches* ∞, *i.e.,* $\lim_{k\to\infty} X_k = \tilde{X}$, *where* $\tilde{X}$ *is the unique positive-semidefinite fixed point of the MARO* (12.8), *i.e.,* $\tilde{X} = g(\tilde{X}),\ \tilde{X} \geq 0$.

Proof (i)⇔(ii) If (i) holds, then we can derive that $X > \phi_1(L, X)$ by setting $L = L_1^*(X)$. On the other hand, when (ii) is true, we have that

$$X > \phi_1(L, X) \geq \phi_1(L_1^*(X), X) = g_1(X).$$

(i)⇐(iii) It follows directly by choosing the same X.
(i)⇒(iii) Suppose that (i) is true. By continuity, it holds that

$$X > AXA^T - AXC^T\Pi\left[\Pi CXC^T\Pi + \Sigma \odot (CXC^T) + \theta I\right]^{-1}\Pi CXA^T + \beta I$$

for sufficiently small scalars $\beta > 0$ and $\theta > 0$. Therefore, for any scalar $\delta > 0$, we have

$$\begin{aligned}\delta X &> \delta AXA^T - \delta AXC^T\Pi\left[\Pi CXC^T\Pi + \Sigma \odot (CXC^T) + \theta I\right]^{-1}\Pi CXA^T + \delta\beta I \\ &= A(\delta X)A^T - A(\delta X)C^T\Pi\left[\Pi C(\delta X)C^T\Pi + \Sigma \odot (C(\delta X)C^T) + \delta\theta I\right]^{-1} \\ &\quad \times \Pi C(\delta X)A^T + \delta\beta I.\end{aligned}$$

By selecting $\delta > 0$ such that $\delta\theta I \geq R$ and $\delta\beta I \geq Q$ and letting $\hat{X} = \delta X > 0$, we obtain

$$\begin{aligned}\hat{X} &> A\hat{X}A^T - A\hat{X}C^T\Pi\left[\Pi C\hat{X}C^T\Pi + \Sigma \odot (C\hat{X}C^T) + \delta\theta I\right]^{-1}\Pi C\hat{X}A^T + \delta\beta I \\ &\geq A\hat{X}A^T - A\hat{X}C^T\Pi\left[\Pi C\hat{X}C^T\Pi + \Sigma \odot (C\hat{X}C^T) + R\right]^{-1}\Pi C\hat{X}A^T + Q \\ &= g(\hat{X}),\end{aligned}$$

which completes the proof.

(iii)⇔(iv) It can be proved similarly to the proof of the equivalence between (i) and (ii).

(i)⇒(v) Assume that (i) holds. In this situation, we can always choose $b_1 \in [0, 1)$, $b_2 \in (0, \infty)$ such that $g_1(X) \le b_1 X$, $X_0 \le b_2 X$, and $L_1^*(X) R L_1^*(X)^T + Q \le b_2 X$. It follows from Lemma 12.1 that

$$\begin{aligned} X_1 &= g(X_0) = \phi(L^*(X_0), X_0) \\ &\le \phi(L_1^*(X), X_0) \\ &\le \phi(L_1^*(X), b_2 X) \\ &= \phi_1(L_1^*(X), b_2 X) + L_1^*(X) R L_1^*(X)^T + Q \\ &= b_2 g_1(X) + L_1^*(X) R L_1^*(X)^T + Q \\ &\le (b_1 b_2 + b_2) X, \\ X_2 &= g(X_1) \\ &\le g((b_1 b_2 + b_2) X) \\ &\le (b_1^2 b_2 + b_1 b_2 + b_2) X. \end{aligned}$$

By induction, it can be shown that

$$X_k = \sum_{i=0}^{t} b_1^i b_2 X \le \frac{b_2}{1 - b_1} X, \ \forall k \ge 0,$$

which implies that the sequence $\{X_k\}_{k \ge 0}$ is bounded from above for any initial $X_0 \ge 0$.

Next, the existence of the limit of $\{X_k\}_{k \ge 0}$ will be shown by using its boundedness proved above and the monotonicity property shown in Lemma 12.1 (iii), which extends the proof of [6, Theorem 1] for the packet-loss case to the general fading case.

Case 1: $X_0 = 0$. It is easy to see by induction that $X_{k+1} \ge X_k$ for all $k \ge 0$. In this situation, since $\{X_k\}_{k \ge 0}$ is monotonically nondecreasing and bounded from above, it is convergent to $\tilde{X}$, which is positive semidefinite by continuity and is also the fixed point of the MARO (12.8).

Case 2: $X_0 \ge \tilde{X}$. We can derive that $X_k \ge \tilde{X}$ for all $k \ge 0$ following an inductive argument. Note that

$$\begin{aligned} X_{k+1} - \tilde{X} &= g(X_k) - g(\tilde{X}) \\ &= \phi(L^*(X_k), X_k) - \phi(L^*(\tilde{X}), \tilde{X}) \\ &\le \phi(L^*(\tilde{X}), X_k) - \phi(L^*(\tilde{X}), \tilde{X}) \\ &= \phi_1(L^*(\tilde{X}), X_k - \tilde{X}) \\ &= \phi_1^{k+1}(L^*(\tilde{X}), X_0 - \tilde{X}). \end{aligned}$$

It follows from $Q > 0$ that $\tilde{X} = \phi(L^*(\tilde{X}), \tilde{X}) > \phi_1(L^*(\tilde{X}), \tilde{X}) \ge 0$, which further implies that $\lim_{k \to \infty} \phi_1^{k+1}(L^*(\tilde{X}), X_0 - \tilde{X}) = 0$ for $X_0 \ge \tilde{X}$, i.e., the sequence $\{X_k\}_{k \ge 0}$ is also convergent to $\tilde{X}$ for Case 2. For any possible X_0, we always have $0 \le X_0 \le$

$X_0+\tilde{X}$. Then, the convergence of $\{X_k\}_{k\geq 0}$ follows from Cases 1 and 2 and the squeeze theorem [7, p. 64]. The uniqueness of $\tilde{X}$ can be easily shown by contradiction. □

Note that the solution $\tilde{X} \geq 0$ to $\tilde{X} = g(\tilde{X})$ can be obtained by solving the optimization [6, Theorem 6]

$$\tilde{X} = \arg\max_{X\geq 0} \operatorname{tr}(X) \tag{12.12}$$

subject to the LMI constraint

$$\begin{bmatrix} X - AXA^T - Q & AXC^T\Pi \\ \Pi CXA^T & -\Pi CXC^T\Pi - \Sigma \odot (CXC^T) - R \end{bmatrix} \leq 0.$$

The next theorem provides explicit conditions in terms of the Mahler measure of the plant for ensuring the existence of X to $X > g_1(X)$. To this end, suppose that the pair (A, C) has the following Wonham decomposition [8]:

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ \star & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \star & \cdots & \star & A_\ell \end{bmatrix}, \quad C = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ \star & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \star & \cdots & \star & c_\ell \end{bmatrix}, \tag{12.13}$$

where $\star$ represents terms that will not be used in the derivation, $A_i \in \mathbb{R}^{n_i\times n_i}$, $c_i \in \mathbb{R}^{1\times n_i}$, $\sum_{i=1}^{\ell} n_i = n$, and each pair (A_i, c_i) is detectable.

Theorem 12.1 *There exists $X > 0$ such that $X > g_1(X)$ if*

$$1 + \frac{[\Pi]_{ii}^2}{[\Sigma_1]_{ii}} > \mathscr{M}(A_i)^2, \quad \forall i = 1, 2, \ldots, \ell, \tag{12.14}$$

and only if

$$\prod_{i=1}^{\ell}\left(1 + \frac{[\Pi]_{ii}^2}{[\Sigma_2]_{ii}}\right) > \mathscr{M}(A)^2, \tag{12.15}$$

where Σ_1 and Σ_2 are any positive-semidefinite diagonal matrices satisfying $0 \leq \Sigma_2 \leq \Sigma \leq \Sigma_1$.

The next result (see [9, Lemma 5.4]) is needed in the proof of Theorem 12.1.

Lemma 12.3 *Suppose that A is unstable, (c, A) is detectable for a row vector $c \neq 0$. Then, there exists $X > 0$ such that*

$$X > AXA^T - \nu AXc^T(cXc^T)^{-1}cXA^T \tag{12.16}$$

if and only if

$$\nu > 1 - \frac{1}{\mathscr{M}(A)^2}. \tag{12.17}$$

Proof of Theorem 12.1 First of all, it is easy to see that the existence of $X > 0$ to $X > g_1(X)$ is invariant under similarity transformations on the pair (C, A). Without loss of generality, let (C, A) have the form

$$A = \begin{bmatrix} A_{\mathrm{s}} & 0 \\ 0 & A_{\mathrm{u}} \end{bmatrix}, \quad C = [C_{\mathrm{s}} \;\; C_{\mathrm{u}}],$$

after a similarity transformation, where A_{s} is stable, all the eigenvalues of A_{u} are either on or outside the unit circle, and $(C_{\mathrm{u}}, A_{\mathrm{u}})$ is observable. In this case, we can prove that the existence of $X > 0$ to $X > g_1(X)$ is further equivalent to the existence of $X_{\mathrm{u}} > 0$ to

$$X_{\mathrm{u}} > A_{\mathrm{u}} X_{\mathrm{u}} A_{\mathrm{u}}^T - A_{\mathrm{u}} X_{\mathrm{u}} C_{\mathrm{u}}^T \Pi \left[\Pi C_{\mathrm{u}} X_{\mathrm{u}} C_{\mathrm{u}}^T \Pi + \Sigma \odot (C_{\mathrm{u}} X_{\mathrm{u}} C_{\mathrm{u}}^T) \right]^{-1} \Pi C_{\mathrm{u}} X_{\mathrm{u}} A_{\mathrm{u}}^T.$$

Therefore, in the sequel we assume that all the eigenvalues of A are either on or outside the unit circle.

First, we will prove the necessity of the condition (12.15). Suppose that there exists $X > 0$ such that $X > g_1(X)$. Based on the property of Hadamard product documented in Lemma A.5, we have $\Sigma \odot (CXC^T) \geq \Sigma_2 \odot (CXC^T)$. The condition (12.15) holds when any $[\Sigma_2]_{ii} = 0$, otherwise it follows from $X > g_1(X)$ that

$$\begin{aligned}
\det(X) &> \det(AXA^T - AXC^T \Pi \left[\Pi CXC^T \Pi + \Sigma \odot (CXC^T) \right]^{-1} \Pi CXA^T) \\
&= \det(A)^2 \det(X) \det \left(I - \left[\Pi CXC^T \Pi + \Sigma \odot (CXC^T) \right]^{-1} \Pi CXC^T \Pi \right) \\
&= \det(A)^2 \det(X) \left(\det(I + \Pi CXC^T \Pi \left[\Sigma \odot (CXC^T) \right]^{-1}) \right)^{-1} \\
&= \det(A)^2 \det(X) \left(\det(I + X^{1/2} C^T \Pi \left[\Sigma \odot (CXC^T) \right]^{-1} \Pi CX^{1/2}) \right)^{-1} \\
&\geq \det(A)^2 \det(X) \left(\det(I + X^{1/2} C^T \Pi \left[\Sigma_2 \odot (CXC^T) \right]^{-1} \Pi CX^{1/2}) \right)^{-1} \\
&= \det(A)^2 \det(X) \\
&\quad \times \left(\det(I + \left[\Sigma_2 \odot (CXC^T) \right]^{-1/2} \Pi CXC^T \Pi \left[\Sigma_2 \odot (CXC^T) \right]^{-1/2}) \right)^{-1} \\
&\geq \mathscr{M}(A)^2 \det(X) \prod_{i=1}^{\ell} \left(\frac{[\Sigma_2]_{ii}}{[\Sigma_2]_{ii} + [\Pi]_{ii}^2} \right),
\end{aligned} \tag{12.18}$$

where the last inequality follows from Hadamard's inequality; see Lemma A.2. Thus, the proof of necessity is completed since (12.18) implies (12.15).

Next, the sufficiency of the condition (12.14) will be shown. According to Lemma 12.3 and letting $\nu = \frac{[\Pi]_{ii}^2}{[\Pi]_{ii}^2+[\Sigma_1]_{ii}}$, if (12.14) is true, then there exists $X_i > 0$ such that

$$X_i > A_i X_i A_i^T - A_i X_i c_i^T [\Pi]_{ii} \left[[\Pi]_{ii} c_i X_i c_i^T [\Pi]_{ii} + [\Sigma_1]_{ii} \odot (c_i X_i c_i^T) \right]^{-1} \times [\Pi]_{ii} c_i X_i A_i^T,$$

for every $i = 1, 2, \ldots, \ell$, which, based on Lemma 12.2, is also equivalent to the existence of $X_i > 0$ and L_i such that

$$X_i > (A_i + L_i[\Pi]_{ii}c_i)X_i(A_i + L_i[\Pi]_{ii}c_i)^T + L_i \left([\Sigma_1]_{ii} \odot (c_i X_i c_i^T) \right) L_i^T. \quad (12.19)$$

The sufficiency for the case $\ell = 1$ is straightforward. We turn to consider the case $\ell = 2$. Since the existence of $X > 0$ to $X > g_1(X)$ is invariant under similarity transformations on the pair (C, A), we let the pair (C, A) have the form given in (12.13), namely

$$A = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}, \quad C = \begin{bmatrix} c_1 & 0 \\ c_{21} & c_2 \end{bmatrix}.$$

Based on (12.19), we can always choose a sufficiently large θ_1 such that

$$\theta_1 X_2 > (A_2 + L_2[\Pi]_{22}c_2)\theta_1 X_2 (A_2 + L_2[\Pi]_{22}c_2)^T + L_2 \left([\Sigma_1]_{22} \odot (c_2 \theta_1 X_2 c_2^T) \right) L_2^T + F(X_1),$$

where

$$\begin{aligned} F(X_1) = {} & L_2 \left([\Sigma_1]_{22} \odot (c_{21} X_1 c_{21}^T) \right) L_2^T \\ & + (A_{21} + L_2[\Pi]_{22}c_{21})X_1(A_1 + L_1[\Pi]_{11}c_1)^T \\ & \times \Big[X_1 - (A_1 + L_1[\Pi]_{11}c_1)X_1(A_1 + L_1[\Pi]_{11}c_1)^T \\ & \quad - L_1 \left([\Sigma_1]_{11} \odot (c_1 X_1 c_1^T) \right) L_1^T \Big]^{-1} \\ & \times (A_1 + L_1[\Pi]_{11}c_1)X_1(A_{21} + L_2[\Pi]_{22}c_{21})^T. \end{aligned}$$

It then follows from Schur's complement in Lemma A.1 that there exist $X = \text{diag}\{X_1, \theta_1 X_2\}$ and $L = \text{diag}\{L_1, L_2\}$ such that

$$X > \phi_1(L, X). \quad (12.20)$$

By induction, for the case $\ell > 2$, there exist $X = \text{diag}\{X_1, \theta_1 X_2, \ldots, \theta_{\ell-1} X_\ell\}$ and $L = \text{diag}\{L_1, L_2, \ldots, L_\ell\}$ with sufficiently large $\theta_1, \theta_2, \ldots, \theta_{\ell-1}$ such that (12.20) holds. It follows from Lemma 12.2 and

$$\Sigma_1 \odot (CXC^T) \geq \Sigma \odot (CXC^T)$$

that $X > g_1(X)$, which completes the proof. □

Remark 12.3 The condition (12.14) depends on the specific Wonham decomposition in (12.13), and both (12.14) and (12.15) are related to the structure of Σ. Obviously, we can put $\Sigma_1 = \Sigma_2 = \Sigma$ when Σ is diagonal, and the bounds in (12.14) and (12.15) are the same for the single-output case ($\ell = 1$).

12.2.2 Mean Covariance Stability

Now, we are in the position to present the main result of this chapter that provides necessary and sufficient conditions on the network for mean covariance stability of a remote Kalman filter with faded observations along with lower and upper bounds for the mean error covariance matrix.

Theorem 12.2 *The Kalman filter over the fading channel(s)* (12.2) *is mean covariance stable if*

$$1 + \frac{[\Pi_\xi]_{ii}^2}{[\Sigma_{\xi 1}]_{ii}} > \mathscr{M}(A_i)^2, \quad \forall i = 1, 2, \ldots, \ell, \tag{12.21}$$

and only if

$$\frac{1}{\alpha} > \rho(A)^2, \tag{12.22}$$

where $\Sigma_{\xi 1}$ is any positive-semidefinite matrix satisfying $\Sigma_{\xi 1} \geq \Sigma_\xi$, and

$$\alpha = \Pr\{\xi_{1,k} = \xi_{2,k} = \cdots = \xi_{\ell,k} = 0\}. \tag{12.23}$$

Moreover, it holds that

$$h^k(P_0) \leq \mathbb{E}\{P_k\} \leq g^k(P_0)$$

for all $P_0 \geq 0$ and $k \geq 0$, where $g(\cdot)$ and $h(\cdot)$ are defined in (12.8) *and* (12.9) *with*

$$\Pi = \Pi_\xi, \ \Sigma = \Sigma_\xi, \ R = \Pi_\xi R_v \Pi_\xi + \Sigma_\xi \odot R_v + R_n, \tag{12.24}$$

and α as defined in (12.23). *If* (12.21) *is true, then $\lim_{k\to\infty} g^k(P_0) = \tilde{P}_{g1} \geq 0$ with $g(\tilde{P}_{g1}) = \tilde{P}_{g1}$. If* (12.22) *is true, then $\lim_{k\to\infty} h^k(P_0) = \tilde{P}_h \geq 0$ with $h(\tilde{P}_h) = \tilde{P}_h$.*

Proof We will first prove the sufficiency of the condition (12.21). It follows from (12.7) that, for any scalar $\varepsilon > 0$,

$$\begin{bmatrix} P_{k+1} - AP_kA^T - Q - \varepsilon I & AP_k\tilde{C}_k^T \\ \tilde{C}_kP_kA^T & -\tilde{C}_kP_k\tilde{C}_k^T - \tilde{R}_k \end{bmatrix} < 0, \tag{12.25}$$

which is affine in both P_k and P_{k+1}. By noting the independence between P_k and ξ_k, taking the expectation on both sides of (12.25) and denoting $\hat{P}_k \triangleq \mathbb{E}\{P_k\}$, we have

$$\begin{bmatrix} \hat{P}_{k+1} - A\hat{P}_kA^T - Q - \varepsilon I & A\hat{P}_kC^T\Pi_\xi \\ \Pi_\xi C\hat{P}_kA^T & -\Pi_\xi C\hat{P}_kC^T\Pi_\xi - \Sigma_\xi \odot (C\hat{P}_kC^T) - R \end{bmatrix} < 0,$$

which further implies that

$$\begin{aligned} \hat{P}_{k+1} < & -A\hat{P}_kC^T\Pi_\xi \left[\Pi_\xi C\hat{P}_kC^T\Pi_\xi + \Sigma_\xi \odot (C\hat{P}_kC^T) + R\right]^{-1} \Pi C\hat{P}_kA^T \\ & + A\hat{P}_kA^T + Q + \varepsilon I. \end{aligned} \tag{12.26}$$

Using the continuity arguments with respect to ε, (12.26) yields $\hat{P}_{k+1} \le g(\hat{P}_k)$ by letting ε approach 0. By taking $X_0 = P_0, X_{k+1} = g(X_k)$ and noting that $\hat{P}_k \le X_k$ for all k, we can easily see from Theorem 12.1 and Lemma 12.2 that the sequence $\{\hat{P}_k\}_{k\ge 0}$ is bounded from above, i.e., the remote Kalman filter is mean covariance stable. In addition, $\lim_{k\to\infty} g^k(P_0) = \tilde{P}_{g1} = g(\tilde{P}_{g1}) \ge 0$ follows from Lemma 12.2.

Next, we will consider the necessity of the condition (12.22). Based on the definition of mathematical expectation, we can derive that $\hat{P}_0 = P_0$,

$$\begin{aligned} \hat{P}_1 &= \mathbb{E}\{P_1\} \\ &= \mathbb{E}\{AP_0A^T + Q - AP_0\tilde{C}_0^T\left[\tilde{C}_0P_0\tilde{C}_0^T + \tilde{R}_0\right]^{-1}\tilde{C}_0P_0A^T\} \\ &\ge \Pr\{\xi_{1,k} = \xi_{2,k} = \cdots = \xi_{\ell,k} = 0\}\mathbb{E}\{AP_0A^T\} + Q \\ &= h(\hat{P}_0), \end{aligned}$$

and

$$\hat{P}_{k+1} \ge h(\hat{P}_k) \ge h^{k+1}(\hat{P}_0) = h^{k+1}(P_0).$$

Since $(A, Q^{1/2})$ is controllable, there exists $\tilde{P}_h > 0$ such that $\tilde{P}_h = h(\tilde{P}_h)$ if and only if $\rho(\sqrt{\alpha}A) < 1$, i.e., the condition (12.22) holds. We will show the necessity by contradiction. Assume that $P_0 = 0$ and (12.22) does not hold. In this case, $\tilde{P}_h$ does not exist. Based on Lemma 12.1 (iii), we can easily show that $\{h^k(0)\}_{k\ge 0}$ is monotonically nondecreasing and thus unbounded, which contradicts the mean covariance stability of the remote Kalman filter. Thus, the proof of the necessity of (12.22) is completed. Furthermore, if (12.22) is true, then, following a similar line of proof as in Lemma 12.2, we obtain $\lim_{k\to\infty} h^k(P_0) = \tilde{P}_h$ for any P_0. □

Note that a tighter sufficient condition than (12.21) can be obtained under the following technical assumption.

Assumption 12.1 There exists a matrix $\tilde{R}_n \geq 0$ such that $\mathbb{E}\{\Omega_k^{-1} R_n \Omega_k^{-1}\} \leq \tilde{R}_n$.

Remark 12.4 The term $\Omega_k^{-1} R_n \Omega_k^{-1}$ is nonlinear in Ω_k, and thus its expectation can only be evaluated case by case for the given distribution of Ω_k.

Proposition 12.2.1 *Under Assumption 12.1, the Kalman filter over the fading channel(s) (12.2) is mean covariance stable if*

$$\frac{1}{\alpha_i} > \mathscr{M}(A_i)^2, \quad \forall i = 1, 2, \ldots, \ell, \tag{12.27}$$

where α_i is as defined in (12.5). Moreover, we have

$$\mathbb{E}\{P_k\} \leq g^k(P_0), \ \forall P_0 \geq 0, \ k \geq 0,$$

where $g(\cdot)$ is defined in (12.8) with

$$\Pi = \Pi_\gamma, \ \Sigma = \Sigma_\gamma, \ R = \Pi_\gamma R_v \Pi_\gamma + \Sigma_\gamma \odot R_v + \tilde{R}_n. \tag{12.28}$$

If (12.27) is true, then $\lim_{k\to\infty} g^k(P_0) = \tilde{P}_{g2} \geq 0$ with $g(\tilde{P}_{g2}) = \tilde{P}_{g2}$.

Proof It follows from (12.25) that

$$\begin{bmatrix} P_{k+1} - AP_kA^T - Q - \varepsilon I & AP_kC^T\gamma_k \\ \gamma_k CP_kA^T & -\gamma_k(CP_kC^T + R_v)\gamma_k - \Omega_k^{-1} R_n \Omega_k^{-1} \end{bmatrix} < 0. \tag{12.29}$$

Under Assumption 12.1, we can derive that $\hat{P}_{k+1} \leq g(\hat{P}_k)$ with Π, Σ, R given in (12.28). The rest of the proof follows analogously to the proof of Theorem 12.2. □

Remark 12.5 As we can see from Proposition 12.2.1, the signal fluctuation does not affect the mean covariance stability of the remote Kalman filter under Assumption 12.1, however, it would influence the upper bound for $\mathbb{E}\{P_k\}$ through the term $\tilde{R}_n$. Under Assumption 12.1, the condition (12.22) is still necessary for mean covariance stability and $h^k(P_0)$ provides a lower bound for $\mathbb{E}\{P_k\}$. It is also direct to see that (12.21) always implies (12.27), and they are consistent if there is no signal fluctuation, i.e., $\Omega_{i,k} \equiv 1$ in (12.4). The optimization (12.12) can be used to compute $\tilde{P}_{g1}$ and $\tilde{P}_{g2}$, while $\tilde{P}_h$ can be obtained by solving a Lyapunov equation.

12.3 A Numerical Example

The numerical example that follows shows the validity of the results in this chapter.

Example 12.1 Consider the plant with $\ell = 1$ and

$$A = \begin{bmatrix} 0 & 0.1 & 1.2 \\ 0 & 0.5 & 0 \\ 1.3 & 0.1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

and let $Q = I_3$, $R_v = R_n = 1$. We have $\rho(A) = 1.2490$, $\mathscr{M}(A) = 1.5600$. Suppose that $\Omega_{1,k}$ is Nakagami distributed with the mean channel power gain $\varpi_1 = 2$ and the severity of fading $\varrho_1 = 2$. In this case, $\mathbb{E}\{\Omega_k^{-1} R_n \Omega_k^{-1}\} = 1$, and we can choose $\tilde{R}_n = 1$ in Assumption 12.1. Based on Theorem 12.2, the sufficient condition for the mean covariance stability of the remote Kalman filter is $\alpha_1 < 0.2050$, while the corresponding necessary condition is $\alpha = \alpha_1 < 0.6410$. According to Proposition 12.2.1, the sufficient condition for the mean covariance stability of the remote Kalman filter becomes $\alpha_1 < 0.4109$, which is less conservative than $\alpha_1 < 0.2050$ obtained from Theorem 12.2.

Next, assume that $\alpha_1 = 0.2$. Figure 12.2 shows the path of $\xi_{1,k}$ for one sample of simulation and the empirical norm of P_k by averaging 10,000 Monte Carlo simulations. As we can see, the Kalman filter with faded observations in this situation is mean covariance stable. In addition, we can compute that

$$\tilde{P}_{g1} = \begin{bmatrix} 29.9731 & 0.1248 & 0.0753 \\ 0.1248 & 1.3332 & 0.0972 \\ 0.0753 & 0.0972 & 20.0949 \end{bmatrix}, \quad \tilde{P}_{g2} = \begin{bmatrix} 12.0830 & 0.1218 & 0.0600 \\ 0.1218 & 1.3331 & 0.0924 \\ 0.0600 & 0.0924 & 7.6722 \end{bmatrix},$$

$$\tilde{P}_h = \begin{bmatrix} 1.4307 & 0.0120 & 0.0039 \\ 0.0120 & 1.0526 & 0.0121 \\ 0.0039 & 0.0121 & 1.4863 \end{bmatrix}.$$

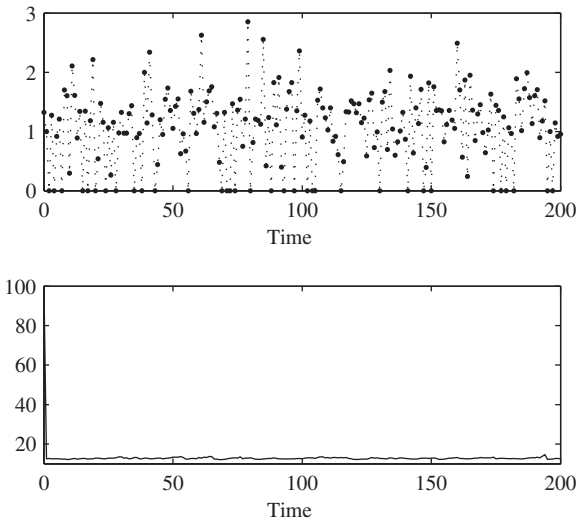

Fig. 12.2 The path of $\xi_{1,k}$ for one sample of simulation (*above*) and the empirical norm of P_k by averaging 10,000 Monte Carlo simulations (*below*) when $\alpha_1 = 0.2$

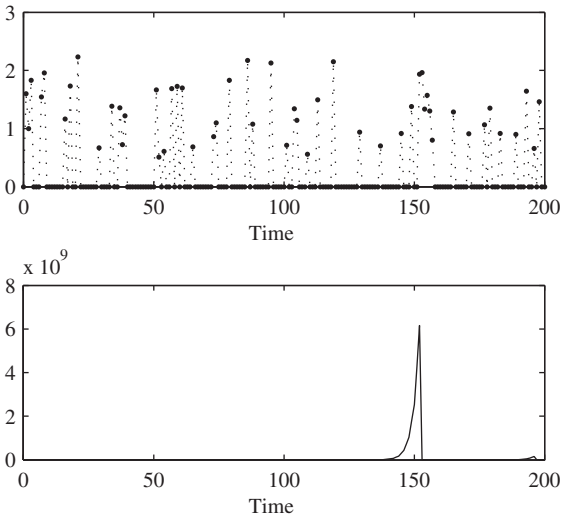

Fig. 12.3 The paths of $\xi_{1,k}$ (*above*) and $\|P_k\|$ (*below*) for one sample of simulation when $\alpha_1 = 0.7$

When $\alpha_1 = 0.7$, the norm of P_k can reach a relatively high value for one sample of simulation as shown in Fig. 12.3.

12.4 Summary

In this chapter, necessary and sufficient conditions on the network for guaranteeing the mean covariance stability of Kalman filtering over fading channels subject to both transmission failure and signal fluctuation have been given in terms of the unstable poles of the plant. It has been shown that the mean error covariance matrix of the remote Kalman filter is bounded from above by a modified Riccati iteration and from below by a modified Lyapunov iteration.

The results in this chapter are based mainly on [10].

References

1. T. Koch, A. Lapidoth, The fading number and degrees of freedom in non-coherent MIMO fading channels: a peace pipe, in *Proceedings of IEEE International Symposium on Information Theory*, pp. 661–665 (2005)
2. A. Goldsmith, *Wireless Communications* (Cambridge University Press, Cambridge, 2005)
3. S. Dey, A. Leong, J. Evans, Kalman filtering with faded measurements. Automatica **45**(10), 2223–2233 (2009)
4. P. Bougerol, Kalman filtering with random coefficients and contractions. SIAM J. Control Optim. **31**, 942–959 (1993)

5. P. Bougerol, Almost sure stabilizability and Riccati's equation of linear systems with random parameters. SIAM J. Control Optim. **33**(3), 702–717 (1995)
6. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
7. R. Bartle, D. Sherbert, *Introduction to Real Analysis* (Wiley, New York, 2000)
8. W. Wonham, On pole assignment in multi-input controllable linear systems. IEEE Trans. Autom. Control **12**(6), 660–665 (1967)
9. L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S. Sastry, Foundations of control and estimation over lossy networks. Proc. IEEE **95**(1), 163–187 (2007)
10. N. Xiao, L. Xie, C. de Souza, Kalman filtering over fading channels with both transmission failure and signal fluctuation, in *Proceedings of the 9th IEEE International Conference on Control and Automation*, pp. 183–188 (2011)

Chapter 13
Kalman Filtering with Packet Losses

In this chapter, we study the Kalman filtering problem with Markovian packet losses with the focus on the stability of estimation error covariance matrices. We introduce the notions of *stability in stopping times* and *stability in sampling times*. The first one deals with stability of a randomly down-sampled system. It is shown that both the aforementioned stability notions are equivalent. This result makes the stability analysis of the estimation error covariance matrices relatively easier because stability in stopping times is generally easier to study. All stability criteria in this chapter are described by simple strict inequalities in terms of the largest eigenvalue of the open loop matrix and transition probabilities of the Markov process.

The chapter is organized as follows. The problem is formulated in Sect. 13.1, where two stability notions are introduced. In Sect. 13.2, a necessary condition for both stability notions of vector systems is derived, from which the equivalence between the two stability notions is established. Necessary and sufficient conditions for the stability of the mean estimation error covariance matrices of second-order systems are provided in Sect. 13.3. The necessary condition presented in Sect. 13.2 is proved to be sufficient for certain classes of higher-order systems in Sect. 13.4. Illustrative examples are presented in Sect. 13.5. Most of proofs in this chapter can be found in Sect. 13.6. Concluding remarks are drawn in Sect. 13.7.

13.1 Networked Estimation

Consider a discrete-time stochastic linear system

$$\begin{cases} x_{k+1} = Ax_k + w_k, \\ y_k = Cx_k + v_k, \end{cases} \tag{13.1}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_13

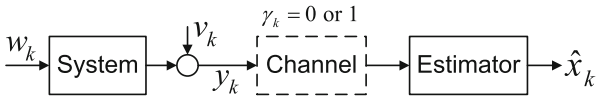

Fig. 13.1 Network configuration

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, i.e., $\text{rank}(C) = \ell \leq n$. The initial state x_0 is assumed to be a random Gaussian vector with mean $\hat{x}_0$ and the covariance matrix $P_0 > 0$. Moreover, w_k, v_k and x_0 are mutually independent.

We focus on an estimation framework where the raw measurements of the system are transmitted to an estimator via an unreliable communication channel, see Fig. 13.1. Due to random fading and/or congestion of the communication channel, packets may be lost while in transit through the channel. Here we do not address other effects such as quantization, transmission errors and delays. The packet loss process is modeled by a time-homogenous binary Markov process $\{\gamma_k\}_{k\geq 0}$, which is more general and realistic than the i.i.d. case due to possible temporal correlation of network conditions. Furthermore, assume that $\{\gamma_k\}_{k\geq 0}$ does not contain any information of the system, and is independent of the system evolution. Let $\gamma_k = 1$ indicate that the packet containing the information of y_k has been successfully delivered to the estimator while $\gamma_k = 0$ corresponds to the loss of the packet. In addition, the Markov process has a transition probability matrix given by

$$\Pi^+ = (\mathbb{P}\{\gamma_{k+1} = j | \gamma_k = i\})_{i,j\in\mathbb{S}} = \begin{bmatrix} 1-q & q \\ p & 1-p \end{bmatrix}, \tag{13.2}$$

where $\mathbb{S} \triangleq \{0, 1\}$ is the state space of the Markov process. To avoid any trivial case, the *failure rate* p and *recovery rate* q are assumed to be strictly positive and less than 1, i.e., $0 < p, q < 1$. This implies that the Markov process $\{\gamma_k\}_{k\geq 0}$ is ergodic. Obviously, a smaller value of p and a larger value of q indicate a more reliable communication link.

Denote $(\Omega, \mathscr{F}, \mathbb{P})$ the common probability space for all random variables in the chapter, where Ω is the space of elementary events, $\mathscr{F}$ is the underlying σ-field on Ω, and $\mathbb{P}$ is a probability measure on $\mathscr{F}$.

Let

$$\mathscr{F}_k \triangleq \sigma(y_i\gamma_i, \gamma_i, i \leq k) \subset \mathscr{F}$$

be an increasing sequence of σ-fields generated by the information received by the estimator up to time k, i.e., all events that are generated by the random variables $\{y_i\gamma_i, \gamma_i, i \leq k\}$. In the sequel, the terminology of almost everywhere (abbreviated as *a.e.*) is always with respect to (w.r.t.) the probability measure $\mathbb{P}$. Similarly to Chap. 5, we define a sequence of stopping times $\{t_k\}_{k\geq 0}$ adapted to the Markov process $\{\gamma_k\}_{k\geq 0}$ as follows:

$$
\begin{aligned}
t_0 &= 0,\\
t_1 &= \inf\{k|k \geq 1, \gamma_k = 1\},\\
t_2 &= \inf\{k|k > t_1, \gamma_k = 1\},\\
&\vdots \quad \vdots\\
t_j &= \inf\{k|k > t_{j-1}, \gamma_k = 1\}.
\end{aligned} \tag{13.3}
$$

By the ergodic property of the Markov process $\{\gamma_k\}_{k\geq 0}$, t_k is finite *a.e.* for any fixed k [1]. Thus, the integer valued sojourn time $\tau_k,\ k > 0$ that denotes the time duration between two successive packet received times is well-defined *a.e.*, where

$$\tau_k \triangleq t_k - t_{k-1} > 0. \tag{13.4}$$

With regard to the probability distribution of sojourn times $\{\tau_k\}_{k>0}$, we recall the following interesting result.

By Lemma 5.1, it follows that conditioned on $\{\gamma_0 = 1\}$, the sojourn times $\{\tau_k\}_{k>0}$ are i.i.d., and the conditional distribution of τ_1 is explicitly expressed as

$$\mathbb{P}\{\tau_1 = i|\gamma_0 = 1\} = \begin{cases} 1-p, & i = 1;\\ pq(1-q)^{i-2}, & i > 1. \end{cases} \tag{13.5}$$

13.1.1 Intermittent Kalman Filter

To this purpose, denote the state estimate and one-step prediction corresponding to the minimum mean square error estimator by $\hat{x}_{k|k} = \mathbb{E}[x_k|\mathscr{F}_k]$ and $\hat{x}_{k+1|k} = \mathbb{E}[x_{k+1}|\mathscr{F}_k]$, respectively. The associated estimation error covariance matrices are defined by

$$P_{k|k} = \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^H|\mathscr{F}_k]$$

and

$$P_{k+1|k} = \mathbb{E}[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^H|\mathscr{F}_k],$$

where A^H is the conjugate transpose of A. By [2], it is known that the Kalman filter is still optimal. That is, the following recursions are in force:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k(y_k - C\hat{x}_{k|k-1}); \tag{13.6}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C P_{k|k-1}, \tag{13.7}$$

where $K_k = P_{k|k-1}C^H(CP_{k|k-1}C^H + R)^{-1}$. In addition, the time update equations continue to hold: $\hat{x}_{k+1|k} = A\hat{x}_{k|k}$, $P_{k+1|k} = AP_{k|k}A^H + Q$ and $\hat{x}_{0|-1} = \bar{x}_0$, $P_{0|-1} = P_0$. For simplicity of exposition, let $P_{k+1} = P_{k+1|k}$ and $M_k = P_{t_k+1}$.

13.1.2 Stability Notions

To analyze the behavior of the estimation error covariance matrices, we introduce two types of stability notions.

Definition 13.1 We say that the mean state estimation error covariance matrices are stable in sampling times if

$$\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$$

while they[1] are stable in stopping times if

$$\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$$

for any $P_0 > 0$, where the expectation is taken w.r.t. packet loss process $\{\gamma_k\}_{k\geq 0}$ with γ_0 being any Bernoulli random variable.

Here $\mathbb{E}[P_k]$ represents the mean of one-step prediction error covariance at the sampling time whereas $\mathbb{E}[M_k]$ denotes the mean of one-step prediction error covariance at the stopping time. To some extent, the former is time-driven while the latter is event-driven. Although the two stability notions have different meanings, they will be shown to be equivalent in Sect. 13.2. Our objective of this chapter is to establish the equivalence between the two stability notions and derive necessary and sufficient conditions for stability. For scalar systems, the stability in sampling times has been discussed in [3] by analyzing a random Riccati recursion. Their approach is quite conservative for vector systems as they leave the system structure unexplored. In this chapter, a completely different method is developed to establish the main results.

Assumption 13.1 For simplicity of presentation, we make the following assumptions.

(a) P_0, Q, R are all identity matrices with compatible dimensions.
(b) All the eigenvalues of A lie outside the unit circle.
(c) (C, A) is observable.

13.2 Equivalence of the Two Stability Notions

Due to the temporal correlations of the packet loss process, it is generically difficult to directly study the notion of stability in sampling times [3]. However, the two stability notions will be shown to be equivalent in this section. Then, it is sufficient to study the stability in stopping times, which is relatively easier as will be demonstrated.

[1] This notation means that there is a positive definite $\bar{P}$ such that $\mathbb{E}[P_k] < \bar{P}$ for all $k \in \mathbb{N}$. Similar meaning applies to the notation $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$.

For any $i \in \mathbb{S}$, denote $\mathbb{E}^i[\cdot]$ the mathematical expectation operator conditioned on the event that $\{\gamma_0 = i\}$.

Lemma 13.1 *The following statements hold:*

(a) $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$ *if and only if* $\sup_{k\in\mathbb{N}} \mathbb{E}^1[P_k] < \infty$ *and* $\sup_{k\in\mathbb{N}} \mathbb{E}^0[P_k] < \infty$.
(b) $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *if and only if* $\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty$ *and* $\sup_{k\in\mathbb{N}} \mathbb{E}^0[M_k] < \infty$.

Proof (a) "⇐:" It is obvious since

$$\mathbb{E}[P_k] \le \mathbb{E}^1[P_k] + \mathbb{E}^0[P_k].$$

"⇒:" Let $\mathbb{P}\{\gamma_0 = 1\} = \mathbb{P}\{\gamma_0 = 0\} = 1/2$. Note that $P_k \ge 0$, then $\mathbb{E}[P_k] \ge \mathbb{E}^1[P_k]/2$ and $\mathbb{E}[P_k] \ge \mathbb{E}^0[P_k]/2$.

(b) Similar to (a).

Theorem 13.1 *Consider the system* (13.1) *satisfying Assumption* 13.1 *and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix* (13.2)*. Then, a necessary condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$\rho(A)^2(1-q) < 1.$$

Proof Define a linear operator $g(\cdot)$ by

$$g(P) = APA^H + Q$$

and the composite function $g \circ g(\cdot)$ by

$$g \circ g(P) = g(g(P)) = g^2(P).$$

Similar definition applies to the notation $g^k(\cdot)$ for all $k \ge 1$. Since t_k is a stopping time, $\mathscr{F}_{t_k} \triangleq \sigma(y_i\gamma_i, \gamma_i, i \le t_k)$ is a well defined σ-field. Noting that $P_k \ge Q = I$ for all $k \in \mathbb{N}$ and $M_k = P_{t_k+1}$, it immediately follows from the property of conditional expectation that

$$\begin{aligned}\mathbb{E}[M_{k+1}] &= \mathbb{E}\left[\mathbb{E}[M_{k+1}|\mathscr{F}_{t_k}]\right] \\ &= \mathbb{E}\left[g^{\tau_{k+1}}(M_k)\right] \ge \mathbb{E}[g^{\tau_{k+1}}(Q)] \\ &= \mathbb{E}[\sum_{j=0}^{\tau_{k+1}} A^j(A^j)^H], \end{aligned} \tag{13.8}$$

where the first inequality is due to that $g(\cdot)$ is a monotonically increasing function.

Let

$$J = \text{diag}(J_1, \ldots, J_d) \in \mathbb{C}^{n\times n}$$

be the Jordan canonical form of A, where $J_i \in \mathbb{C}^{n_i\times n_i}$ is the Jordan block corresponding to the eigenvalue λ_i. That is, there exists a nonsingular matrix $U \in \mathbb{R}^{n\times n}$ such that $A = UJU^{-1}$. Then, it follows that

$$\begin{aligned}\sum_{j=0}^{\tau_{k+1}} A^j (A^j)^H &= U \sum_{j=0}^{\tau_{k+1}} J^j U^{-1} U^{-H} (J^j)^H U^H \\ &\geq \lambda_{\min}(U^{-1}U^{-H}) U \sum_{j=0}^{\tau_{k+1}} J^j (J^j)^H U^H, \end{aligned} \tag{13.9}$$

where $\lambda_{\min}(U^{-1}U^{-H}) > 0$ is the smallest eigenvalue of $U^{-1}U^{-H}$. In view of (13.8), (13.9) and Lemma 13.1, it is clear that

$$\sup_{k\in\mathbb{N}} \mathbb{E}[M_{k+1}] < \infty$$

implies that

$$\sup_{k\in\mathbb{N}} \mathbb{E}^1[\sum_{j=0}^{\tau_{k+1}} J^j (J^j)^H] < \infty. \tag{13.10}$$

Note that the (n_i, n_i)th element of $\mathbb{E}^1[\sum_{j=0}^{\tau_{k+1}} J_i^j (J_i^j)^H]$ is computed by

$$\mathbb{E}^1[\sum_{j=0}^{\tau_{k+1}} |\lambda_i|^{2j}] = \frac{|\lambda_i^2|\mathbb{E}^1[|\lambda_i|^{2\tau_1}] - 1}{|\lambda_i|^2 - 1}.$$

By (13.10) and the equivalence property of norms on a finite-dimensional vector space, it follows that

$$\frac{|\lambda_i^2|\mathbb{E}^1[|\lambda_i|^{2\tau_1}] - 1}{|\lambda_i|^2 - 1} < \infty.$$

Together with Lemma 5.1, we have that $|\lambda_i|^2(1-q) < 1$. Since λ_i is an arbitrary eigenvalue of A, this completes the proof.

Theorem 13.2 *Consider the system* (13.1) *satisfying Assumption* 13.1 *and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix* (13.2)*. Then, a necessary condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$ *is that* $\rho(A)^2(1-q) < 1$.

Proof Since $P_k - P_k C^H (CP_k C^H + R)^{-1} CP_k \geq 0$, see [4], we obtain that for any $k > 3$,

$$P_{k+1} \geq (1-\gamma_k) A P_k A^H + Q \geq \sum_{j=1}^{k} (\prod_{i=j}^{k} (1-\gamma_i)) A^{k-j} (A^{k-j})^H, \quad (13.11)$$

where the second inequality is due to that $Q = I$ by A1.

Denote $\pi_j^i = \mathbb{P}\{\gamma_j = i\}, i \in \{0, 1\}$ and $\pi_j = [\pi_j^0, \pi_j^1]$. By (13.2), we have that $\pi_{j+1} = \pi_j \Pi^+$ for any $j \in \mathbb{N}$. Together with $0 < p, q < 1$, one can test that for any finite $j > 1$, $\pi_j^i > 0$ for all $i \in \{0, 1\}$. In addition, the Markov process $\{\gamma_k\}_{k\in\mathbb{N}}$ has a unique stationary distribution $[\pi^0, \pi^1]$, i.e.,

$$\lim_{j\to\infty} \pi_j^i = \pi^i, i \in \mathbb{S}.$$

By (13.2), we further obtain that $\pi^0 = \frac{p}{p+q} > 0$. It follows that

$$\underline{\pi}^0 \triangleq \inf_{j\geq 1} \pi_j^0 > 0,$$

which further implies that for all $j \geq 2$,

$$\begin{aligned} \mathbb{E}[\prod_{i=j}^{k} (1-\gamma_i)] &\geq \mathbb{E}[\prod_{i=j}^{k} (1-\gamma_i) | \gamma_{j-1} = 0] \mathbb{P}(\gamma_{j-1} = 0) \\ &\geq \underline{\pi}^0 (1-q)^{k-j}. \end{aligned} \quad (13.12)$$

In view of (13.11), we obtain that

$$\mathbb{E}[P_{k+1}] \geq \underline{\pi}^0 \sum_{j=0}^{k-2} (1-q)^j A^j (A^j)^H.$$

By following a similar line of the proof in Theorem 13.1, it immediately yields that

$$\rho(A)^2 (1-q) < 1.$$

Remark 13.1 Let $\bar{q} = \max\{q, 1-p\}$, [4] provides a necessary condition, i.e.,

$$\rho^2(A)(1-\bar{q}) < 1$$

for $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$, which is obviously weaker than Theorem 13.2 if $p + q < 1$.

By the above results, the equivalence between the two stability notions is established in the following result, whose proof is given in Sect. 13.6.

Theorem 13.3 *Consider the system* (13.1) *satisfying Assumption* 13.1 *and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix* (13.2)*. Then, the notions of stability in stopping times and stability in sampling times are equivalent.*

Thus, there is no loss of generality for the rest of the chapter to focus on the stability in stopping times.

13.3 Second-Order Systems

Consider second-order systems with the following structure:

Assumption 13.2 $A = \mathrm{diag}(\lambda_1, \lambda_2)$ and $\mathrm{rank}(C) = 1$, where $\lambda_2 = \lambda_1 \exp(\frac{2\pi r}{d}\mathbb{i})$, $\mathbb{i}^2 = -1$, $d > r \geq 1$ and $r, d \in \mathbb{N}$ are irreducible.

Under A4, it is easy to verify that (C, A^d) is not an observable pair. This essentially indicates that the measurements received at times kd for all $k \in \mathbb{N}$ do not help to reduce the estimation error, which will become clear shortly. Thus, it is intuitive that with a smaller d, it may require a stronger condition to ensure stability of the mean estimation error covariance matrices as observability may be lost relatively easily, which is confirmed in Theorem 13.4.

Theorem 13.4 *Consider the second-order system* (13.1) *satisfying Assumption* 13.1 *and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix* (13.2)*. Then,*

(a) *if* (C, A) *satisfies A4, a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$(1 + \frac{pq}{(1-q)^2})(\rho(A)^2(1-q))^d < 1;$$

(b) *otherwise, a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$\rho(A)^2(1-q) < 1.$$

The proof is delivered in Sect. 13.6. By Theorem 13.3, the results in Theorem 13.4 apply to the notion of stability in sampling times as well. Some remarks are included below.

Remark 13.2 Since $d \geq 2$, the function

$$(1 + \frac{pq}{(1-q)^2})(1-q)^d$$

is decreasing w.r.t. $q \in (0, 1)$ but increasing w.r.t. $p \in (0, 1)$. For a communication link with a smaller p and a larger q, which corresponds to a more reliable network, a

more unstable system can be tolerated without losing stability of the estimation error covariance matrices. This is consistent with our intuition.

Remark 13.3 If the conjugate complex eigenvalues satisfy that $\lambda_2 = \lambda_1 \exp(2\pi\varphi\mathbb{i})$, where φ is an irrational number, Assumption 13.2 does not hold. A necessary and sufficient condition for both the types of stability is that

$$|\lambda_1|^2(1-q) < 1.$$

Under this situation, the pair (C, A^k) remains observable for all $k \geq 1$. Then, the failure rate p becomes immaterial. In Sect. 13.4, we show that even for certain classes of higher-order systems with scalar measurements, the failure rate is of little importance for stability as well.

Remark 13.4 In [3], they establish the equivalence of the usual stability (stability in sampling times) and the so-called peak covariance stability of the estimation error covariance matrices only for scalar systems. But for vector systems, they give a conservative sufficient condition for the peak covariance stability and do not consider the usual stability.

Remark 13.5 If the packet loss process is an i.i.d. process, corresponding to $q = 1-p$ in the transition probability matrix of the Markov process, the stability criterion under A4 in Theorem 13.4 is reduced to that

$$q > 1 - \rho(A)^{-\frac{2d}{d-1}},$$

which recovers the result in [5]. Note that under i.i.d. packet losses, a lower bound for the critical packet loss rate given in [2] is interpreted as

$$q > 1 - \rho(A)^{-2},$$

which is obviously not tight for systems satisfying Assumption 13.2.

13.4 Higher-Order Systems

Under an i.i.d. packet loss assumption, an explicit characterization of necessary and sufficient conditions for stability of filtering error covariance for general vector linear systems is known to be extremely challenging [2, 5, 6]. Fortunately, for certain classes of higher-order systems, where each unstable eigenvalue of A^{-1} associates with only one Jordan block and has a distinct magnitude or (C, A) is a non-degenerate pair, it is possible to give a simple necessary and sufficient condition for stability of the estimation error covariance matrices. This section shows that the condition in Theorem 13.1 is also sufficient under certain classes of higher-order systems, whose proofs are given in Sect. 13.6.

13.4.1 Non-degenerate Systems

Some definitions introduced in [5] are adopted.

Definition 13.2 The pair (C, A) is one step observable if C is of full column rank.

Definition 13.3 Assume that (C, A) is in diagonal standard form, i.e.,

$$A = \text{diag}(\lambda_1, \ldots, \lambda_n) \text{ and } C = [C_1, \ldots, C_n].$$

An equi-block of the system is defined as the subsystem corresponding to the block $(C_{\mathscr{I}}, A_{\mathscr{I}})$, where $\mathscr{I} = \{i_1, \ldots, i_l\} \subset \{1, \ldots, n\}$ is an index set such that $|\lambda_{i_1}| = \ldots = |\lambda_{i_l}|$ and $A_{\mathscr{I}} = \text{diag}(\lambda_{i_1}, \ldots, \lambda_{i_l})$, $C_{\mathscr{I}} = [C_{i_1}, \ldots, C_{i_l}]$.

Definition 13.4 The system (C, A) is non-degenerate if every equi-block of the system is one step observable. Conversely, the system (C, A) is degenerate if there exists an equi-block of the system that is not one step observable.

The concept of non-degenerate is weaker than that of one step observable system but stronger than observable one.

Assumption 13.3 (C, A) is a non-degenerate pair.

Theorem 13.5 *Consider the system* (13.1) *satisfying Assumptions* 13.1 *and* 13.3*, and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix* (13.2)*. Then, a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$\rho(A)^2(1-q) < 1.$$

It should be noted that Theorem 7 of [5] provides a necessary and sufficient condition for stability in sampling times for non-degenerate systems under i.i.d. packet losses. Their results indicate that the lower bound for the critical packet loss rate in [2] is tight for non-degenerate systems. While in Theorem 13.5, we give a necessary and sufficient condition for stability of non-degenerate systems under Markovian packet losses. Next, the necessary condition in Theorem 13.1 is proved to be sufficient for another class of higher-order systems with the following structure.

Assumption 13.4 $A^{-1} = \text{diag}(J_1, \ldots, J_r)$ and $\text{rank}(C) = 1$, 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 an identity matrix with a compatible dimension and the (j, k)th element of N_i is 1 if $k = j + 1$ and 0, otherwise.

Theorem 13.6 *Consider the system* (13.1) *satisfying Assumptions* 13.1 *and* 13.4*, and the packet loss process of the measurements governed by a time-homogeneous*

Markov process with transition probability matrix (13.2). *Then, a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$\rho(A)^2(1-q) < 1.$$

Remark 13.6 Note that except for the case that A has n eigenvalues and each of them is with a distinct magnitude, Assumptions 13.3 and 13.4 define two disjoint classes of higher-order systems.

13.5 Illustrative Examples

Example 13.1 Let a second-order system be specified by

$$A = \begin{bmatrix} 1.5 & 0 \\ 0 & -1.5 \end{bmatrix} \text{ and } C = [1 \;\; 1]. \tag{13.13}$$

In order to achieve stability, the *failure rate* p and *recovery rate* q should satisfy that

$$\left(1 + \frac{pq}{(1-q)^2}\right)(1-q)^2 < 1.5^{-4} = 0.198$$

by Theorem 13.4. Two sample paths with different recovery rates are shown in Figs. 13.2 and 13.3, which illustrate that with a smaller recovery rate, the estimation error covariance matrices have more chances to reach a high level, even diverge. Actually, it can be verified that with $q = 0.6$ and $p = 0.1$, the inequality in Theorem 13.4 is violated.

Example 13.2 The results on higher-order systems in Sect. 13.4 are applied to target tracking over a packet loss network. The dynamic of target is expressed by [7]

$$x_{k+1} = \begin{bmatrix} 1 & h & h^2 \\ 0 & 1 & h \\ 0 & 0 & 1 \end{bmatrix} x_k + w_k, \tag{13.14}$$

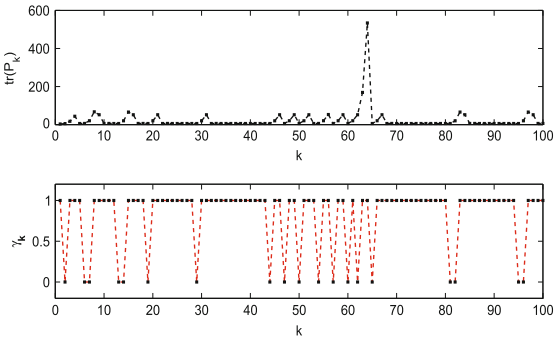

Fig. 13.2 A sample path with $q = 0.8$ and $p = 0.1$

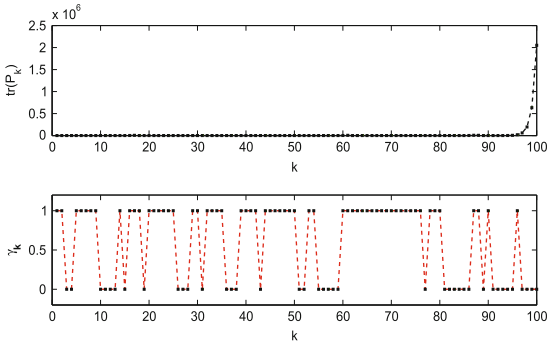

Fig. 13.3 A sample path with $q = 0.6$ and $p = 0.1$

where h is the sampling period and x_k denotes the target state at time kh, including the target position, speed and acceleration. The input random signal w_k is an additive white Gaussian noise. When the sampling period h is sufficiently small, the covariance of w_k is given by

$$Q = 2\alpha\sigma_m^2 \begin{bmatrix} h^5/20 & h^4/8 & h^3/6 \\ h^4/8 & h^3/3 & h^2/2 \\ h^3/6 & h^2/2 & h \end{bmatrix}, \tag{13.15}$$

where σ_m^2 is the variance of the target acceleration and α is the reciprocal of the maneuver time constant. The sensor periodically measures the target position with the following output equation:

$$y_k = [1,\ 0,\ 0]x_k + v_k, \tag{13.16}$$

where the measurement noise v_k is an additive white noise with variance R and independent of w_k. The initial state x_0 is a Gaussian random vector with zero mean and covariance as follows [7]:

$$P_0 = \begin{bmatrix} R & R/h & 0 \\ R/h & 2R/h^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In this example, set $h = 0.1$s, $\alpha = 0.1$, $\sigma_m^2 = 1$ and $R = 0.01$. Although here A is marginally unstable, a scaling on A can be made as in Theorem 8 of [5]. Jointly with Theorem 13.6, it follows that $q > 0$ is sufficient to guarantee the stability of the estimation error covariance matrices. Let $q = 0.2$ and $p = 0.5$, one sample path for the tracking error variance of position is shown in Fig. 13.4, which illustrates that an bounded tracking error is achieved.

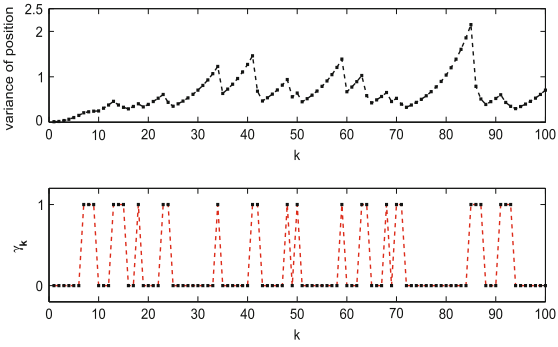

Fig. 13.4 A sample path with $q = 0.2$ and $p = 0.5$

13.6 Proofs

Since the Markov process is temporally correlated, the proof would be more challenging than the case with i.i.d. packet losses. Before proceeding further, we need some technical lemmas.

Lemma 13.2 [8] *For any $A \in \mathbb{R}^{n\times n}$ and $\varepsilon > 0$, it holds that*

$$\|A^k\| \le N\eta^k, \quad \forall k \ge 0, \tag{13.17}$$

where $N = \sqrt{n}(1 + \frac{2}{\varepsilon})^{n-1}$ and $\eta = \rho(A) + \varepsilon\|A\|$.

If A is invertible, define $\phi(k, i) = A^{t_i - t_k}$ if $k > i$ and $\phi(k, i) = I$ if $k \le i$. Let

$$\Theta_k = \sum_{i=0}^{k} \gamma_i (A^{i-k})^H C^H C A^{i-k} + (A^{-k})^H A^{-k}, \tag{13.18}$$

$$\Lambda_k = \sum_{j=0}^{k} \phi^H(k, j) C^H C \phi(k, j) + \phi^H(k, 0)\phi(k, 0), \tag{13.19}$$

$$\Xi_k = \sum_{j=0}^{k} \phi^H(j, 0) C^H C \phi(j, 0) + \phi^H(k, 0)\phi(k, 0), \tag{13.20}$$

$$\Xi = \sum_{j=0}^{\infty} \phi^H(j, 0) C^H C \phi(j, 0). \tag{13.21}$$

Lemma 13.3 *Under A1–A3, there exist strictly positive constant numbers α and β such that for any $k \in \mathbb{N}$,*

$$\alpha A \Lambda_k^{-1} A^H \le M_k \le \beta A \Lambda_k^{-1} A^H. \tag{13.22}$$

Proof By revising Lemma 2 in [5] and the fact that $\gamma_j = 0$ if $j \notin \{t_k, k \in \mathbb{N}\}$, the proof can be readily established and the details are omitted.

By (13.2), it is easy to check that Ξ is invertible *a.e.* Thus, except on a set with zero probability, the inverse of Ξ is well defined. On this exceptional set, we can set Ξ^{-1} to be any value, e.g., zero matrix, as its value on a zero probability set does not affect the expectation of $\mathbb{E}[\Xi^{-1}]$.

Lemma 13.4 *Under Assumption 13.1, there exist strictly positive constant numbers $\tilde{\alpha}$ and $\tilde{\beta}$ such that*

$$\tilde{\alpha} A \mathbb{E}^1[\Xi^{-1}] A^H \le \sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] \le \tilde{\beta} A \mathbb{E}^1[\Xi^{-1}] A^H. \tag{13.23}$$

Proof By Lemma 5.1, it is clear that conditioned on the event $\{\gamma_0 = 1\}$, the following random vectors are with an identical distribution, e.g.,

$$(\tau_k, \tau_k + \tau_{k-1}, \ldots, \tau_k + \cdots + \tau_1) \overset{d}{=} (\tau_1, \tau_1 + \tau_2, \ldots, \tau_1 + \cdots + \tau_k),$$

where $\overset{d}{=}$ means equal in distribution on its both sides. Thus, it yields that

$$\mathbb{E}^1[\Lambda_k^{-1}] = \mathbb{E}^1[\Xi_k^{-1}]$$

by (13.19) and (13.20). Jointly with Lemma 13.3, it follows that

$$\mathbb{E}^1[M_k] \le \beta A \mathbb{E}^1[\Xi_k^{-1}] A^H. \tag{13.24}$$

Under Assumption 13.1, it is possible to select a positive $\varepsilon < \frac{1-\rho(A^{-1})}{\|A^{-1}\|}$ and

$$\eta = \rho(A^{-1}) + \varepsilon \|A^{-1}\| < 1,$$

then it follows from Lemma 13.2 that for any $k \in \mathbb{N}$,

$$\begin{aligned} \sum_{j=k+1}^{\infty} \phi^H(j,k) C^H C \phi(j,k) &\le \|C\|^2 \sum_{j=k+1}^{\infty} \|A^{(t_j - t_k)}\|^2 I \\ &\le N\|C\|^2 \sum_{j=k+1}^{\infty} \eta^{2(t_k - t_j)} I \le \frac{N\|C\|^2}{1-\eta^2} I \triangleq \beta_0 I, \end{aligned} \tag{13.25}$$

where the last inequality is due to that $\tau_k \ge 1$ for all $k \in \mathbb{N}$.

Let $\beta_1 = \min(1, \beta_0^{-1})$ and $\tilde{\beta} = \beta\beta_1$, then $\Xi_k \ge \sum_{j=0}^{k} \phi^H(j,0) C^H C \phi(j,0) + \beta_0^{-1} \phi^H(k,0) (\sum_{j=k+1}^{\infty} \phi^H(j,k) C^H C \phi(j,k)) \phi(k,0) \ge \beta_1 \Xi$, where the second inequality is due to (13.25). Thus, the right hand side of the inequality of (13.23) trivially follows from (13.24). Similar to (13.24), the left hand side of (13.23) can be shown by using Fatou's Lemma [9].

13.6.1 Proof of Theorem 13.3

Proof On one hand, assume that $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$. By (13.2), the Markov process has a unique stationary distribution given as follows,

$$\mathbb{P}\{\gamma_\infty = i\} = \lim_{k\to\infty} \mathbb{P}\{\gamma_k = i\} = \frac{p^{1-i}q^i}{p+q}, \quad \forall i \in \mathbb{S}. \tag{13.26}$$

Consider a special case that the Markov process starts at its stationary distribution, i.e.,

$$\mathbb{P}\{\gamma_0 = i\} = \frac{p^{1-i}q^i}{p+q}$$

for all $i \in \mathbb{S}$. Then, the distribution of γ_k is the same as that of γ_0. Under this case, it can be verified that

$$\Pi^- = (\mathbb{P}\{\gamma_k = j|\gamma_{k+1} = i\})_{i,j\in\mathbb{S}} = \begin{bmatrix} 1-q & q \\ p & 1-p \end{bmatrix}. \tag{13.27}$$

Given a measurable function $f : \mathbb{R}^{k+1} \to \mathbb{R}^{n\times n}$, we obtain that:

$$\begin{aligned}
&\mathbb{E}[f(\gamma_k, \ldots, \gamma_0)] \\
&\quad = \sum_{i_j\in\mathbb{S}, 0\le j\le k} f(i_k, \ldots, i_0)\mathbb{P}\{\gamma_k = i_k, \ldots, \gamma_0 = i_0\} \\
&\quad = \sum_{i_j\in\mathbb{S}, 0\le j\le k} f(i_k, \ldots, i_0)\mathbb{P}\{\gamma_0 = i_0\} \prod_{j=0}^{k-1} \mathbb{P}\{\gamma_{j+1} = i_{j+1}|\gamma_j = i_j\} \quad (13.28) \\
&\quad = \sum_{i_j\in\mathbb{S}, 0\le j\le k} f(i_k, \ldots, i_0)\mathbb{P}\{\gamma_k = i_0\} \prod_{j=0}^{k-1} \mathbb{P}\{\gamma_j = i_{j+1}|\gamma_{j+1} = i_j\} \quad (13.29) \\
&\quad = \mathbb{E}[f(\gamma_0, \ldots, \gamma_k)] = \mathbb{E}[f(\gamma_1, \ldots, \gamma_{k+1})]. \quad (13.30)
\end{aligned}$$

In the above, (13.28) follows from the Markov property of $\{\gamma_k\}_{k\ge0}$ while (13.29) is due to (13.2), (13.27) and that the distribution of γ_k is the same as that of γ_0. The last equality is due to the strict stationarity of the Markov process starting from its stationary distribution. By Lemma 3 of [5], there exists a positive constant α_1 such that

$$P_{k+1} \ge \alpha_1 (\sum_{i=1}^{k+1} \gamma_{k+1-i}(A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1}.$$

Together with (13.30), we have that

$$\begin{aligned}\mathbb{E}[P_{k+1}] &\geq \alpha_1 \mathbb{E}(\sum_{i=1}^{k+1} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1} \\ &\geq \alpha_1 \mathbb{E}(\sum_{i=1}^{\infty} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1}. \end{aligned} \tag{13.31}$$

Under Assumption 13.1, the term in (13.31) is decreasing w.r.t. k. It follows from the monotone convergence theorem [9] that

$$\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] \geq \alpha_1 \mathbb{E}[(\sum_{i=1}^{\infty} \gamma_i (A^{-i})^H C^H C A^{-i})^{-1}] \geq \alpha_1 \mathbb{E}[\Xi^{-1}],$$

where the last equality follows from the definition of Ξ in (13.21). Define a stopping time μ as the time at which the first packet is received, i.e.,

$$\mu = \inf\{k \,|\gamma_k = 1, \forall k \in \mathbb{N}\}.$$

Since μ is a stopping time adapted to the Markov process $\{\gamma_k\}_{k\geq 0}$, we know that

$$\mathscr{G}_\mu \triangleq \sigma(\gamma_0, \ldots, \gamma_\mu)$$

is a well defined σ-field. Furthermore, it follows from the property of conditional expectation that

$$\begin{aligned}\mathbb{E}[\Xi^{-1}] &= \mathbb{E}[A^\mu (\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} (A^\mu)^H] \\ &= \mathbb{E}[A^\mu \mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \mathscr{G}_\mu] (A^\mu)^H].\end{aligned}$$

By (13.2), it is clear that γ_k is a strong Markov process [1]. This implies that

$$\mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \mathscr{G}_\mu] = \mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \gamma_\mu].$$

By the definition of μ, it yields that $\gamma_\mu = 1$. Again, by the strong Markov property, it follows that the transition probability matrix of $\{\gamma_{k+\mu}\}_{k\geq 0}$ is the same as that of the original Markov process $\{\gamma_k\}_{k\geq 0}$. Combining the above, we obtain that $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$ implies

$$\mathbb{E}^1[\Xi^{-1}] = \mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \gamma_\mu = 1] < \infty.$$

By Lemma 13.4, it follows that $\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty$. In view of Theorem 13.2, it implies that $\rho(A)^2(1-q) < 1$ and

$$\mathbb{E}^0[A^\mu (A^\mu)^H] = q \sum_{i=1}^{\infty} A^i (A^i)^H (1-q)^{i-1} < \infty.$$

Together with $\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty$, it can be easily established that

$$\sup_{k\in\mathbb{N}} \mathbb{E}^0[M_k] < \infty.$$

By Lemma 13.1, we finally obtain that $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$.

On the other hand, assume that $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$. By Lemmas 13.1 and 13.4, we obtain that $\mathbb{E}^1[\Xi^{-1}] < \infty$. By Theorem 13.1, it follows that $\rho(A)^2(1-q) < 1$. Then, one can easily show that $\mathbb{E}^0[\Xi^{-1}] < \infty$. As in the first part, consider the special case that the Markov process $\{\gamma_k\}_{k\geq 0}$ starts at its stationary distribution. By Lemma 3 of [5], there exists a positive constant β_2 such that

$$P_{k+1} \leq \beta_2 (\sum_{i=1}^{k+1} \gamma_{k+1-i} (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1}.$$

Together with (13.30), we have that

$$\begin{aligned} \mathbb{E}[P_{k+1}] &\leq \beta_2 \mathbb{E}(\sum_{i=1}^{k+1} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1} \\ &= \beta_2 A \mathbb{E}[\Theta_k^{-1}]\, A^H, \end{aligned} \tag{13.32}$$

where the last equality is due to the strict stationarity of the Markov process as it starts from its stationary distribution and Θ_k is defined (13.18).

Similar to (13.25), there is a positive β_3 such that

$$\sum_{j=1}^{\infty} \gamma_{k+j} (A^{-j})^H C^H C A^{-j} \leq \beta_3 I.$$

Let $\beta_4 = \min(1, \beta_3^{-1})$, we obtain that $\Theta_k \geq \beta_4 \Xi$. By (13.32), it follows that

$$\mathbb{E}[P_{k+1}] \leq \beta_2 \beta_4^{-1} A \mathbb{E}[\Xi^{-1}] A^H < \beta_2 \beta_4^{-1} A (\mathbb{E}^0[\Xi^{-1}] + \mathbb{E}^1[\Xi^{-1}]) A^H < \infty$$

for all $k \in \mathbb{N}$. Note that here $\mathbb{E}[P_k]$ is taken w.r.t. the Markov process $\{\gamma_k\}_{k\geq 0}$ with the distribution of γ_0 being the stationary distribution. Jointly with (13.26), we obtain that $\mathbb{E}^0[P_k] < \infty$ and $\mathbb{E}^1[P_k] < \infty$ for all $k \in \mathbb{N}$. By Lemma 13.1, the proof is completed.

13.6.2 Proof of Theorem 13.4

Proof Define the integer valued set $\mathscr{S}_d = \{kd | \forall k \in \mathbb{N}\}$ and

$$\theta = \sum_{j \in \mathscr{S}_d} \mathbb{P}\{\tau_1 = j | \gamma_0 = 1\}.$$

Let $E_k, k \geq 1$ be a sequence of events defined as follows:

$$E_1 = \{\tau_1 \notin \mathscr{S}_d\}, E_k \triangleq \{\tau_1 \in \mathscr{S}_d, \ldots, \tau_{k-1} \in \mathscr{S}_d, \tau_k \notin \mathscr{S}_d\},$$

for all $k \geq 2$. By Lemma 5.1, it is obvious that

$$\mathbb{P}(E_k | \gamma_0 = 1) = \theta^{k-1}(1-\theta)$$

and $E_i \bigcap E_j = \emptyset$ if $i \neq j$.

Let $F_k = \bigcup_{j=1}^{k} E_j$ and $F = \bigcup_{j=1}^{\infty} E_j$, it follows that F_k asymptotically increases to F and

$$\mathbb{P}(F | \gamma_0 = 1) = \mathbb{P}(\bigcup_{j=1}^{\infty} E_j | \gamma_0 = 1) = \sum_{j=1}^{\infty} \mathbb{P}(E_j | \gamma_0 = 1) = 1.$$

Define the indicator function $1_{F_k}(w)$ which is one if $w \in F_k$, otherwise 0. It is clear that $1_{F_k} = \sum_{j=1}^{k} 1_{E_j}$ asymptotically increases to 1_F. Since $\mathbb{P}(F | \gamma_0 = 1) = 1$, then $1_F = 1$ *a.e.* on $\{\gamma_0 = 1\}$. Together with the monotone convergence theorem [9], it follows that

$$\mathbb{E}^1[\Xi^{-1}] = \mathbb{E}^1[\Xi^{-1} 1_F] = \mathbb{E}^1[\Xi^{-1}(\lim_{k \to \infty} 1_{F_k})] = \lim_{k \to \infty} \sum_{j=1}^{k} \mathbb{E}^1[\Xi^{-1} 1_{E_j}].$$

Proof of part (a).
"$\Leftarrow$:" By (13.21), it is clear that

$$\mathbb{E}^1[\Xi^{-1} 1_{E_j}] \leq \mathbb{E}^1[(\sum_{i=j-1}^{j} \phi^H(i,0) C^H C \phi(i,0))^{-1} 1_{E_j}].$$

Define $C = [c_1, c_2]$, we can compute that

$$\sum_{i=j-1}^{j} \phi^H(i,0) C^H C \phi(i,0) \tag{13.33}$$

$$= \phi^H(j-1,0) \begin{bmatrix} c_1 \\ & c_2 \end{bmatrix} \begin{bmatrix} 1 + \lambda_1^{-2\tau_j} & 1 + \lambda_1^{-\tau_j} \lambda_2^{-\tau_j} \\ 1 + \lambda_1^{-\tau_j} \lambda_2^{-\tau_j} & 1 + \lambda_2^{-2\tau_j} \end{bmatrix} \begin{bmatrix} c_1 \\ & c_2 \end{bmatrix} \phi(j-1,0).$$

Define

$$\Sigma_j = \begin{bmatrix} 1+\lambda_1^{-2\tau_j} & 1+\lambda_1^{-\tau_j}\lambda_2^{-\tau_j} \\ 1+\lambda_1^{-\tau_j}\lambda_2^{-\tau_j} & 1+\lambda_2^{-2\tau_j} \end{bmatrix},$$

then if $\tau_j \notin \mathscr{S}_d$, it yields that

$$\Sigma_j^{-1} \leq \frac{4}{\lambda_1^{-2\tau_j} + \lambda_2^{-2\tau_j} - 2\lambda_1^{-\tau_j}\lambda_2^{-\tau_j}} I \leq \frac{2|\lambda_1|^{2\tau_j}}{1-\cos(\frac{2\pi}{d})} I.$$

Let $c = \max(c_1^{-2}, c_2^{-2})$, it follows from (13.33) that if $\tau_j \notin \mathscr{S}_d$, then

$$(\sum_{i=j-1}^{j} \phi^H(i,0) C^H C \phi(i,0))^{-1} \leq \frac{2c|\lambda_1|^{2t_j}}{1-\cos(\frac{2\pi}{d})} I.$$

Thus, we get that

$$\mathbb{E}^1[\Xi^{-1}] \leq \frac{2cI}{1-\cos(\frac{2\pi}{d})} \lim_{k\to\infty} \sum_{j=1}^{k} \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}].$$

By Lemma 5.1, the following statements are in force:

$$\begin{aligned} \lim_{k\to\infty} \sum_{j=1}^{k} \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}] &= \lim_{k\to\infty} \sum_{j=1}^{k} \mathbb{E}^1[(\prod_{i=1}^{j-1} |\lambda_1|^{2\tau_i} 1_{\{\tau_i \in \mathscr{S}_d\}}) |\lambda_1|^{2\tau_j} 1_{\{\tau_j \notin \mathscr{S}_d\}}] \\ &\leq \lim_{k\to\infty} \mathbb{E}^1[|\lambda_1|^{2\tau_1}] \sum_{j=1}^{k} (\mathbb{E}^1[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}])^{j-1}, \quad (13.34) \end{aligned}$$

which is finite if and only if $\mathbb{E}^1[|\lambda_1|^{2\tau_1}] < \infty$ and

$$\mathbb{E}^1[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}] < 1.$$

After some algebraic manipulations, it is easy to verify that

$$(1 + \frac{pq}{(1-q)^2})(|\lambda_1|^2(1-q))^d < 1$$

is equivalent to that $|\lambda_1|^2(1-q) < 1$ and

$$\frac{pq}{(1-q)^2} \frac{(|\lambda_1|^2(1-q))^d}{1-(|\lambda_1|^2(1-q))^d} < 1.$$

Together with Lemma 5.1, it implies that $\mathbb{E}^1[|\lambda_1|^{2\tau_1}] < \infty$ and

$$\mathbb{E}^1[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}] = \frac{pq}{(1-q)^2} \frac{(|\lambda_1|^2(1-q))^d}{1-(|\lambda_1|^2(1-q))^d} < 1.$$

Then, we conclude that $\mathbb{E}^1[\Xi^{-1}] < \infty$. By Lemma 13.4, it follows that

$$\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty.$$

Observe that $|\lambda_1|^2(1-q) < 1$, it is easy to show that $\sup_{k\in\mathbb{N}} \mathbb{E}^0[M_k] < \infty$. By Lemma 13.1, we obtain that $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$.

"$\Rightarrow$:" Denote $\Xi'_k = \sum_{j=0}^k \phi^H(j,0)C^H C\phi(j,0)$. In view of (13.25), it is easy to derive that

$$\begin{aligned}\Xi &= \Xi'_{j-1} + \phi^H(j,0)(C^H C + \sum_{i=j+1}^{\infty} \phi^H(i,j)C^H C\phi(i,j))\phi(j,0) \\ &\le \Xi'_{j-1} + \phi^H(j,0)(C^H C + \beta_0 I)\phi(j,0),\end{aligned}$$

where β_0 is given in (13.25).

Let

$$\beta_5^{-1} = \max\{\frac{1}{1-|\lambda_1|^{-2}}, 1, \beta_0\},$$

it follows that if $1_{E_j} = 1$, then

$$\begin{aligned}\Xi^{-1} &\ge (\Xi'_{j-1} + \phi^H(j,0)(C^H C + \beta_0 I)\phi(j,0))^{-1} \\ &= (\sum_{i=0}^{j-1} |\lambda_1|^{-2t_i} C^H C + \phi^H(j,0)(C^H C + \beta_0 I)\phi(j,0))^{-1} \\ &\ge (\frac{1}{1-|\lambda_1|^{-2}} C^H C + \phi^H(j,0)(C^H C + \beta_0 I)\phi(j,0))^{-1} \\ &\ge \beta_5(C^H C + \phi^H(j,0)(C^H C + I)\phi(j,0))^{-1}. \qquad (13.35)\end{aligned}$$

By the definition of the indicator function, it is clear that

$$\Xi^{-1} 1_{E_j} \ge \beta_5(C^H C + \phi^H(j,0)(C^H C + I)\phi(j,0))^{-1} 1_{E_j}.$$

In view of Lemma 13.4, then $\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty$ is equivalent to that $\mathbb{E}^1[\Xi^{-1}] < \infty$. This implies that

$$\lim_{k\to\infty} \sum_{j=1}^k \mathbb{E}^1[(C^H C + \phi^H(j,0)(C^H C + I)\phi(j,0))^{-1} 1_{E_j}] < \infty.$$

By some manipulations, there is a positive $\beta_6 > 0$ such that

$$\mathrm{tr}(C^H C + \phi^H(j,0)(C^H C + I)\phi(j,0))^{-1} 1_{E_j}) \geq \beta_6 |\lambda_1|^{2t_j} 1_{E_j}.$$

Thus, we obtain that

$$\lim_{k\to\infty} \sum_{j=1}^{k} \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}] = \mathbb{E}^1[|\lambda_1|^2 1_{\{\tau_1 \notin \mathscr{S}_d\}}] \lim_{k\to\infty} \sum_{j=1}^{k} (\mathbb{E}^1[|\lambda_1|^2 1_{\{\tau_1 \in \mathscr{S}_d\}}])^{j-1} < \infty.$$

Finally, as in the proof of sufficiency, one can easily derive that

$$(1 + \frac{pq}{(1-q)^2})(|\lambda_1|^2(1-q))^d < 1.$$

Proof of part (b).

"$\Leftarrow$:" Without loss of generality, only the following cases need to be discussed.

(i) If $\mathrm{rank}(C) = 2$ or A has two eigenvalues but with distinct magnitudes, this indicates that (C, A) is a non-degenerate pair. It is proved in Theorem 13.5.

(ii) If $\mathrm{rank}(C) = 1$ and A contains two identical eigenvalues, it is proved in Theorem 13.6. Note that for this case, A can not be of the form $A = \lambda_1 I$ for it leads to the pair (C, A) unobservable. Thus, A must contain exactly an elementary Jordan block.

(iii) If $\mathrm{rank}(C) = 1$ and $A = \mathrm{diag}(\lambda_1, \lambda_2)$, where $\lambda_2 = \lambda_1 \exp(2\pi\varphi\mathrm{i})$ and φ is an irrational number. Since φ is an irrational number and the set of rational numbers is dense, we can find a sequence of rational numbers

$$\{\varphi_k = \frac{r_k}{d_k}\}_{k\geq 0}$$

such that $\lim_{k\to\infty} \varphi_k = \varphi$, the integers r_k and d_k are irreducible and d_k goes into infinity as $k \to \infty$. Note that $|\lambda_1|^2(1-q) < 1$, there must exist a positive integer, denoted by d_{k_0}, such that

$$(1 + \frac{pq}{(1-q)^2})(|\lambda_1|^2(1-q))^{d_{k_0}} < 1.$$

Then, the rest of the proof follows similarly as the proof of sufficiency of part (a).

"$\Rightarrow$:" It directly follows from Theorem 13.1.

13.6.3 Proofs of Results in Sect. 13.4

Lemma 13.5 [5] *Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A and $|\lambda_1| \geq \cdots \geq |\lambda_n|$. If the pair (C, A) satisfies Assumptions 13.1 and 13.3, then the following inequality holds*

$$\limsup_{\Delta_1,\ldots,\Delta_n\to\infty} \frac{(\sum_{j=1}^{n}(A^{-k_j})^H C^H C A^{-k_j})^{-1}}{\prod_{j=1}^{n}|\lambda_j|^{2\Delta_j}} \le \beta_7 I, \tag{13.36}$$

where β_7 is a positive constant, $k_1 < k_2 < \cdots < k_n \in \mathbb{N}$, $\Delta_1 = k_1$, $\Delta_j = k_j - k_{j-1}$ for all $j \in \{2, \ldots, n\}$.

Proof of Theorem 13.5

"$\Leftarrow$:" By Lemma 13.5, there exists a sufficiently large $\Delta > 0$ such that for all $\Delta_j > \Delta$, it holds

$$(\sum_{j=1}^{n}(A^{-k_j})^H C^H C A^{-k_j})^{-1} \le \beta_7 \prod_{j=1}^{n}|\lambda_j|^{2\Delta_j} I.$$

Now, select $k_j = t_{i_j}$, where $i_1 > \Delta$, $i_j - i_{j-1} > \Delta$ for all $j \in \{2, \ldots, n\}$ and t_{i_j} is a stopping time defined in (13.4). Then, it is obvious that

$$t_{i_j} - t_{i_{j-1}} \ge i_j - i_{j-1} > \Delta,$$

which jointly with (13.21), implies that

$$\begin{aligned}\mathbb{E}^1[\Xi^{-1}] &\le \beta_7 \mathbb{E}^1[\prod_{j=1}^{n}|\lambda_j|^{2\Delta_j}]I = \beta_7 \mathbb{E}^1[\prod_{j=1}^{n}|\lambda_j|^{2(t_{i_j}-t_{i_{j-1}})}]I \\ &= \beta_7 \prod_{j=1}^{n}(\mathbb{E}^1[|\lambda_j|^{2\tau_1}])^{i_j-i_{j-1}} I < \infty,\end{aligned} \tag{13.37}$$

where the last equality is due to Lemma 5.1 and we use the fact that $|\lambda_1|^2(1-q) < 1$ in the last inequality. By Lemma 13.4, it follows that $\sup_{k\in\mathbb{N}} \mathbb{E}^1[M_k] < \infty$. Together with that $|\lambda_1|^2(1-q) < 1$, it is easy to establish that $\sup_{k\in\mathbb{N}} \mathbb{E}^0[M_k] < \infty$. The rest of proof follows from Lemma 13.1.

"$\Rightarrow$:" It is proved in Theorem 13.1. □

The proof of Theorem 13.6 is much more involved and depends on the following lemmas, which are devoted to establishing a similar result as (13.36) under Assumptions 13.1 and 13.4.

Lemma 13.6 *For any integer k_i such that $k_{i+1} > k_i$, let $\mathscr{B} \in \mathbb{R}^{n\times n}$ be a matrix with its (i,j)th element given by $\mathscr{B}_{ij} = \binom{k_i}{j-1}$. Then, the determinant of $\mathscr{B}$ is computed as*

$$\det(\mathscr{B}) = \frac{1}{\prod_{i=0}^{n-1} i!} \prod_{1\le j<i\le n}(k_i - k_j),$$

where $i!$ is the factorial of a positive integer i.

Proof It is clear that $\det(\mathscr{B})$ is an alternative, i.e., swapping the ith and jth rows is the same as changing values of k_i and k_j. Moreover, $\det(\mathscr{B})$ is an $(n-1)$th order multivariate polynomial in $k_1, \ldots, k_n$. For example, $\det(\mathscr{B})$ is an $(n-1)$th order polynomial in k_i when all $k_j, j \neq i$ are fixed. Combining those two properties, we obtain that $\det(\mathscr{B})$ contains $\prod_{1\leq j<i\leq n}(k_i - k_j)$ as a factor. Furthermore, $\prod_{1\leq j<i\leq n}(k_i - k_j)$ is the only factor of $\det(\mathscr{B})$, modulo a constant α_n, due to that $\prod_{1\leq j<i\leq n}(k_i - k_j)$ and $\det(\mathscr{B})$ are both $(n-1)$th order, from which we get the following equality:

$$\det(\mathscr{B}) = \alpha_n \prod_{1\leq j<i\leq n}(k_i - k_j). \tag{13.38}$$

It remains to show that $\alpha_n = 1/\prod_{i=0}^{n-1} i!$. We do so by mathematical induction. For $n = 1$, $\det \mathscr{B}(k_1) = 1$. The factor $\prod_{1\leq j<i\leq n}(k_i - k_j)$ is void and $1/\prod_{i=0}^{n-1} i! = 1$. Thus, $\alpha_1 = 1$, which is correct.

Given $n = t$, suppose it holds that $\alpha_t = 1/\prod_{i=0}^{t-1} i!$. For $n = t+1$, let $k_1, \ldots, k_t$ be fixed and k_{t+1} go to infinity.

Note that

$$\lim_{k_{t+1}\to\infty} \frac{\binom{k_{t+1}}{i}}{k_{t+1}^i} = \frac{1}{i!},$$

we have that

$$\begin{aligned}
\lim_{k_{t+1}\to\infty} \frac{\det(\mathscr{B}(k_1, \ldots, k_{t+1}))}{k_{t+1}^t} &= \lim_{k_{t+1}\to\infty} \left(\frac{\binom{k_{t+1}}{t} \det(\mathscr{B})}{k_{t+1}^t} + \sum_{i=0}^{t-1} \frac{O(k_{t+1}^i)}{k_{t+1}^t} \right) \\
&= \det(\mathscr{B}(k_1, \ldots, k_t))/t! \\
&= \frac{1}{\prod_{i=0}^{t} i!} \prod_{1\leq j<i\leq t} (k_i - k_j),
\end{aligned} \tag{13.39}$$

where $O(k_{t+1}^i)$ in the first equality means that

$$\lim_{k_{t+1}\to\infty} \frac{O(k_{t+1}^i)}{k_{t+1}^i} < \infty.$$

In light of (13.38), it yields that

$$\lim_{k_{t+1}\to\infty} \frac{\det(\mathscr{B}(k_1, \ldots, k_{t+1}))}{k_{t+1}^t} = \alpha_{t+1} \prod_{1\leq j<i\leq t} (k_i - k_j).$$

Combining the above, we immediately obtain that $\alpha_{t+1} = 1/\prod_{i=0}^{t} i!$. Hence, $\alpha_n = 1/\prod_{i=0}^{n-1} i!$ holds for all $n \geq 1$. □

Lemma 13.7 *For any integer k_i such that $k_{i+1} > k_i$, let $\mathscr{B}' \in \mathbb{R}^{n\times n}$ be a matrix such that the (i,j)th element is given by $\mathscr{B}'_{ij} = \binom{k_i}{j}$. Then, the determinant of $\mathscr{B}'$ is computed as*

$$\det(\mathscr{B}') = (\prod_{i=1}^{n} \frac{k_i}{i!}) \prod_{1\le j<i\le n} (k_i - k_j).$$

Proof Similar to the proof of Lemma 13.6, $\det(\mathscr{B}')$ is an alternative and nth order multivariate polynomial. It is straightforward that $\det(\mathscr{B}')$ contains $\prod_{i=1}^{n} k_i$ as a factor. Thus, we further obtain that $\det(\mathscr{B}')$ contains

$$\prod_{i=1}^{n} k_i \prod_{1\le j<i\le n} (k_i - k_j)$$

as a factor, which is also the only factor containing k_i. Hence, the following is in force:

$$\det(\mathscr{B}') = \alpha'_n (\prod_{i=1}^{n} k_i \prod_{1\le j<i\le n} (k_i - k_j)).$$

Using similar induction arguments as in Lemma 13.6, one can easily show that $\alpha'_n = 1/\prod_{i=1}^{n} i!$. □

Lemma 13.8 *Given any integer k_i such that $k_{i+1} > k_i > n$, let $\Delta_1 = k_1, \Delta_i = k_i - k_{i-1}$ if $i \ge 2$. Denote*

$$\mathscr{O}(\{k_i\}_1^{n-1}) = [C^H, (A^{-k_1})^H C^H, \ldots, (A^{-k_{n-1}})^H C^H]^H$$

and

$$D_\lambda(\{k_i\}_1^{n-1}) = \prod_{v=1}^{r} \lambda_v^{-k(v)+\frac{n_v(n_v-1)}{2}},$$

where

$$k(1) = k_1 + \cdots + k_{n_1-1}$$

and

$$k(v) = k_{n_1+\cdots+n_{v-1}} + \cdots + k_{n_1+\cdots+n_v-1}$$

if $v \ge 2$. Under A2 and A6, we can asymptotically compute the determinant of $\mathscr{O}(\{k_i\}_1^{n-1})$. In particular, there exists a multivariate polynomial $\psi(\{k_i\}_1^{n-1})$ w.r.t. $\{k_i\}_1^{n-1}$ and independent of λ_i such that

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{\det \mathscr{O}(\{k_i\}_1^{n-1})}{D_\lambda(\{k_i\}_1^{n-1})\psi(\{k_i\}_1^{n-1})} = 1.$$

Proof Under A6, partition the observation matrix C in conformity with the block diagonal matrix A. Let $C_i = [c_{i1}, \ldots, c_{in_i}]$, it is easy to verify that

$$C_i N_i^k = [\underbrace{0, \ldots, 0}_{k}, c_{i1}, \ldots, c_{i(n_i-k)}]$$

for any $k \leq n_i - 1$. We further obtain that for any $k > n_i$,

$$\begin{aligned} C_i(\lambda_i^{-1} I_i + N_i)^k &= [\frac{c_{i1}}{\lambda_i^k}, \frac{c_{i2}}{\lambda_i^k} + \binom{k}{1}\frac{c_{i1}}{\lambda_i^{k-1}}, \ldots, \sum_{j=0}^{n_i-1}\binom{k}{j}\frac{c_{i(n_i-j)}}{\lambda_i^{k-j}}] \\ &\triangleq \lambda_i^{-k}\left[1, \binom{k}{1}, \ldots, \binom{k}{n_i-1}\right]\widetilde{C}_i, \end{aligned}$$

where $\widetilde{C}_i$ is defined as

$$\widetilde{C}_i = \mathrm{diag}(1, \lambda_i, \ldots, \lambda_i^{n_i-1}) \begin{bmatrix} c_{i1} & c_{i2} & \ldots & c_{in_i} \\ & c_{i1} & \ldots & c_{i(n_i-1)} \\ & & \ddots & \\ & & & c_{i1} \end{bmatrix}$$

and $\det(\widetilde{C}_i) = \lambda_i^{n_i(n_i-1)/2} c_{i_1}^{n_i}$. By using the above property, it follows that

$$\begin{aligned} \det \mathscr{O}(\{k_i\}_1^{n-1}) &= \det \begin{bmatrix} C_1 & \ldots & C_r \\ C_1(\lambda_1^{-1} I_1 + N_1)^{k_1} & \ldots & C_r(\lambda_r^{-1} I_r + N_r)^{k_1} \\ \ddots & & \ddots \\ C_1(\lambda_1^{-1} I_1 + N_1)^{k_{n-1}} & \ldots & C_r(\lambda_r^{-1} I_r + N_r)^{k_{n-1}} \end{bmatrix} \\ &= \det \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \binom{k_1}{0}\lambda_1^{-k_1} & \binom{k_1}{1}\lambda_1^{-k_1} & \ldots & \binom{k_1}{n_1-1}\lambda_1^{-k_1} \\ & & \ddots & \\ \binom{k_{n-1}}{0}\lambda_1^{-k_{n-1}} & \binom{k_{n-1}}{1}\lambda_1^{-k_{n-1}} & \ldots & \binom{k_{n-1}}{n_1-1}\lambda_1^{-k_{n-1}} \end{bmatrix} \\ &\quad \begin{bmatrix} \ldots & 1 & 0 & \ldots & 0 \\ \ldots & \binom{k_1}{0}\lambda_r^{-k_1} & \binom{k_1}{1}\lambda_r^{-k_1} & \ldots & \binom{k_1}{n_m-1}\lambda_r^{-k_1} \\ \ddots & & & \ddots & \\ \ldots & \binom{k_{n-1}}{0}\lambda_r^{-k_{n-1}} & \binom{k_{n-1}}{1}\lambda_r^{-k_{n-1}} & \ldots & \binom{k_{n-1}}{n_r-1}\lambda_r^{-k_{n-1}} \end{bmatrix} \\ &\quad \times \det(\mathrm{diag}(\widetilde{C}_1, \ldots, \widetilde{C}_r)) \\ &\triangleq (D_1 + \cdots + D_r)\det(\mathrm{diag}(\widetilde{C}_1, \ldots, \widetilde{C}_r)), \end{aligned} \tag{13.40}$$

where D_i is the determinant of the minor of the first matrix in the previous equation, obtained by eliminating the first row and the first column in the ith block. For example, the first block consists of the first n_1 columns and the followed n_2 columns forms the second block.

Let $\sigma = [\sigma_1, \ldots, \sigma_r]$ be a permutation of $\{1, \ldots, n-1\}$ such that $\#\sigma_1 = n_1 - 1$, $\#\sigma_2 = n_2, \ldots, \#\sigma_r = n_r$, where $\#\sigma_i$ denotes the order of the permutation σ_i.

Then, it follows from the Leibnitz formula [10] for the determinant of a matrix that

$$D_1 = \sum_{\sigma} sgn(\sigma) h(k_{\sigma_j}) \Big(\prod_{j=1}^{r} \lambda_j^{-k_{\sigma_j}}\Big), \tag{13.41}$$

where the signature of permutation σ is denoted as $sgn(\sigma)$, which is $+1$ for even permutation and -1 for odd permutations, $\lambda_j^{-k_{\sigma_j}} = \lambda_j^{-\sum_{i\in\sigma_j} k_i}$ and $h(k_{\sigma_i})$ is a polynomial function of k_i for all $i \in \sigma_j$. The summation is taken w.r.t. all permutations with order $n-1$.

Due to that $|\lambda_1| > \cdots > |\lambda_m|$, it is clear that $\lambda^{\sigma} \triangleq |\prod_{j=1}^{r} \lambda_j^{-k_{\sigma_j}}|$ achieves the maximum when $\lambda_1^{-k_{\sigma_1}} = \lambda_1^{-\sum_{i\in\{1,\ldots,n_1-1\}} k_i}$, $\lambda_2^{-k_{\sigma_2}} = \lambda_2^{-\sum_{i\in\{n_1,\ldots,n_1+n_2-1\}} k_i}$, $\ldots$, $\lambda_r^{-k_{\sigma_r}} = \lambda_r^{-\sum_{i\in\{n_1+\cdots+n_{r-1},\ldots,n-1\}} k_i}$. Thus, denote the set of permutations having the above property by $\mathscr{P}_{\sigma}^{*}$. Given any permutation σ which does not belong to $\mathscr{P}_{\sigma}^{*}$, we always have

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{\lambda^{\sigma}}{\lambda^{\sigma^*}} = 0$$

for all $\sigma^* \in \mathscr{P}_{\sigma}^{*}$ and $\sigma \notin \mathscr{P}_{\sigma}^{*}$. Consequently,

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{D_1}{D_{1\infty}} = 1,$$

where $D_{1\infty} = \prod_{j=1}^{r} D_{1j}$ and

$$\begin{aligned} D_{11} &= \det \begin{bmatrix} \binom{k_1}{1}\lambda_1^{-k_1} & \cdots & \binom{k_1}{n_1-1}\lambda_1^{-k_1} \\ & \ddots & \\ \binom{k_{n_1-1}}{1}\lambda_1^{-k_{n_1-1}} & \cdots & \binom{k_{n_1-1}}{n_1-1}\lambda_1^{-k_{n_1-1}} \end{bmatrix} \\ &= \lambda_1^{-(k_1+\ldots+k_{n_1-1})} \det \begin{bmatrix} \binom{k_1}{1} & \cdots & \binom{k_1}{n_1-1} \\ & \ddots & \\ \binom{k_{n_1-1}}{1} & \cdots & \binom{k_{n_1-1}}{n_1-1} \end{bmatrix} \\ &= \Big(\prod_{i=1}^{n_1-1} \frac{k_i}{i!} \prod_{1\le j<i\le n_1-1} (k_i - k_j)\Big) \lambda_1^{-k(1)}, \end{aligned}$$

where the last equality follows from Lemma 13.7. Using Lemma 13.6, we can similarly derive that

$$D_{12} = \det \begin{bmatrix} \binom{k_{n_1}}{0}\lambda_2^{-k_{n_1}} & \cdots & \binom{k_{n_1}}{n_2-1}\lambda_2^{-k_{n_1}} \\ & \ddots & \\ \binom{k_{n_1+n_2-1}}{0}\lambda_2^{-k_{n_1+n_2-1}} & \cdots & \binom{k_{n_1+n_2-1}}{n_2-1}\lambda_2^{-k_{n_1+n_2-1}} \end{bmatrix}$$

$$= \lambda_2^{-(k_{n_1}+\cdots+k_{n_1+n_2-1})} \det \begin{bmatrix} \binom{k_{n_1}}{0} & \cdots & \binom{k_{n_1}}{n_2-1} \\ & \ddots & \\ \binom{k_{n_1+n_2-1}}{0} & \cdots & \binom{k_{n_1+n_2-1}}{n_2-1} \end{bmatrix}$$

$$= \Big(\prod_{i=0}^{n_2-1} \frac{1}{i!} \prod_{0\le j<i\le n_2-1} (k_{n_1+i} - k_{n_1+j})\Big)\lambda_2^{-k(2)}.$$

Applying the same arguments to the rest of D_{1j}, we obtain that

$$D_{1\infty} = \prod_{i=1}^{n_1-1} \frac{k_i}{i!} \left(\prod_{v=2}^{r} \prod_{i=0}^{n_v-1} \frac{1}{v!} \prod_{0\le j<i\le n_v-1} (k_{n_1+\cdots+n_{v-1}+i} - k_{n_1+\cdots+n_{v-1}+j}) \right) \prod_{v=1}^{r} \lambda_v^{-k(v)}$$

$$\triangleq \psi_1(\{k_i\}_1^{n-1}) D_\lambda^1(\{k_i\}_1^{n-1}), \tag{13.42}$$

where $\psi_1(\{k_i\}_1^{n-1})$ is a multivariate polynomial in k_i and $D_\lambda^1(\{k_i\}_1^{n-1})$ contains k_i as its exponential component. Continuing with the same fashion, one can show that

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{D_2}{D_{2\infty}} = 1$$

with $D_{2\infty} = \psi_2(\{k_i\}_1^{n-1}) D_\lambda^2(\{k_i\}_1^{n-1})$, where $\psi_2(\{k_i\}_1^{n-1})$ is a multivariate polynomial in k_i and

$$D_\lambda^2(\{k_i\}_1^{n-1}) = \frac{\prod_{v=3}^{r} \lambda_v^{-k(v)}}{\lambda_1^{k_1+\cdots+k_{n_1}} \lambda_2^{k_{n_1+1}+\cdots+k_{n_1+n_2-1}}}.$$

Then, it follows that

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{D_{2\infty}}{D_{1\infty}} = \left(\frac{\lambda_2}{\lambda_1}\right)^{k_{n_1}} \frac{\psi_2(\{k_i\}_1^{n-1})}{\psi_1(\{k_i\}_1^{n-1})} = 0.$$

due to that $\psi_i(\{k_i\}_1^{n-1})$ is a multivariate polynomial in k_i and $|\lambda_1| > |\lambda_2|$.

Similar conclusion is reached for $D_3, \ldots, D_r$, e.g.,

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{D_{v\infty}}{D_{1\infty}} = 0, \quad \forall v \geq 3.$$

To sum up, we finally get that

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{D_1 + \cdots + D_r}{D_{1\infty}} = 1.$$

Let $\psi(\{k_i\}_1^{n-1}) = (\prod_{i=1}^r c_{i1}^{n_i})\psi_1(\{k_i\}_1^{n-1})$ and

$$D_\lambda(\{k_i\}_1^{n-1}) = (\prod_{i=1}^m \lambda_i^{n_i(n_i-1)/2}) D_\lambda^1(\{k_i\}_1^{n-1}),$$

it follows from (13.40) and (13.42) that

$$\lim_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{\det \mathscr{O}(\{k_i\}_1^{n-1})}{\psi(\{k_i\}_1^{n-1}) D_\lambda(\{k_i\}_1^{n-1})} = 1. \tag{13.43}$$

Proof of Theorem 13.6
"⇐:" In order to simplify notation, we remove the dependence of $\{k_i\}_1^{n-1}$ for quantities in Lemma 13.8, i.e., rewrite $\mathscr{O}(\{k_i\}_1^{n-1})$ as $\mathscr{O}$. Then, it yields that

$$(\mathscr{O}^H\mathscr{O})^{-1} \leq \mathrm{tr}(\mathscr{O}^H\mathscr{O})^{-1} I = \sum_{i,j} \left(\frac{[\mathrm{adj}(\mathscr{O})]_{ij}}{\det(\mathscr{O})} \right)^2 I. \tag{13.44}$$

Here $\mathrm{adj}(\mathscr{O})$ is the adjoint matrix of $\mathscr{O}$ and $[\mathrm{adj}(\mathscr{O})]_{ij}$ is the (i,j)th element of $\mathrm{adj}(\mathscr{O})$. Following a similar line of Lemma 13.8, we can show that there exist constant numbers $\beta_{i,j} = \beta_{i,j}(\lambda_1, \ldots, \lambda_r)$ such that for sufficiently large Δ_j, we have that

$$\mathrm{adj}(\mathscr{O})_{ij} \leq \beta_{i,j} |\psi| \prod_{v=1}^r |\lambda_v|^{-k'(v)},$$

where $k'(v) = k_1 + \cdots + k_{n_1-2}$ and

$$k'(v) = k_{n_1+\cdots+n_{v-1}-1} + \cdots + k_{n_1+\cdots+n_v-2}, \quad v \geq 2.$$

In light of (13.44), it follows that there exist constant numbers

$$\tilde{\beta}_{ij} = \tilde{\beta}_{ij}(\lambda_1, \ldots, \lambda_r)$$

such that

$$\limsup_{\Delta_1,\ldots,\Delta_{n-1}\to\infty} \frac{(\mathscr{O}^H\mathscr{O})^{-1}}{\prod_{i=1}^{r}|\lambda_v|^{2\Delta(v)}} \le (\sum_{i,j}\tilde{\beta}_{i,j})I, \tag{13.45}$$

where $\Delta(1) = \Delta_1 + \cdots + \Delta_{n_1-1}$ and

$$\Delta(v) = \Delta_{n_1+\cdots+n_{v-1}} + \cdots + \Delta_{n_1+\cdots+n_v-1}, \quad v \ge 2.$$

The rest of the proof directly follows from that of Theorem 13.5.

"$\Rightarrow$:" It is proved in Theorem 13.1.

13.7 Summary

The emergence of NCSs has attracted considerable interest on the intermittent Kalman filter, which is an optimal MMSE estimator. It is of paramount importance to derive the network condition for its stability. In particular, it was proved in [2] that there exists a critical packet loss rate to separate the possibility and impossibility of the channel to obtain a stable estimator. This is certainly consistent with our intuition, and gives rise to a fundamental problem on the quantification of the critical rate. Since this seminal work, a fair number of research works have been devoted to this problem. The result of this chapter is one of them, and is different from most of the existing results by exploring the system structure and introducing a new stability notion.

We have examined the stability of Kalman filtering with Markovian packet losses. To analyze the random estimation error covariance matrices, two stability notions have been introduced and shown to be equivalent, which makes relatively easier to analyze the stability of the estimation error covariance matrices. For second-order systems, necessary and sufficient conditions were obtained for ensuring stability with respect to different system structures. For certain classes of higher-order systems, a necessary and sufficient condition has been derived to guarantee stability of estimation error covariance matrices. All results can recover the related results in the existing literature.

References

1. S. Meyn, R. Tweedie, J. Hibey, *Markov Chains and Stochastic Stability* (Springer, London, 1996)
2. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
3. M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses. Automatica **43**(4), 598–607 (2007)

4. L. Xie, L. Xie, Stability of a random Riccati equation with Markovian binary switching. IEEE Trans. Autom. Control **53**(7), 1759–1764 (2008)
5. Y. Mo, B. Sinopoli, Towards finding the critical value for Kalman filtering with intermittent observations (2010). http://arxiv.org/abs/1005.2442
6. K. Plarre, F. Bullo, On Kalman filtering for detectable systems with intermittent observations. IEEE Trans. Autom. Control **54**(2), 386–390 (2009)
7. R. Singer, Estimating optimal filter tracking performance for manned maneuvering targets. IEEE Trans. Aerosp. Electron. Syst. **6**(4), 473–483 (1970)
8. V. Solo, One step ahead adaptive controller with slowly time-varying parameters. Technical report, Department of ECE (John Hopkins University, Baltimore, 1991)
9. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)
10. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)

Chapter 14
Kalman Filtering with Scheduled Measurements

Sensor nodes in a WSN are usually battery driven and hence operate on an extremely frugal energy budget. Experimental studies show that communication is a major source of energy consumption in sensor nodes. Thus, it is of paramount importance to reduce the communication load in the network. One natural way is to reduce the length of transmitted message using quantization. Another effective approach for energy saving in WSNs is to minimize the number of communications for sensor nodes under a prescribed performance requirement. This chapter presents an estimation framework under scheduled measurements for linear discrete-time stochastic systems to reduce the number of measurement transmissions.

The chapter is organized as follows. The networked estimation framework under scheduled measurement is formulated in Sect. 14.1. A sequentially controllable scheduler is devised in Sect. 14.2 where the stability analysis of the associated estimator under the scheduled measurements is investigated as well. In Sect. 14.3, an uncontrollable scheduler is introduced and the mean stability condition of the error covariance matrix is derived. Concluding remarks are drawn in Sect. 14.4.

14.1 Networked Estimation

Consider a linear discrete-time stochastic system:

$$\begin{cases} x_{k+1} = Ax_k + w_k; \\ y_k = Cx_k + v_k, \end{cases} \tag{14.1}$$

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 $\Sigma_w > 0$ and $\Sigma_v = I$, respectively. C is of full rank, i.e., $\text{rank}(C) = \ell \leq n$. The initial state x_0 is 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.

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_14

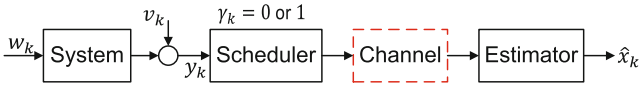

Fig. 14.1 Network configuration

Remark 14.1 If Σ_v is not an identity matrix but a general positive definite matrix, we use a transformed measurement $\widetilde{y}_k = \Sigma_v^{-1/2} y_k$ and $\widetilde{C} = \Sigma_v^{-1/2} C$ instead of y_k and C, where $\Sigma_v^{1/2}$ is a square root matrix of Σ_v and can be obtained by the Cholesky decomposition [1]. Then,

$$\widetilde{y}_k = \widetilde{C} x_k + \Sigma_v^{-1/2} v_k$$

and the covariance matrix of $\Sigma_v^{-1/2} v_k$ is an identity matrix.

We focus on an estimation framework consisting of a scheduler, a remote estimator and a wireless communication channel as shown in Fig. 14.1. Due to limited energy and communication resources, the scheduler is deployed for reducing the number of communications in the sensor node. For example, let $\gamma_k = 1$ indicate that the sensor node is triggered to communicate with the estimator at time k and the packet containing the information of y_k is delivered to the estimator while $\gamma_k = 0$ means that there is no communication between the sensor and the estimator. That is, the scheduler will decide whether y_k is to be transmitted or not. To evaluate the transmission frequency, we introduce the *scheduled transmission rate* R as follows:

$$R \triangleq \limsup_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} \gamma_j. \tag{14.2}$$

Obviously, a larger scheduled transmission rate corresponds to a higher ratio of the number of transmitted measurements to the total number of sensor measurements and a higher communication cost. In information theory [2], the transmission rate is used to measure how fast information is processed by a transmission facility. Its unit is usually expressed as bits per second. Here the scheduled transmission rate measures how frequently the sensor measurements will be transmitted.

14.1.1 Scheduling Problems

We ignore quantization, data corruption and communication delays. Under such an estimation framework, the following questions are naturally raised.

- Given a scheduling protocol, what is the corresponding scheduled transmission rate? Can we find necessary and sufficient conditions on the scheduled transmission rate for the existence of a stable estimator? Here, by stable estimator, we mean

that mean square estimation error is uniformly bounded. Furthermore, how the scheduling protocol affects the performance of the optimal estimator?
- What is the minimum scheduled transmission rate (denoted by R_m) required for the existence of a stable estimator? Namely, find R_m such that for any scheduled transmission rate $R \leq R_m$, there does not exist any scheduling protocol to obtain a stable estimator.
- Given a scheduled transmission rate, what is the optimal scheduling protocol such that the estimation error covariance is minimized?

This chapter focuses on the first two questions while the last one is left to our future work. In particular, two stochastic scheduling protocols will be provided and the stability of the estimators are analyzed.

14.2 Controllable Scheduler

Since measurement innovation represents new information of the measurement that is not contained in historical measurements, a small innovation indicates that the predicted measurement is already close to the new measurement. Motivated by this observation, we devise a scheduler as a function of the innovation process, which compares the innovation with a threshold to decide whether the associated measurement is to be transmitted to the estimator. In [3], the same philosophy has been employed where the normalized innovation outside a deadzone will be quantized and transmitted to the estimator. Ignoring the quantization effect, this idea has also been adopted in [4] for the sensor data scheduling.

In this section, we propose a sequential scheduling algorithm. Let

$$y_k^T = [y_k^1, \ldots, y_k^\ell],$$

the scheduler sequentially decides whether y_k^i is to be transmitted to the estimator. Likewise, the binary variable γ_k^i is used to indicate the transmission of y_k^i. While the scheduler in [4] checks whether the vector measurement y_k is to be sent to the estimator or not, our sequential scheduling algorithm takes the different importance of each element of y_k into consideration.

14.2.1 An Approximate MMSE Estimator

Suppose that at each transmission, only a scalar measurement is to be communicated to the estimator and multiple transmissions are permitted within one sampling interval. To be specific, the sampling interval is equally divided into m (the dimension of y_k) time slots. In each slot, the sensor node decides the transmission of y_k^i to the estimator.

Algorithm 14.1: Controllable Scheduler

1: Prior estimate and error covariance matrix:
$\hat{x}_{0|0} = \hat{x}_0$ and $P_{0|0} = P_0$.
2: Time update:
Given $\hat{x}_{k-1|k-1}$ and $P_{k-1|k-1}$, do

$$\begin{aligned}\hat{x}_{k|k-1} &= A\hat{x}_{k-1|k-1},\\ P_{k|k-1} &= AP_{k-1|k-1}A^T + \Sigma_w.\end{aligned}$$

3: Scheduling, transmission and measurement update:
Define $C^T = [C_1^T, \ldots, C_\ell^T]$, $\hat{x}_{k|k}^0 = \hat{x}_{k|k-1}$ and $P_{k|k}^0 = P_{k|k-1}$. For $i \geq 1$, let

$$\begin{aligned}\sigma_k^i &= \sqrt{C^i P_{k|k}^{i-1}(C^i)^T + 1},\\ z_k^i &= (y_k^i - C^i\hat{x}_{k|k}^{i-1})/\sigma_k^i.\end{aligned}$$

Sensor scheduling: Let the scheduling variable be given by

$$\gamma_k^i = \begin{cases} 0, & \text{if } |z_k^i| < \eta;\\ 1, & \text{otherwise.}\end{cases}$$

Data transmission: If $\gamma_k^i = 1$, send y_k^i to the estimator.
Measurement update: For $i = 1$ to ℓ, let

$$K_k^i = P_{k|k}^{i-1}(C^i)^T(C^iP_{k|k}^{i-1}(C^i)^T + 1)^{-1},$$

do

$$\begin{aligned}\hat{x}_{k|k}^i &= \hat{x}_{k|k}^{i-1} + \gamma_k^i K_k^i(y_k^i - C^i\hat{x}_{k|k}^{i-1});\\ P_{k|k}^i &= P_{k|k}^{i-1} - h(\gamma_k^i, \eta)K_k^iC^iP_{k|k}^{i-1},\end{aligned}$$

where $h(\gamma_k^i, \eta) = \gamma_k^i + (1 - \gamma_k^i)\sqrt{\frac{2}{\pi}} \times \frac{\eta\exp(-\eta^2/2)}{1-2Q(\eta)}$.
End, do $\hat{x}_{k|k} = \hat{x}_{k|k}^\ell$ and $P_{k|k} = P_{k|k}^\ell$.

Given a scheduling threshold $\eta > 0$, a sequential scheduler and the corresponding filter are proposed in Algorithm 14.1. If there is no scheduler applied, i.e., $\eta = 0$, it always holds that $\gamma_k^i = 1$. One can easily verify that Algorithm 14.1 is the same as the standard Kalman filter,[1] which is an MMSE estimator. However, if $\eta > 0$, it becomes a nonlinear filtering problem as the scheduler is typically a nonlinear function of the measurement. It is known that the exact MMSE estimator under a nonlinear measurement function is computationally intractable [5]. To derive a recursive filter, a common technique in the literature is to assume a Gaussian distribution of the predicted density to obtain an approximate MMSE estimator [3–8]. We will adopt

[1] Note that the noise covariance matrix $\Sigma_v = I$.

this method to derive Algorithm 14.1 as well. For compactness of notation, we introduce that $\gamma_k^0 = y_k^0 = 0$. Let

$$\mathscr{I}_k^i = \{\gamma_1^1 y_1^1, \gamma_1^1, \gamma_1^2 y_1^2, \gamma_1^2, \ldots, \gamma_k^i y_k^i, \gamma_k^i\}, \forall i \in \{0, \ldots, \ell\}.$$

Proposition 14.2.1 *If the predicted density in Algorithm 14.1 is approximately Gaussian, i.e.,*

$$x_k | \mathscr{I}_k^{i-1} \sim \mathscr{N}(\hat{x}_{k|k}^{i-1}, P_{k|k}^{i-1}),$$

then $\hat{x}_{k|k}^i$ is an MMSE estimator.

Proof To elaborate it, suppose that $\hat{x}_{k|k}^{i-1}$ is already an MMSE estimator, i.e.,

$$\hat{x}_{k|k}^{i-1} = \mathbb{E}[x_k | \mathscr{I}_k^{i-1}]$$

and

$$P_{k|k}^{i-1} = \mathbb{E}[(x_k - \hat{x}_{k|k}^{i-1})(x_k - \hat{x}_{k|k}^{i-1})^T | \mathscr{I}_k^{i-1}].$$

Let

$$\mathscr{G}_k^i = \{\gamma_1^1 y_1^1, \gamma_1^1, \gamma_1^2 y_1^2, \gamma_1^2, \ldots, \gamma_k^{i-1} y_k^{i-1}, \gamma_k^{i-1}, y_k^i\}.$$

Using the (assumed) Gaussian approximation and the technique in Kalman filtering [9], it follows that

$$\begin{aligned} \breve{x}_{k|k}^i &\triangleq \mathbb{E}[x_k | \mathscr{G}_k^i] = \hat{x}_{k|k}^{i-1} + K_k^i (y_k^i - C^i \hat{x}_{k|k}^{i-1}); \\ \breve{P}_{k|k}^i &\triangleq \mathbb{E}[(x_k - \breve{x}_{k|k}^i)(x_k - \breve{x}_{k|k}^i)^T | \mathscr{G}_k^i] \\ &= P_{k|k}^{i-1} - K_k^i C^i P_{k|k}^{i-1}. \end{aligned} \tag{14.3}$$

By iterative conditioning, it follows that

$$\begin{aligned} \hat{x}_{k|k}^i &= \mathbb{E}[x_k | \mathscr{I}_k^i] = \mathbb{E}[\mathbb{E}[x_k | \mathscr{G}_k^i] | \mathscr{I}_k^i] = \mathbb{E}[\breve{x}_{k|k}^i | \mathscr{I}_k^i] \\ &= \hat{x}_{k|k}^{i-1} + K_k^i \mathbb{E}[y_k^i - C^i \hat{x}_{k|k}^{i-1} | \mathscr{I}_k^i]. \end{aligned}$$

Thus, $\hat{x}_{k|k}^i = \hat{x}_{k|k}^{i-1} + K_k^i (y_k^i - C^i \hat{x}_{k|k}^{i-1})$ if $\gamma_k^i = 1$ and $\hat{x}_{k|k}^i = \hat{x}_{k|k}^{i-1}$ if $\gamma_k^i = 0$. Similarly, the estimation error covariance matrix is computed as follows.

$$\begin{aligned} P_{k|k}^i &= \mathbb{E}[\mathbb{E}[(x_k - \hat{x}_{k|k}^i)(x_k - \hat{x}_{k|k}^i)^T | \mathscr{G}_k^i] | \mathscr{I}_k^i] \\ &= \mathbb{E}[\mathbb{E}[(x_k - \breve{x}_{k|k}^i + \breve{x}_{k|k}^i - \hat{x}_{k|k}^i)(x_k - \breve{x}_{k|k}^i + \breve{x}_{k|k}^i - \hat{x}_{k|k}^i)^T | \mathscr{G}_k^i] | \mathscr{I}_k^i] \\ &= \breve{P}_{k|k}^i + K_k^i \mathbb{E}[(1 - \gamma_k^i)^2 (y_k^i - C^i \hat{x}_{k|k}^{i-1})^2 | \mathscr{I}_k^i] (K_k^i)^T, \end{aligned} \tag{14.4}$$

where the last equality uses the fact that

$$\mathbb{E}[(x_k - \breve{x}_{k|k}^i)(\breve{x}_{k|k}^i - \hat{x}_{k|k}^i)^T | \mathscr{G}_k^i] = 0.$$

If $\gamma_k^i = 1$, we have that

$$\mathbb{E}[(1-\gamma_k^i)^2(y_k^i - C^i\hat{x}_{k|k}^{i-1})^2|\mathscr{I}_k^i] = 0$$

and if $\gamma_k^i = 0$, it follows that

$$\begin{aligned}\mathbb{E}[(1-\gamma_k^i)^2(y_k^i - C^i\hat{x}_{k|k}^{i-1})^2|\mathscr{I}_k^i] &= \mathbb{E}[(y_k^i - C^i\hat{x}_{k|k}^{i-1})^2|\mathscr{I}_k^i](\sigma_k^i)^2 \\ &= \mathbb{E}[(z_k^i)^2|\mathscr{I}_k^{i-1}, |z_k^i| < \eta] \\ &= \frac{(\sigma_k^i)^2}{1-2Q(\eta)}\int_{-\eta}^{\eta}\frac{t^2}{\sqrt{2\pi}}\exp(-t^2/2)dt \qquad (14.5) \\ &= (\sigma_k^i)^2\left(1-\sqrt{\frac{2}{\pi}}\times\frac{\eta\exp(-\eta^2/2)}{1-2Q(\eta)}\right) \qquad (14.6)\end{aligned}$$

where (14.5) is due to that the conditional random variable $z_k^i|\mathscr{I}_k^{i-1}$ is a standard Gaussian random variable. Inserting (14.6) into (14.4), it finally yields that

$$P_{k|k}^i = P_{k|k}^{i-1} - h(\gamma_k^i, \eta)K_k^iC^iP_{k|k}^{i-1}.$$

If $\eta = 0$, it indeed holds that

$$x_k|\mathscr{I}_k^{i-1} \sim \mathscr{N}(\hat{x}_{k|k}^{i-1}, P_{k|k}^{i-1})$$

since Algorithm 14.1 is reduced to the Kalman filter. Intuitively, a smaller η will result in a more accurate Gaussian approximation. Moreover, under the Gaussian approximation, $\{\gamma_k^1, \ldots, \gamma_k^\ell\}_{k\geq 0}$ form an i.i.d. process [4]. Together with the strong law of large numbers [10], the scheduled transmission rate is evaluated by

$$R = \limsup_{k\to\infty}\frac{\sum_{j=1}^k(\gamma_j^1 + \cdots + \gamma_j^\ell)}{\ell k} = \mathbb{E}[\gamma_1^1] = 2Q(\eta). \qquad (14.7)$$

14.2.2 An Illustrative Example

We use a target tracking system [11] to test the above results via simulations. The dynamics of the target is expressed by

$$x_{k+1} = \begin{bmatrix}1 & h & h^2\\ 0 & 1 & h\\ 0 & 0 & 1\end{bmatrix}x_k + w_k, \qquad (14.8)$$

where h is the sampling period and x_k denotes the target state at time kh, including the target position, speed and acceleration. The input random signal w_k is an additive white Gaussian noise. When the sampling period h is sufficiently small, the covariance of w_k is given by

$$\Sigma_w = 2\alpha\sigma_m^2 \begin{bmatrix} h^5/20 & h^4/8 & h^3/6 \\ h^4/8 & h^3/3 & h^2/2 \\ h^3/6 & h^2/2 & h \end{bmatrix}, \tag{14.9}$$

where σ_m^2 is the variance of the target acceleration and α is the reciprocal of the maneuver time constant. The sensor periodically measures the target position with the following output equation:

$$y_k = [1, 0, 0]x_k + v_k, \tag{14.10}$$

where the measurement noise v_k is an additive white noise with unit variance and independent of w_k. The initial state x_0 is a Gaussian random vector with zero mean and covariance as follows [11]:

$$P_0 = \begin{bmatrix} 1 & 1/h & 0 \\ 1/h & 2/h^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In this example, set $h = 0.1$s, $\alpha = 0.1$ and $\sigma_m^2 = 1$. The objective is to track the target under a scheduled transmission rate $R = 0.6$. To achieve it, we compute that $\eta = 0.5$ by (14.7). The estimation error of position using Algorithm 14.1 is illustrated in Fig. 14.2.

Define $R_k = 1/k \cdot \sum_{j=1}^{k} \gamma_j$, which is drawn as the solid pink line in Fig. 14.2. Observe that R_k asymptotically converges to $R = 0.6$, which is consistent with (14.7). Figure 14.3 shows the model consistency of the Guassian approximation using the Kolmogorov-Smirnov (KS) goodness-of-fit tests [12] on the normalized estimation error (NEE), which is defined as

$$NEE_k = \mathbf{1}^T P_{k|k-1}^{-1/2} \cdot (\hat{x}_{k|k-1} - x_k)/\sqrt{n}. \tag{14.11}$$

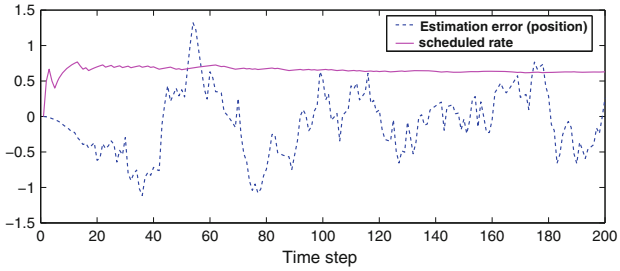

Fig. 14.2 Estimation error of position and the scheduled transmission rate

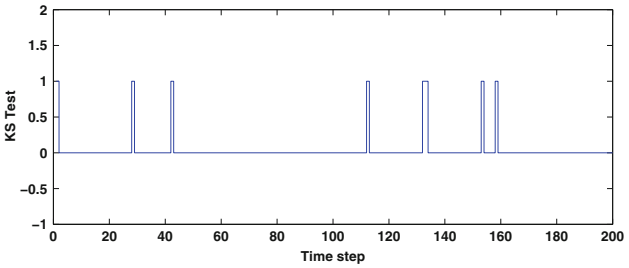

Fig. 14.3 Consistency test for the Gaussian approximation: $R = 0.6$

Here **1** is a column vector with all elements equal to one. If the estimator is consistent with the Gaussian approximation, NEE_k is postulated to be a standard Gaussian random variable (since $\hat{x}_{k|k-1} - x_k$ is assumed to be Gaussian with zero mean and covariance $P_{k|k-1}$). This is tested by running Monte Carlo simulations with $L = 300$ samples at each time, and a data set with size L at each time, i.e., $N_k := \{NNE_k^i\}_{i=1}^L$ is obtained. Then, we perform a KS test on N_k. The result is 1 if the test rejects the null hypothesis that N_k admits a standard Gaussian distribution at the 5 % significance level, 0 otherwise. We observe from Fig. 14.3 that the Gaussian hypothesis at only $4\,\% = 8/200$ of 200 times is rejected. Similarly, we use the KS test under the scheduling threshold η ranging from 0.5 to 1.3 with step size 0.1. By (14.7), we obtain the corresponding scheduled transmission rate, which is plotted as the horizontal axis in Fig. 14.4. The percentage of the times at which the Gaussian approximation is rejected by the KS test is shown in the vertical axis. It is observed that the percentage of the rejected times is less than 6 %, even when the scheduled transmission rate is as low as $R = 0.2$. □

Corollary 14.1 *Under a communication constraint, i.e., $R \leq r$, the optimal scheduling threshold to minimize $tr(P_{k|k})$ in Algorithm 14.1 is given by $\eta^* = Q^{-1}(r/2)$, where $Q^{-1}(\cdot)$ is the inverse function of $Q(\cdot)$.*

Proof By Algorithm 1, η^* should maximize $h(0, \eta)$ as well as satisfy the given communication constraint. This is exactly obtained by solving the following optimization:

$$\eta^* = \arg\max_{\{R \leq r\}} h(0, \eta). \tag{14.12}$$

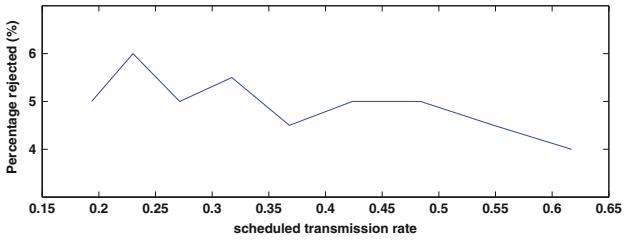

Fig. 14.4 Percentage of the rejected times and scheduled transmission rate

Since both $Q(\eta)$ and

$$h(0,\eta) = 1 - \int_{-\eta}^{\eta} \frac{t^2}{\sqrt{2\pi}} \exp(-t^2/2)dt$$

are decreasing in η, the optimization is attained at η^* satisfying $R = 2Q(\eta^*) = r$. □

14.2.3 Stability Analysis

For notational simplicity, let $P_k = P_{k|k-1}$. We study the mean stability of the approximate prediction error covariance, i.e., $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$.

Proposition 14.2.2 *Consider system* (14.1) *where A is unstable,* $(A, \Sigma_w^{1/2})$ *is controllable and* (C, A) *is observable. Under the Gaussian approximation, there exists a critical threshold* η^c *such that if* $\eta \geq \eta^c$*, then there exits a* $P_0 \geq 0$ *such that*

$$\lim_{k\to\infty} \mathbb{E}[P_k] = \infty,$$

and if $\eta < \eta^c$*, then for all* $P_0 \geq 0$ *it holds that*

$$\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty.$$

Proof By the Gaussian approximation, it follows that γ_k^i in Algorithm 14.1 forms an i.i.d. process. This implies that $h(\gamma_k^i, \eta)$ is also an i.i.d. process. Note that $h(\gamma_k^i, \eta)$ is a decreasing function in η. The rest of the proof follows from that of Theorem 2 in [13]. □

Another interesting problem is how to quantify the critical threshold η_c. To this purpose, define

$$\gamma(\eta) = \mathbb{E}[\gamma_k^1] = 2Q(\eta) \quad \text{and} \quad \zeta(\eta) = \frac{\eta \exp(-\eta^2/2)}{\sqrt{\pi/2} \cdot (1 - 2Q(\eta))}. \tag{14.13}$$

Then, $h(\gamma_k^i, \eta) = \gamma_k^i + (1 - \gamma_k^i)\zeta(\eta)$. Some interesting remarks are drawn below.

Remark 14.2 (a) Compared with the packet loss model in [13–15], $\gamma_k^i = 0$ in their works corresponds to the occurrence of a packet loss. Since the packet loss is independent of system evolution, the estimator can not obtain any measurement information at this time. Then, there is no measurement update and $P_{k|k}^i = P_{k|k}^{i-1}$. In the controllable scheduling algorithm, $\gamma_k^i = 0$ does provide partial information of y_k^i to the estimator, i.e., the normalized innovation lies in a certain bounded

region, which is uniquely determined by the threshold η, although y_k^i is not communicated to the estimator. This additional information helps us to improve our estimate and reduce the estimation error. The effect of this information on the estimation performance can be quantified by $\zeta(\eta)$. In fact, it follows from (14.6) that

$$\zeta(\eta) = 1 - 1/\sqrt{2\pi} \cdot \int_{-\eta}^{\eta} t^2 \exp(-t^2/2)dt \in [0,\ 1]$$

is strictly decreasing in η and

$$\lim_{\eta\to\infty} \zeta(\eta) = 1 - \lim_{\eta\to 0^+} \zeta(\eta) = 0.$$

(b) In comparison with the Kalman filter [9], the performance degradation can be characterized by

$$1 - h(\gamma_k^i, \eta) = (1 - \gamma_k^i)(1 - \zeta(\eta)).$$

If $\gamma_k^i = 1$, the estimator receives y_k^i. There is no information loss at this time and thus the measurement update is just the same as the Kalman filter, i.e., $1 - h(\gamma_k^i, \eta) = 0$. However, if $\gamma_k^i = 0$, the complete measurement information is unavailable. Thus, our estimator can not reduce the prediction error covariance matrix as much as the standard Kalman filter and the performance degradation is expressed by

$$1 - h(\gamma_k^i, \eta) = 1 - \zeta(\eta).$$

Proposition 14.2.3 *The critical threshold η^c in Lemma* 14.2.2 *satisfies that*

$$\eta^c \le \eta^*,$$

where η^* is the unique solution to the following equation:

$$(1 - \gamma(\eta))(1 - \zeta(\eta)) = \left(\prod_{i=1}^{n} \max\{|\lambda_i|, 1\}\right)^{-\frac{2}{\ell}}. \tag{14.14}$$

Here $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, and $\gamma(\eta)$, $\zeta(\eta)$ are defined in (14.13).

Remark 14.3 (a) By the Gaussian approximation, it follows that $h(\gamma_k^i, \eta)$ is an i.i.d. process and

$$1 - \mathbb{E}[h(\gamma_k^i, \eta)] = (1 - \gamma(\eta))(1 - \zeta(\eta)).$$

For a smaller η, it is more likely for the scheduler to trigger a measurement transmission. This is confirmed by $\gamma(\eta)$ as $Q(\eta)$ is a strictly deceasing function. Even if $\gamma_k^i = 0$, a smaller η is associated with less information loss since $1 - \zeta(\eta)$ is an increasing function. From this perspective, $1 - \gamma(\eta)$ controls the

measurement information loss rate while $1-\zeta(\eta)$ quantifies the amount of information loss if the complete information of y_k^i is unavailable to the estimator. Loosely speaking, the quantity

$$(1-\gamma(\eta))(1-\zeta(\eta))$$

characterizes the performance degradation as compared with the Kalman filter and we call it *performance degenerating factor* in this chapter.

(b) The quantity

$$M(A)=\prod_{i=1}^{n}\max\{|\lambda_i|,1\}$$

characterizes the degree of instability of the system and represents the intrinsic uncertainty growth rate generated by the system. In fact, $M(A)$ was introduced in 1960 [16], and named as Mahler measure. Note that the stability of Algorithm 14.1 is guaranteed only if $\eta<\eta^c$. This reveals that to achieve stability of Algorithm 14.1, the information loss rate needs to be strictly less than the uncertainty growth rate, which is consistent with our intuition.

(c) Since the *performance degenerating factor* is increasing in η, a larger m may result in a larger η^*. Observe that a larger m involves a larger number of iterations to be performed in the measurement update of Algorithm 14.1.

Proof of Proposition 14.2.3

By a coordinate transformation, we can write $A=\mathrm{diag}(A_s,A_u)$, where A_s is stable and all eigenvalues of A_u lie outside or on the unit circle. By Lemma 2 in [17], there is no loss of generality to focus on the unstable part because the mean square stability of the state variables corresponding to A_s will be achieved spontaneously. Thus, we directly assume that all eigenvalues of A lie outside or on the unit circle.

Since the *performance degenerating factor* on the left hand side of (14.14) is increasing in η and

$$\lim_{\eta\to\infty}\zeta(\eta)=1-\lim_{\eta\to 0^+}\zeta(\eta)=0,$$

there exists a positive and unique solution to (14.14), which is denoted by η^*. By Proposition 14.2.2, it is equivalent to showing that a necessary condition for

$$\sup_{k\in\mathbb{N}}\mathbb{E}[P_k]<\infty,\ \forall P_0\geq 0$$

is that $0\leq\eta<\eta^*$. Note that $\sup_{k\in\mathbb{N}}\mathbb{E}[P_k]<\infty$ if and only if $\sup_{k\in\mathbb{N}}\mathbb{E}[\mathrm{tr}(P_k)]<\infty$.

It follows from the inequality of arithmetic and geometric means that

$$(\det(P_k))^{1/n}\leq \mathrm{tr}(P_k)/n.$$

This implies that if $\sup_{k\in\mathbb{N}} \mathbb{E}[\mathrm{tr}(P_k)] < \infty$, then $\sup_{k\in\mathbb{N}} \mathbb{E}[\det(P_k)] < \infty$. By measurement update in Algorithm 14.1, it follows that

$$\begin{aligned}
\det(P_{k|k}^i) &= \det(P_{k|k}^{i-1} - h(\gamma_k^i, \eta)P_{k|k}^{i-1}(C^i)^T(C^iP_{k|k}^{i-1}(C^i)^T + 1)^{-1}C^iP_{k|k}^{i-1}) \\
&\geq \det(P_{k|k}^{i-1} - h(\gamma_k^i, \eta)P_{k|k}^{i-1}(C^i)^T(C^iP_{k|k}^{i-1}(C^i)^T)^{-1}C^iP_{k|k}^{i-1}) \\
&= \det(P_{k|k}^{i-1})\det(I - h(\gamma_k^i, \eta)(C^i)^T(C^iP_{k|k}^{i-1}(C^i)^T)^{-1}C^iP_{k|k}^{i-1})) \\
&= \det(P_{k|k}^{i-1})\det(1 - h(\gamma_k^i, \eta)(C^iP_{k|k}^{i-1}(C^i)^T)^{-1}C^iP_{k|k}^{i-1}(C^i)^T) \\
&= (1 - h(\gamma_k^i, \eta))\det(P_{k|k}^{i-1}),
\end{aligned}$$

where the third equality was obtained by applying the fact that

$$\det(I - EF) = \det(I - FE)$$

for matrices E and F with appropriate dimensions. Thus, we have that

$$\det(P_{k|k}) = \det(P_{k|k}^{\ell}) \geq \left(\prod_{i=1}^{\ell}(1 - h(\gamma_k^i, \eta))\right)\det(P_{k|k}^0).$$

By Minkowski inequality [1, Theorem 7.8.8], it follows that

$$\begin{aligned}
\det(P_{k+1|k}) &\geq \det(AP_{k|k}A^T) + \det(\Sigma_w) \\
&= \det(A^2)\det(P_{k|k}) + \det(\Sigma_w).
\end{aligned}$$

Combining the above and using the Gaussian approximation, we obtain that

$$\mathbb{E}[\det(P_{k+1})] \geq ((1 - \gamma(\eta))(1 - \zeta(\eta)))^{\ell}\det(A^2)\mathbb{E}[\det(P_k)] + \det(\Sigma_w).$$

By $\sup_{k\in\mathbb{N}} \mathbb{E}[\det(P_k)] < \infty$, it yields that

$$((1 - \gamma(\eta))(1 - \zeta(\eta)))^{\ell}\det(A^2) < 1,$$

which is equivalent to that

$$(1 - \gamma(\eta))(1 - \zeta(\eta)) < (M(A))^{-\frac{2}{\ell}}.$$

Since $(1 - \gamma(\eta))(1 - \zeta(\eta))$ is strictly increasing in η, it follows that $\eta < \eta^*$. □

Remark 14.4 By (14.7), the scheduled transmission rate should be greater than $2Q(\eta^*)$ to achieve stability of the filter in Algorithm 1.

14.3 Uncontrollable Scheduler

In the previous section, the scheduler is driven by the innovation process. It requires that the sensor node keep an identical copy of the state estimate as the remote estimator. The sensor thus needs to be embedded with computing capability. While in some scenarios, sensor nodes have no or very limited computing power. They only passively transmit their measurements to the estimato, and randomly wake up to take measurements. To investigate this phenomenon, we consider an *uncontrollable* scheduler driven by an i.i.d. random process, which determines the time duration between consecutive measurement transmissions of the sensor node without referring to the measurements.

14.3.1 Intermittent Kalman Filter

Mathematically, $\{\gamma_k\}_{k\geq 0}$ is independent of $\{x_k\}_{k\geq 0}$ and $\{y_k\}_{k\geq 0}$. The estimation problem is then cast as Kalman filtering with intermittent observations. By [13], it is known that the Kalman update still results in an MMSE estimator, which is computed by the following recursions.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k(y_k - C\hat{x}_{k|k-1}); \tag{14.15}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C P_{k|k-1}, \tag{14.16}$$

where the Kalman gain is given by

$$K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + \Sigma_v)^{-1}.$$

Moreover, the time update continues to hold:

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k},\ P_{k+1|k} = AP_{k|k}A^T + \Sigma_w$$

and

$$\hat{x}_{0|0} = \hat{x}_0,\ P_{0|0} = P_0.$$

Assume that $\gamma_0 = 1$ and $t_0 = 0$, define a sequence of measurement transmission times $\{t_k\}_{k\geq 0}$ at which y_{t_k} is communicated to the estimator as follows:

$$\begin{aligned} t_1 &= \inf\{k|k \geq 1, \gamma_k = 1\}, \\ t_2 &= \inf\{k|k > t_1, \gamma_k = 1\}, \\ &\vdots \quad \vdots \\ t_j &= \inf\{k|k > t_{j-1}, \gamma_k = 1\}. \end{aligned} \tag{14.17}$$

If there exists a positive probability such that $t_k = \infty$ for some $k < \infty$, one can easily verify that the filter will eventually diverge in the average sense if A is unstable. Thus, a necessary condition for any stabilizing scheduler is that t_k should be finite with probability one. The integer valued sojourn process $\{\tau_k\}_{k\geq 1}$ with τ_k denoting the time duration between consecutive sensor transmissions is thus well defined, where $\tau_k \triangleq t_k - t_{k-1}$.

For simplicity of exposition, let $P_{k+1} = P_{k+1|k}$ and $M_k = P_{t_k+1}$. We obtain the following random Riccati recursion:

$$P_{k+1} = AP_kA^T + \Sigma_w - \gamma_k AP_kC^T(CP_kC^T + \Sigma_v)^{-1}CP_kA^T. \tag{14.18}$$

Due to the randomness of γ_k, the stability analysis of P_k is generically challenging. In the literature, it is assumed that γ_k follows an i.i.d. process [13] or Markov process [15] to model the random variation of the underlying network for transmitting raw measurements to the remote estimator. As in [15], we introduce two types of stability notion.

Definition 14.1 We say that the state estimation error covariance matrices are stable in sampling times if $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$ while they are stable in stopping times if $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ for any $P_0 > 0$.

Here $\mathbb{E}[P_k]$ represents the mean of one-step prediction error covariance at the sampling time, whereas $\mathbb{E}[M_k]$ denotes the mean of one-step prediction error covariance matrix at the stopping time. To some extent, the former is time-driven while the latter is event-driven.

In a real application, if the estimator does not receive any measurement from the sensor for a long time, this sensor will be usually treated as a dead sensor and no longer be used. To incorporate this, it is sensible to assume that τ_k can only take values from a finite set, denoted by $\mathscr{L} \triangleq \{1, \ldots, L\}$ with $L < \infty$. Moreover, we make the following assumption to explicitly characterize the effect of the uncontrollable scheduler on the stability of the error covariance matrix.

Assumption 14.1 τ_k is an i.i.d. process with $\mathbb{P}\{\tau_1 = i\} = p_i \in [0, 1]$ and $\sum_{i=1}^{L} p_i = 1$.

Under the Markovian packet loss model, i.e., $\{\gamma_k\}_{k\geq 0}$ is a Markov process, the sojourn time τ_k is an i.i.d. process taking an infinite number of values [15]. This gives us an impression that the stability analysis of the Kalman filter with an uncontrollable scheduler under Assumption 14.1 appears to be easier since $L < \infty$. This is not the case as here the statistics of the underlying process γ_k is unrecoverable from the sojourn process $\{\tau_k\}_{k\geq 1}$. For an illustration, we construct an example as follows. Consider the process $\{\gamma_k\}_{k\geq 1}$ with an initial distribution $\mathbb{P}\{\gamma_1 = 1\} = 1 - \mathbb{P}\{\gamma_1 = 0\} = p$. For $k \geq 1$, let $\gamma_{2k} = 1 - \gamma_1$ and $\gamma_{2k+1} = \gamma_1$. It is straightforward to verify that $\mathbb{P}\{\tau_1 = 1\} = \mathbb{P}\{\tau_k = 1\} = 1$ for all $k \geq 1, p \in (0, 1)$. Clearly, the statistics of γ_k varies over p but the statistics of the sojourn process does not depend on the exact value of p.

Thus, it is impossible to recover the statistics of γ_k from the sojourn process. This implies that the approach by using the statistics of γ_k in the packet loss model [13–15] is inapplicable in the current framework. Note that the Markovian properties of γ_k are vital to [15].

Remark 14.5 Define $j_k \triangleq \max\{j | t_j \leq k\}$, then for a sufficiently large k, it follows that

$$\frac{k-L}{k} \leq \frac{t_{j_k}}{k} \leq 1.$$

This implies that

$$\limsup_{k\to\infty} \frac{t_{j_k}}{k} = 1$$

and the scheduled transmission rate is computed by

$$\begin{aligned} R &= \limsup_{k\to\infty} \frac{j_k}{k} = \limsup_{k\to\infty} \frac{j_k}{\tau_1 + \cdots + \tau_{j_k}} \cdot \limsup_{k\to\infty} \frac{t_{j_k}}{k} \\ &= \frac{1}{\mathbb{E}[\tau_1]}, \end{aligned} \tag{14.19}$$

where the last equality is due to the strong law of large numbers [10]. By fixing p_L to be a positive constant and letting L be sufficiently large, it follows from (14.19) that the scheduled transmission rate can be made arbitrarily small.

14.3.2 Second-Order System

If C is of full column rank, then for any k, a suboptimal and stable estimate of x_k is constructed as

$$\breve{x}_k = A^{k-t_{j_k}} C^{-1} y_{t_{j_k}}. \tag{14.20}$$

By the optimality of the MMSE estimator, it follows that

$$\begin{aligned} P_k &\leq A\mathbb{E}[(x_k - \breve{x}_k)(x_k - \breve{x}_k)^T | \mathscr{F}_k] A^T \\ &\leq A^{k-t_{j_k}+1} C^{-1} \Sigma_v C^{-T} (A^{k-t_{j_k}+1})^T < \infty, \forall k. \end{aligned}$$

Thus, it is trivial to achieve a stable filter for a full column rank C under Assumption 14.1. To this purpose, only column rank deficient C is to be considered. In particular, let $\text{rank}(C) = 1$ in the second-order system (14.1). The open-loop matrix A is categorized into the following cases:

C1: A is not diagonalizable;
C2: $A = \text{diag}(\lambda_1, \lambda_2)$ and $|\lambda_1| \neq |\lambda_2|$;

C3: $A = \text{diag}(\lambda_1, \lambda_2)$ and $\lambda_2 = \lambda_1 \exp(2\pi\varphi\mathbb{i})$ with $\mathbb{i}^2 = -1$ and φ is an irrational number;

C4: $A = \text{diag}(\lambda_1, \lambda_2)$ and $\lambda_2 = \lambda_1 \exp(\frac{2\pi r}{d}\mathbb{i})$, where $d > r \geq 1$ and $r, d \in \mathbb{N}$ are irreducible.

Note that if $A = \text{diag}(\lambda_1, \lambda_2)$ with $\lambda_1 = \lambda_2$ and $\text{rank}(C) = 1$, then (C, A) is unobservable. Under such a case, the Kalman filter will diverge if $|\lambda_1| \geq 1$ [9].

Under C1–C3, the pair $(C, A^k), \forall k > 1$ continues to be observable if (C, A) is observable while the pair (C, A^{kd}) will lose observability in case C4. This intuitively implies that the stability condition for C4 may be stronger than the other cases. In fact, under C4, the measurements $\{y_{kd}, k \in \mathbb{N}\}$ contain redundant information and only provide information for observing one mode. Thus to obtain a stable filter, the estimator has to resort to measurements not belonging to $\{y_{kd}, k \in \mathbb{N}\}$. The above statement is rigorously confirmed in Theorem 14.1.

Theorem 14.1 *Consider a second-order system* (14.1) *satisfying that A is unstable,* $(A, \Sigma_w^{1/2})$ *is controllable and* (C, A) *is observable. Suppose that the uncontrollable scheduler satisfies Assumption* (14.1)*. Then,*

(a) *if A satisfies C4, rank*$(C) = 1$ *and* $d \leq L$*, a necessary and sufficient condition for* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$|\lambda_1|^{2d} p_d + \cdots + |\lambda_1|^{2\mu d} p_{\mu d} < 1,$$

where $\mu = \max\{j \in \mathbb{N} | jd \leq L\}$*;*

(b) *otherwise, it always holds that* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *and* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$.

The proof depends on a lemma given in [15]. If A is invertible, define $\phi(k, i) = A^{t_i - t_k}$ if $k \geq i$ and $\phi(k, i) = I$ if $k < i$. Let

$$\Lambda_k = \sum_{j=0}^{k} \phi^H(k, j) C^H C \phi(k, j) + \phi^H(k, 0)\phi(k, 0).$$

Proof of Theorem 14.1

(a) "⇐:" Define an integer valued set $\mathscr{S}_d = \{kd | \forall k \in \mathbb{N}\}$ and let $C = [c_1, c_2]$, then for any $j \in \{1, \ldots, k\}$, it follows that

$$\sum_{i=j-1}^{j} \phi^H(k, i) C^H C \phi(k, i) \tag{14.21}$$
$$= \phi^H(k, j) \begin{bmatrix} c_1 & \\ & c_2 \end{bmatrix} \begin{bmatrix} 1 + \lambda_1^{-2\tau_j} & 1 + \lambda_1^{-\tau_j}\lambda_2^{-\tau_j} \\ 1 + \lambda_1^{-\tau_j}\lambda_2^{-\tau_j} & 1 + \lambda_2^{-2\tau_j} \end{bmatrix} \begin{bmatrix} c_1 & \\ & c_2 \end{bmatrix} \phi(k, j).$$

Define

$$\mathscr{R}_j = \begin{bmatrix} 1 + \lambda_1^{-2\tau_j} & 1 + \lambda_1^{-\tau_j}\lambda_2^{-\tau_j} \\ 1 + \lambda_1^{-\tau_j}\lambda_2^{-\tau_j} & 1 + \lambda_2^{-2\tau_j} \end{bmatrix},$$

then if $\tau_j \notin \mathscr{S}_d$, it yields that

$$\mathscr{R}_j^{-1} \leq \frac{4}{\lambda_1^{-2\tau_j} + \lambda_2^{-2\tau_j} - 2\lambda_1^{-\tau_j}\lambda_2^{-\tau_j}} I \leq \frac{2|\lambda_1|^{2\tau_j}}{1 - \cos(\frac{2\pi}{d})} I.$$

Let $c = \max\{c_1^{-2}, c_2^{-2}\}$ and $\alpha_1 = \max\{\frac{2c}{1-\cos(\frac{2\pi}{d})}, 1\}$, it follows from (14.21) that if $\tau_j \notin \mathscr{S}_d$ and $\tau_{j+1}, \ldots, \tau_k \in \mathscr{S}_d$, then

$$\Lambda_k^{-1} \leq \frac{2c|\lambda_1|^{2(t_k - t_{j-1})}}{1 - \cos(\frac{2\pi}{d})} I \leq \alpha_1 |\lambda_1|^{2(t_k - t_{j-1})} I.$$

Similarly, if $\tau_k \notin \mathscr{S}_d$, then $\Lambda_k^{-1} \leq \alpha_1 |\lambda_1|^{2\tau_k} I$. If $\tau_j \in \mathscr{S}_d$, $\forall j \in \{1, \ldots, k\}$, it obtains that

$$\Lambda_k^{-1} \leq (\phi^H(k,0)\phi(k,0))^{-1} \leq |\lambda_1|^{2t_k} I \leq \alpha_1 |\lambda_1|^{2t_k} I.$$

Define $E_k = \{\tau_k \notin \mathscr{S}_d\}$, $E_j = \{\tau_j \notin \mathscr{S}_d, \tau_{j+1}, \ldots, \tau_k \in \mathscr{S}_d\}$, $1 \leq j < k$ and $E_0 = \{\tau_1 \in \mathscr{S}_d, \ldots, \tau_k \in \mathscr{S}_d\}$. Let $\theta = \mathbb{P}\{\tau_1 \notin \mathscr{S}_d\}$, then by Assumption 14.1, it follows that $\mathbb{P}(E_j) = \theta(1-\theta)^{k-j}$, $1 \leq j \leq k$ and $\mathbb{P}(E_0) = (1-\theta)^k$. Jointly with that $\sum_{j=0}^k \mathbb{P}(E_j) = 1$ and Assumption 14.1, we obtain that

$$\begin{aligned}
\mathbb{E}[\Lambda_k^{-1}] &= \sum_{j=0}^{k} \mathbb{E}[\Lambda_k^{-1} 1_{E_j}] \leq \alpha_1 \sum_{j=0}^{k} \mathbb{E}[|\lambda_1|^{2(t_k - t_j)} 1_{E_j}] I \\
&= \alpha_1 \sum_{j=1}^{k-1} \mathbb{E}[(\prod_{i=j+1}^{k-1} |\lambda_1|^{2\tau_i} 1_{\{\tau_i \in \mathscr{S}_d\}}) |\lambda_1|^{2\tau_j} 1_{\{\tau_j \notin \mathscr{S}_d\}}] I \\
&\quad + \alpha_1 \mathbb{E}[\prod_{i=1}^{k} |\lambda_1|^{2\tau_i} 1_{\{\tau_i \in \mathscr{S}_d\}}] I + \alpha_1 \mathbb{E}[|\lambda_1|^{2\tau_k} 1_{\{\tau_k \notin \mathscr{S}_d\}}] I \\
&\leq \alpha_1 |\lambda_1|^{2L} (\sum_{j=1}^{k} (\mathbb{E}[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}])^{j-1} + 1) I
\end{aligned} \tag{14.22}$$

which is finite if and only if

$$\mathbb{E}[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}] = |\lambda_1|^{2d} p_d + \cdots + |\lambda_1|^{2\mu d} p_{\mu d} < 1.$$

The last inequality is due to that $\tau_k \leq L$ for any $k \geq 1$.

"$\Rightarrow$:" Under Assumption 14.1, it is easy to verify that the following random vectors are with an identical distribution, e.g.,

$$(\tau_k, \tau_k + \tau_{k-1}, \ldots, \tau_k + \cdots + \tau_1) \stackrel{d}{=} (\tau_1, \tau_1 + \tau_2, \ldots, \tau_1 + \cdots + \tau_k),$$

where $\stackrel{d}{=}$ means equal in distribution on its both sides. Let

$$\Xi_k = \sum_{j=0}^{k} \phi^H(j,0) C^H C \phi(j,0) + \phi^H(k,0)\phi(k,0),$$

this implies that

$$\mathbb{E}[\Lambda_k^{-1}] = \mathbb{E}[\Xi_k^{-1}]. \tag{14.23}$$

Let $F_0 = \{\tau_1 \notin \mathscr{S}_d\}, F_j = \{\tau_1 \in \mathscr{S}_d, \ldots, \tau_j \in \mathscr{S}_d, \tau_{j+1} \notin \mathscr{S}_d\}$ for $1 \le j < k$ and $F_k = \{\tau_1 \in \mathscr{S}_d, \ldots, \tau_k \in \mathscr{S}_d\}$. Similarly, it is easy to verify that $\sum_{j=0}^{k} \mathbb{P}(F_j) = 1$. For $j < k$, denote

$$\begin{aligned}
\Theta_j &= \sum_{i=1}^{k-j} \phi^H(j+i,j) C^H C \phi(j+i,j) \\
&\le \|C^H C\| \sum_{i=1}^{\infty} |\lambda_1|^{-2(t_{i+j}-t_j)} I \\
&\le \frac{\|C^H C\|}{1-|\lambda_1|^{-2}} I \triangleq \beta_0 I.
\end{aligned} \tag{14.24}$$

Combining the above, we obtain that

$$\begin{aligned}
\Xi_k^{-1} 1_{F_j} &= \Big(\sum_{i=0}^{j} |\lambda_1|^{-2t_i} C^H C + \phi^H(j,0)\Theta_j \phi(j,0) + \phi^H(k,0)\phi(k,0)\Big)^{-1} 1_{F_j} \\
&\ge \Big(\frac{1}{1-|\lambda_1|^{-2}} C^H C + \beta_0 \phi^H(j,0)\phi(j,0) + \phi^H(j,0)\phi(j,0)\Big)^{-1} 1_{F_j} \\
&\ge \beta_1 (C^H C + \phi^H(j,0)\phi(j,0))^{-1} 1_{F_j},
\end{aligned} \tag{14.25}$$

where

$$\beta_1 = \max\{\beta_0 + 1, \frac{1}{1-|\lambda_1|^{-2}}\}.$$

Thus, it is not difficult to show that there exists a positive constant β_2 such that

$$\mathrm{tr}(\Xi_k^{-1} 1_{F_j}) \ge \beta_2 |\lambda_1|^{2t_j} 1_{F_j}.$$

By (14.23), it follows that

$$\mathrm{tr}(\mathbb{E}[\Lambda_k^{-1}]) = \mathrm{tr}(\mathbb{E}[\Xi_k^{-1}])$$

$$= \sum_{j=0}^{k} \mathrm{tr}(\mathbb{E}[\Xi_k^{-1} 1_{F_j}]) \geq \beta_2 \sum_{j=1}^{k-1} \mathbb{E}[|\lambda_1|^{2t_j} 1_{F_j}]$$
$$= \beta_2 \theta \sum_{j=1}^{k-1} \left(\mathbb{E}[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}] \right)^j. \tag{14.26}$$

In the above, the last equality follows from Assumption 14.1. It follows from Lemma 13.3 that $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ is equivalent to $\sup_{k\in\mathbb{N}} \mathrm{tr}(\mathbb{E}[\Lambda_k^{-1}]) < \infty$. This implies

$$\mathbb{E}[|\lambda_1|^{2\tau_1} 1_{\{\tau_1 \in \mathscr{S}_d\}}] = |\lambda_1|^{2d} p_d + \cdots + |\lambda_1|^{2\mu d} p_{\mu d} < 1.$$

(b) For any $k > 2L$, the estimator must receive at least two measurements during the time interval from $k - 2L$ to k.

Define $k_1 = \max\{j \leq k | \gamma_j = 1\}$ and $k_2 = \max\{j < k_1 | \gamma_j = 1\}$, i.e., y_{k_1} and y_{k_2} are two latest consecutively received packets at time k. By iterations, it follows that

$$\begin{aligned} x_{k_1} &= A^{k_1-k_2} x_{k_2} + A^{k_1-k_2-1} w_{k_2} + \cdots + w_{k_1-1} \\ &\triangleq A^{k_1-k_2} x_{k_2} + w_{k_2:k_1}. \end{aligned}$$

Then, a stable estimator for x_k can be obtained by the least square method. To be precise, we write y_{k_2} and y_{k_1} as follows:

$$\begin{aligned} y_{k_2} &= C x_{k_2} + v_{k_2}; \\ y_{k_1} &= C A^{k_1-k_2} x_{k_2} + C w_{k_2:k_1} + v_{k_1}. \end{aligned} \tag{14.27}$$

Moreover, the above can be rewritten in a compact form by $Y_k = \Phi x_{k_2} + V_k$, where

$$Y_k = \begin{bmatrix} y_{k_2} \\ y_{k_1} \end{bmatrix}, \Phi = \begin{bmatrix} C \\ CA^{k_1-k_2} \end{bmatrix} \text{ and } V_k = \begin{bmatrix} v_{k_2} \\ Cw_{k_2:k_1} + v_{k_1} \end{bmatrix}.$$

Then, one can verify that under C1–C3, Φ is always of full rank. Similarly, Φ is also of full rank under C4 with $L < d$. Note that $k_1 - k_2 \leq L < d$, this implies that the covariance matrix of V_k is uniformly bounded, i.e., there exists a $\bar{V} > 0$ such that $cov(V_k) < \bar{V}$. A least mean square estimator of x_k is now given by

$$\breve{x}_k = A^{k-k_2} (\Phi^H \Phi)^{-1} \Phi^H Y_k.$$

By the optimality of the Kalman filter, it follows that

$$\begin{aligned} P_k &\leq \mathbb{E}[(x_k - \breve{x}_k)(x_k - \breve{x}_k)^T | \mathscr{F}_k] \\ &\leq A^{k-k_2} (\Phi^H \Phi)^{-1} \Phi^H \bar{V} \Phi (\Phi^H \Phi)^{-1} (A^{k-k_2})^T, \end{aligned}$$

which is uniformly bounded due to $k - k_2 < 2L$.

Thus, it is obvious that

$$\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty.$$

Similarly, one can establish that $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$. □

Remark 14.6 It is worthy mentioning that if $L = \infty$, the condition in (a) of Theorem 14.1 is simply modified by

$$\sum_{j=1}^{\infty} |\lambda_1|^{2j} p_{2dj} < 1.$$

14.3.3 Higher-Order System

In general, the stability analysis of random Riccati recursion (14.18) for higher-order systems with vector measurements is quite involved and depends on the so-called eigenvalue cycles of the open-loop matrix [18]. Heuristically, scalar measurement is the worst since a vector measurement generally supplies more information than a scalar observation. For this reason, we focus on the scalar measurement in the stability analysis of the random Riccati recursion (14.18). In particular, the following two cases are to be considered.

C1': $A^{-1} = \text{diag}(J_1, \ldots, J_q)$ and $\text{rank}(C) = 1$, where $J_i = J(\lambda_i^{-1})$ is an elementary Jordan block associated with the eigenvalue of λ_i^{-1} [1] and $|\lambda_i| > |\lambda_{i+1}|$.

C2': $A = \text{diag}(A_1, \ldots, A_q)$ and $\text{rank}(C) = 1$, where $A_i = \text{diag}(\lambda_{i,1}, \ldots, \lambda_{i,l_i})$ and there exists a minimum positive integer d_i such that $\lambda_{i,1}^{d_i} = \cdots = \lambda_{i,l_i}^{d_i}$. In addition, $\lambda_{i_1,1}^{k} \neq \lambda_{i_2,1}^{k}$ for all $k \geq 1$ if $i_1 \neq i_2$.

Theorem 14.2 *Consider system* (14.1) *satisfying that A is unstable,* $(A, \Sigma_w^{1/2})$ *is controllable and* (C, A) *is observable. Suppose that the scheduler satisfies Assumption* 14.1.

(a) *If the pair* (C, A) *satisfies C1', then* $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *and* $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$.
(b) *If the pair* (C, A) *satisfies C2', define* $\mu_i = \max\{j \in \mathbb{N} | jd_i \leq L\}$ *and*

$$\rho_i = \begin{cases} \sum_{j=1}^{\mu_i} |\lambda_{i,1}|^{2jd_i} p_{jd_i}, & \text{if } \mu_i \geq 1, d_i > 1; \\ 0, & \text{otherwise.} \end{cases} \tag{14.28}$$

The necessary and sufficient condition for $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ *is that*

$$\max_{i\in\{1,\ldots,q\}} \rho_i < 1.$$

Proof (a) The idea of the proof is similar to that in Theorem 14.1 (b). For a sufficiently large k, define a sequence of $\{k_i\}_{i=1}^{n}$ as follows.

Let $k_1 = \max\{j \le k | \gamma_j = 1\}$ and for $j \in \{2, \ldots, n\}$,

$$k_j = \max\{j < k_{j-1} | \gamma_j = 1, \text{rank} \begin{pmatrix} C \\ CA^{k_1-k_2} \\ \vdots \\ CA^{k_1-k_j} \end{pmatrix} = j\}.$$

The remaining problem is whether there exists such a sequence. This is guaranteed by the fact that $L < \infty$ and Lemma 21 in [15]. In addition, it can be easily proved that $k_1 - k_n$ is uniformly bounded irrespect to k. Note that $k - k_1 \le L$, we can construct a uniformly stable estimator for x_k by the least square method based on $\{y_{k_1}, \ldots, y_{k_n}\}$ as in the proof of Theorem 14.1 (b). Thus, it follows that $\sup_{k\in\mathbb{N}} \mathbb{E}[P_k] < \infty$. The proof of $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ can be established as well.

(b) "⇒:" Partitioning $\phi(k, i)$ and C in conformity with A, we obtain that there exists a constant $\alpha_2 \ge 1$ satisfying that

$$C^H C \le \alpha_2 \cdot \text{diag}(C_1^H C_1, \ldots, C_q^H C_q).$$

Then, it is trivial that

$$\Lambda_k \le \alpha_2 \cdot \text{diag}(\Lambda_{k,1}, \ldots, \Lambda_{k,q}),$$

where for $i \in \{1, \ldots, q\}$,

$$\Lambda_{k,i} = \sum_{j=0}^{k} \phi_i^H(k, j) C_i^H C_i \phi_i(k, j) + \phi_i^H(k, 0)\phi_i(k, 0).$$

Since $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$, it follows from Lemma 13.3 that $\sup_{k\in\mathbb{N}} \mathbb{E}[\Lambda_k^{-1}] < \infty$. This implies that $\sup_{k\in\mathbb{N}} \mathbb{E}[\Lambda_{k,i}^{-1}] < \infty$ for all $i \in \{1, \ldots, q\}$. The rest of the proof directly follows from the necessity proof of Theorem 14.1 (a).

"⇐:" According to the partition of A, the state vector is decomposed into q blocks, denoted as $x_k^{(1)}, \ldots, x_k^{(q)}$. It is sufficient to prove that the mean of the state estimation error covariance matrix of each block is bounded if $\max_i\{\rho_i\} < 1$. Two cases are to be separately considered.

Case 1: $d_1 = \cdots = d_q = 1$, the proof trivially follows from that of part (a).
Case 2: $d = \prod_{i=1}^{q} d_i > 1$. Only this case needs to be elaborated.

Let $z_k^{(i)} = C_i x_{kd}^{(i)}$. Since $A_i^d = \lambda_{i,1}^d I$, it follows from (14.1) that

$$z_{k+1}^{(i)} = \lambda_i^d z_k^{(i)} + \widetilde{w}_k^{(i)},$$

where

$$\widetilde{w}_k^{(i)} = C_i \sum_{j=1}^{d-1} A_i^{d-1-j} w_{kd+j}^{(i)}.$$

Then, we obtain the following down-sampled system:

$$z_{k+1} = \text{diag}(\lambda_{1,1}^d, \ldots, \lambda_{q,1}^d) z_k + \widetilde{w}_k;$$

$$s_k \triangleq y_{kd} = [1, \ldots, 1] z_k + v_{kd}.$$

Note that $\sup_{k\in\mathbb{N}} \mathbb{E}[\widetilde{w}_k \widetilde{w}_k^T] < \infty$ and $\lambda_{i_1,1}^k \neq \lambda_{i_2,1}^k$ for all $k \geq 1$ if $i_1 \neq i_2$, it follows from Case 1 that z_k can be estimated with a uniformly bounded error covariance matrix. This implies that it is able to estimate $C_i x_k^{(i)}$ with a bounded error as $d < \infty$.

Next, we prove that $x_k^{(i)}$ can be estimated with a uniformly bounded error if $\rho_i < 1$. To this purpose, we can focus on the following decoupled sub-system

$$\begin{aligned} x_{k+1}^{(i)} &= A_i x_k^{(i)} + w_k^{(i)}; \\ r_k^{(i)} &\triangleq y_k - \sum_{j\neq i} C_j \hat{x}_k^{(j)} = C_i x_k^{(i)} + v_k + \sum_{j\neq i} C_j \widetilde{x}_k^{(j)}, \\ &\triangleq C_i x_k^{(i)} + \widetilde{v}_k, \end{aligned} \tag{14.29}$$

where $C_j \hat{x}_k^{(j)}$ is an estimator of $C_j x_k^{(j)}$ with a uniformly bounded estimation error and $C_j \widetilde{x}_k^{(j)} = C_j x_k^{(j)} - C_j \hat{x}_k^{(j)}$. Hence, $\sup_{k\in\mathbb{N}} \mathbb{E}[\widetilde{v}_k^2] < \infty$.

In summary, there is no loss of generality to focus on that $q = 1$ under C2'. Since $q = 1$, the system has a similar structure as the second-order case under C4. The proof thus essentially follows from the sufficiency proof of Theorem 14.1 (a). The details are omitted. □

Remark 14.7 Under C2', when each $\lambda_{i,j}$ is replaced by an elementary Jordan block, i.e., $J_{i,j} = J(\lambda_{i,j})$, the necessary and sufficient condition for $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k] < \infty$ is the same as that in Theorem 14.2 (b). Nonetheless, the proof is tedious and is not included in this chapter due to the risk of diverting the readers' attention.

14.4 Summary

Motivated by the scarcity of computational capacity and energy in WSNs, the state estimation problems of dynamical systems under scheduled measurements have been investigated. Both controllable and uncontrollable schedulers were considered under different scenarios. In particular, the controllable scheduling protocol sequentially decides the transmission of each element of the vector measurement. The stability analysis was performed as well. While for the uncontrollable scheduler, some necessary and sufficient conditions for the stability of the Kalman filter with scheduled measurements were investigated. We expect that the estimation framework would be useful to reduce the communication cost in WSNs.

References

1. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)
2. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
3. K. You, L. Xie, S. Sun, W. Xiao, Multiple-level quantized innovation Kalman filter, in *Proceedings of the 17th IFAC World Congress*, pp. 1420–1425 (2008)
4. J. Wu, Q. Jia, K. Johansson, L. Shi, Event-based sensor data scheduling: trade-off between sensor communication rate and estimation quality. preprint (2011)
5. K. Ito, K. Xiong, Gaussian filters for nonlinear filtering problems. IEEE Trans. Autom. Control **45**(5), 910–927 (2000)
6. M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
7. S. Julier, J. Uhlmann, Unscented filtering and nonlinear estimation. Proc. IEEE **92**(3), 401–422 (2004)
8. A. Ribeiro, G. Giannakis, S. Roumeliotis, SOI-KF: distributed Kalman filtering with low-cost communications using the sign of innovations. IEEE Trans. Signal Process. **54**(12), 4782–4795 (2006)
9. B. Anderson, B. Moore, *Optimal Filtering*. Systems Sciences Series (Prentice-hall, New Jersey, 1979)
10. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)
11. R. Singer, Estimating optimal filter tracking performance for manned maneuvering targets. IEEE Trans. Aerosp. Electron. Syst. **6**(4), 473–483 (1970)
12. M. Stephens, EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**(347), 730–737 (1974)
13. B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, S. Sastry, Kalman filtering with intermittent observations. IEEE Trans. Autom. Control **49**(9), 1453–1464 (2004)
14. M. Huang, S. Dey, Stability of Kalman filtering with Markovian packet losses. Automatica **43**(4), 598–607 (2007)
15. K. You, M. Fu, L. Xie, Mean square stability for Kalman filtering with Markovian packet losses. Automatica **47**(12), 2647–2657 (2011)
16. K. Mahler, An application of Jensen's formula to polynomials. Mathematica **7**, 98–100 (1960)
17. N. Xiao, L. Xie, L. Qiu, Mean square stabilization of multi-input systems over stochastic multiplicative channels, in *Proceedings of the 48th IEEE Conference on Decision and Control*, pp. 6893–6898 (2009)
18. S. Park, A. Sahai, Intermittent Kalman filtering: eigenvalue cycles and nonuniform sampling, in *American Control Conference*, pp. 3692–3697 (2011)

Chapter 15
Parameter Estimation with Scheduled Measurements

In the previous chapter, we only discuss the stability of estimator of a dynamical system under scheduled measurements, leaving performance evaluation untouched. In comparison, this chapter considers the parameter estimation of a linear system under scheduled measurements.

Given a scheduled transmission rate constraint, we jointly design a near optimal scheduler and an estimator. The effect of the scheduler on the estimation performance can be precisely evaluated. Secondly, we conduct an asymptotic stochastic analysis on the estimation algorithm with respect to the number of sensor measurements where conditions for strong consistency and asymptotic normality of the estimated parameters are established. Thirdly, we propose an adaptive scheduler and a recursive estimation algorithm to show that even though the scheduler significantly reduces the communication cost, our estimation algorithm and its performance in terms of mean square estimation error are comparable to the standard least square estimator (LSE) which is based on the full set of measurements. Moreover, it is illustrated that the computational complexity of both estimators is almost identical.

The chapter is organized as follows. The problem formulation is described in Sect. 15.1. To obtain the CRLB, we investigate the maximum likelihood estimation with scheduled measurements in Sect. 15.2. A naive estimation algorithm is proposed in Sect. 15.3. Then, we proceed to the design of an iterative maximum likelihood estimator in Sect. 15.4. Its statistical properties are studied as well. In Sect. 15.6, a recursive estimator is proposed based on the expectation maximization (EM) algorithm. Simulation results are included in Sect. 15.7. Finally, Some concluding remarks are drawn in Sect. 15.8.

15.1 Innovation Based Scheduler

Consider a linear system as follows:

$$y_k = \mathbf{h}_k^T \theta + v_k, k = 1, 2, \ldots \tag{15.1}$$

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3_15

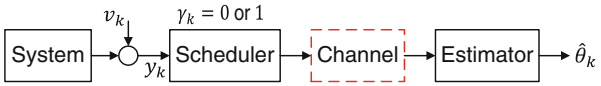

Fig. 15.1 Estimation framework

where $\theta \in \mathbb{R}^p$ is a vector of unknown parameters to be estimated, $\mathbf{h}_k \in \mathbb{R}^p$ is a vector of known regressors, and v_k is an independent, identically distributed, zero-mean Gaussian noise with variance σ^2.

We are concerned with an estimation framework, similar to that in Chap. 14, consisting of a scheduler, a remote estimator and a wireless communication channel, see Fig. 15.1 for an illustration. Due to limited communication resources between the estimator and the sensor, the scheduler is deployed for reducing the number of measurement transmission. For example, let $\gamma_k = 1$ indicate that the sensor is triggered to communicate with the estimator at time k and the packet containing the information of y_k is perfectly delivered to the estimator while $\gamma_k = 0$ means that there is no communication between the sensor and the estimator. That is, the scheduler will decide whether y_k is to be transmitted or not. A natural question is how to design the scheduling variable γ_k to achieve a good estimation performance. We envision that it might be possible to access some priori information of y_k and construct its "predictor" $\widehat{y}_k$ on both sides of the channel. Then, it is reasonable to focus on an "innovation" based scheduling strategy taking a generic form as follows:

$$\gamma_k = \begin{cases} 0, & \text{if } |\frac{y_k - \widehat{y}_k}{\sigma}| \leq \delta; \\ 1, & \text{otherwise,} \end{cases} \tag{15.2}$$

where $\widehat{y}_k$ and δ are to be designed. To some extent, $\widehat{y}_k$ may be treated as a predictor of y_k, and σ is a normalization factor. Loosely speaking, the difference, i.e., $\widetilde{y}_k := y_k - \widehat{y}_k$, is thus called innovation. Intuitively, if $\widetilde{y}_k$ is small, it indicates that $\widehat{y}_k$ is a good approximation of y_k. For this case, it is expected that the absence of y_k for the estimator may not affect the estimation performance significantly. Thus, we perceive such a measurement as "non-useful" one, which will not be transmitted to the estimator for saving communication and energy resources of the sensor.

Here we do not consider other communication effects such as packet losses, transmission delay, data quantization, channel noise and etc. This implies that if there is no measurement sent from the sensor at time k, the estimator does not receive anything from the channel, and understands that y_k is censored. Therefore, the information accessed by the estimator at time k is given by

$$z_k = \{\gamma_k y_k, \gamma_k\}. \tag{15.3}$$

While for noisy channels, the binary valued γ_k can also be available to the estimator by sending merely one bit message from the sensor node.

Remark 15.1 It is worthy mentioning that the effect of channel medium between the sensor and the estimator has not been considered here. Nonetheless, consider the following channel model

$$y_k^{out} = g_k y_k^{in} + w_k,$$

where y_k^{in} and y_k^{out} denote the channel input and output respectively. $g_k \in \mathbb{R}$ is the channel gain and w_k is the additive white Gaussian noise, which is independent of the measurement noise v_k.

In light of (15.1), we further obtain that

$$y_k^{out} = g_k \mathbf{h}_k^T \theta + g_k v_k + w_k.$$

Let $\widetilde{\mathbf{h}}_k = g_k \mathbf{h}_k$ and $\widetilde{v}_k = g_k v_k + w_k$, it follows that

$$y_k^{out} = \widetilde{\mathbf{h}}_k^T \theta + \widetilde{v}_k.$$

The information received by the estimator at time k is then given by $\widetilde{z}_k = \{\gamma_k y_k^{out}, \gamma_k\}$. Thus, we obtain a model similar to the case without incoperating the channel medium effect. This essentially implies that the analysis in the sequel can be generalized to the above channel model. □

Given an arbitrary scheduler in the form of (15.2), the *scheduled transmission rate* at time K in the average sense is defined by

$$R_K = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\gamma_k].$$

Intuitively, R_K characterizes the frequency of measurement transmission in the average sense. The goal of this chapter is to asymptotically quantify the effect of the scheduler with the form (15.2) on the estimation performance as K tends to infinity. The main challenge to derive a good estimator under scheduled measurements lies in the non-linearity of the scheduler.

15.2 Maximum Likelihood Estimation

15.2.1 ML Estimator

Given a sequence of $\widehat{Y}_K := \{\widehat{y}_1, \ldots, \widehat{y}_K\}$, the joint probability density function (pdf) of $Z_K := \{z_1, \ldots, z_K\}$ is given by

$$\prod_{k=1}^{K} p_\theta(z_k) = \prod_{k=1}^{K} [\mathcal{N}(y_k; \mathbf{h}_k^T \theta, \sigma^2)]^{\gamma_k} [\mathbb{P}\{\gamma_k = 0\}]^{1-\gamma_k}, \tag{15.4}$$

where $\mathcal{N}(y_k; \mathbf{h}_k^T\theta, \sigma^2)$ is the pdf of a Gaussian random variable with mean $\mathbf{h}_k^T\theta$ and variance σ^2.

For notational simplicity, let

$$\tau_k(\theta) = \frac{\mathbf{h}_k^T\theta - \widehat{y}_k}{\sigma}; \tag{15.5}$$

$$z_k^1(\theta) = -\delta - \tau_k(\theta),\ z_k^2(\theta) = \delta - \tau_k(\theta). \tag{15.6}$$

Remark 15.2 To simplify the notation, we shall drop the dependence on the unknown parameter θ when it is obvious from the context in the sequel. For instance, we may simply write z_k^1 instead of $z_k^1(\theta)$. □

It is easy to obtain that the probability of the event $\{\gamma_k = 0\}$ is evaluated by

$$\mathbb{P}\{\gamma_k = 0\} = Q(z_k^1(\theta)) - Q(z_k^2(\theta)), \tag{15.7}$$

where $Q(\cdot)$ is the complementary cumulative distribution function of a standard Gaussian random variable, i.e.,

$$Q(x) = \int_x^\infty \phi(t)dt \text{ and } \phi(t) = 1/\sqrt{2\pi}\exp(-t^2/2).$$

Then, the log-likelihood function is computed as

$$\ell_K(\theta) = \sum_{k=1}^K \log p_\theta(z_k) \tag{15.8}$$

$$= \sum_{k=1}^K \{-\frac{\gamma_k}{2}\log(2\pi\sigma^2) - \frac{\gamma_k}{2\sigma^2}(y_k - \mathbf{h}_k^T\theta)^2$$

$$+ (1-\gamma_k)\log(Q(z_k^1(\theta)) - Q(z_k^2(\theta)))\}, \tag{15.9}$$

from which the gradient and the Hessian of the log-likelihood function are obtained as

$$\mathbf{g}_K(\theta) := \frac{\partial \ell_K(\theta)}{\partial\theta} = \sum_{k=1}^K \alpha_k(\theta)\mathbf{h}_k, \tag{15.10}$$

$$\mathbf{H}_K(\theta) := \frac{\partial^2 \ell_K(\theta)}{\partial\theta\partial\theta^T} = \sum_{k=1}^K \beta_k(\theta)\mathbf{h}_k\mathbf{h}_k^T, \tag{15.11}$$

where $\alpha_k(\theta)$ and $\beta_k(\theta)$ are expressed by

$$\alpha_k(\theta) = \frac{\gamma_k(y_k - \mathbf{h}_k^T\theta)}{\sigma^2} + \frac{1-\gamma_k}{\sigma}\frac{\phi(z_k^1) - \phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)},$$

$$\beta_k(\theta) = -\frac{\gamma_k}{\sigma^2} - \frac{1-\gamma_k}{\sigma^2}\left\{\left(\frac{\phi(z_k^1) - \phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)}\right)^2 - \frac{z_k^1\phi(z_k^1) - z_k^2\phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)}\right\}. \quad (15.12)$$

For the ease of notation, we further define

$$s_k(\theta) = \frac{\phi(z_k^1) - \phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)} = \int_{z_k^1}^{z_k^2} x\widetilde{\phi}(x)dx \in [z_k^1, z_k^2],$$

where $\widetilde{\phi}(\cdot)$ is the pdf of a truncated Gaussian random variable with support $[z_k^1, z_k^2]$. We have the following property of the log-likelihood function.

Lemma 15.1 *The log-likelihood function $\ell_K(\theta)$ is a concave function.*

Proof If $s_k \leq 0$, then

$$\left(\frac{\phi(z_k^1) - \phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)}\right)^2 - \frac{z_k^1\phi(z_k^1) - z_k^2\phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)} \geq s_k(s_k - z_k^2) \geq 0. \quad (15.13)$$

Similarly, if $s_k > 0$, then it follows that

$$\left(\frac{\phi(z_k^1) - \phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)}\right)^2 - \frac{z_k^1\phi(z_k^1) - z_k^2\phi(z_k^2)}{Q(z_k^1) - Q(z_k^2)} \geq s_k(s_k - z_k^1) \geq 0. \quad (15.14)$$

The above implies that $\beta_k(\theta) \leq 0$. Thus, $\mathbf{H}_K(\theta) \leq 0$, which completes the proof. □

In light of Lemma 15.1, the maximum likelihood estimator (MLE) is uniquely obtained as

$$\widehat{\theta}_K^{ML} = \arg\max_{\theta\in\mathbb{R}^p} \ell_K(\theta). \quad (15.5)$$

Remark 15.3 Different from the quantized parameter estimation in [1], the maximum likelihood estimator (MLE) can be solved explicitly if a constant quantizer threshold is applied. Since τ_k is a function of θ, it is evident that there is no closed form for $\widehat{\theta}_K^{ML}$ by solving $\mathbf{g}_K(\theta) = 0$. Nonetheless, $\ell_K(\theta)$ is a concave function in θ. This implies that the gradient based algorithm can be adopted to numerically find the MLE, see the Newton method in Algorithm 15.1. □

Algorithm 15.1: Maximum Likelihood Estimation

Initialization:

$$\widehat{\theta}_K = (\sum_{k=1}^{K} \gamma_k \mathbf{h}_k \mathbf{h}_k^T)^+ \sum_{k=1}^{K} \gamma_k y_k \mathbf{h}_k, \tag{15.16}$$

where the superscript $^+$ denotes the Moore-Penrose pseudoinverse [2].

Repeat

Calculate $\mathbf{g}_K(\widehat{\theta}_K)$ and $\mathbf{H}_K(\widehat{\theta}_K)$;

Find μ by the backtracking line search algorithm [3];

Update: $\widehat{\theta}_K \leftarrow \widehat{\theta}_K + \mu[\mathbf{H}_K(\widehat{\theta}_K)]^+ \mathbf{g}_K(\widehat{\theta}_K)$.

Until $\|\mathbf{g}_K(\widehat{\theta}_K)\| \leq \varepsilon$, where $\varepsilon > 0$ and let $\widehat{\theta}_K^{ML} \leftarrow \widehat{\theta}_K$.

15.2.2 Estimation Performance

Obviously, it follows from the gradient of the log-likelihood function that the fisher information matrix [4] is computed as

$$\begin{aligned}\mathbf{I}_K(\theta) &:= \mathbb{E}[\mathbf{g}_K(\theta)\mathbf{g}_K(\theta)^T] \\ &= \sum_{k=1}^{K} \mathbb{E}[\alpha_k(\theta)^2]\mathbf{h}_k\mathbf{h}_k^T + \sum_{i\neq j} \mathbb{E}[\alpha_i(\theta)\alpha_j(\theta)]\mathbf{h}_i\mathbf{h}_j^T.\end{aligned}$$

Since y_k is an independent process, it follows that

$$\mathbb{E}[\alpha_i(\theta)\alpha_j(\theta)] = \mathbb{E}[\alpha_i(\theta)]\mathbb{E}[\alpha_j(\theta)]$$

for all $i \neq j$. We further obtain that

$$\begin{aligned}\sigma \cdot \mathbb{E}[\alpha_k(\theta)] &= \int_{\{\gamma_k=1\}} x\phi(x)dx + (\phi(z_k^1) - \phi(z_k^2)) \\ &= \int_{\{\gamma_k=1\}} x\phi(x)dx + \int_{\{\gamma_k=0\}} x\phi(x)dx \\ &= 0.\end{aligned} \tag{15.17}$$

Hence, the cross term of $\mathbf{I}_K(\theta)$ disappears. The square term of $\mathbf{I}_K(\theta)$ is evaluated by

$$\begin{aligned}\eta_k(\theta) &:= \mathbb{E}[\alpha_k(\theta)^2] \\ &= \frac{\mathbb{E}[\gamma_k(y_k - \mathbf{h}_k^T\theta)^2]}{\sigma^4} + \frac{1}{\sigma^2}\frac{(\phi(z_k^1) - \phi(z_k^2))^2}{Q(z_k^1) - Q(z_k^2)}.\end{aligned}$$

Note that the above can also be rewritten as

$$\eta_k(\theta) = \frac{1}{\sigma^2}(1 - \int_{z_k^1}^{z_k^2} (x - s_k)^2 \phi(x) dx) \leq \frac{1}{\sigma^2}. \tag{15.18}$$

Since

$$\begin{aligned} \int_{z_k^1}^{z_k^2} (x - s_k)^2 \phi(x) dx &= \int_{z_k^1}^{z_k^2} x^2 \phi(x) dx - \frac{(\phi(z_k^1) - \phi(z_k^2))^2}{Q(z_k^1) - Q(z_k^2)} \\ &\leq \int_{z_k^1}^{z_k^2} x^2 \phi(x) dx \leq \int_{-\delta}^{\delta} x^2 \phi(x) dx, \end{aligned}$$

one readily derives that

$$\eta_k(\theta) \geq \frac{1}{\sigma^2}(1 - \int_{-\delta}^{\delta} x^2 \phi(x) dx) > 0, \quad \forall \delta < \infty. \tag{15.19}$$

Lemma 15.2 *For any unbiased estimator $\widehat{\theta}_K$ based on the scheduled measurements Z_K, then*

$$\mathbb{E}[(\theta - \widehat{\theta}_K)(\theta - \widehat{\theta}_K)^T] - [\mathbf{I}_K(\theta)]^+ \geq 0, \tag{15.20}$$

where $A \succeq B$ means that $A - B$ is a positive semi-definite matrix.

Proof It follows from the property of Cramér-Rao lower bound (CRLB) [4]. □

15.2.3 Optimal Scheduler

By (15.7), the scheduled transmission rate of the scheduler with the form (15.2) is computed as

$$R_K = 1 - \frac{1}{K} \sum_{k=1}^{K} [Q(-\delta - \tau_k) - Q(\delta - \tau_k)]. \tag{15.21}$$

Then, the optimal scheduler of the form (15.2) under a transmission rate constraint is defined in the following way.

Definition 15.1 Under a scheduled transmission rate constraint, i.e. $R_K \leq \gamma$, the optimal scheduling parameters $\{\widehat{y}_k, \delta\}$ are solved by

$$\min_{\{\widehat{y}_k,\delta\}} tr([\mathbf{I}_K(\theta)]^{-1}), \quad \text{s.t.} \quad R_K \leq \gamma, \tag{15.22}$$

where tr(A) returns the summation of all diagonal elements of A.

By (15.21), it is evident that it is very difficult to characterize the feasible set for $\{\widehat{y}_k, \delta\}$ such that $R_K \leq \gamma$. Thus, the optimization problem (15.22) is generally intractable. In what follows, a suboptimal scheduler is proposed and its estimation performance is evaluated.

Let $\widehat{y}_k = \mathbf{h}_k^T\theta$ and $\delta = Q^{-1}(\gamma/2)$, where $Q^{-1}(\cdot)$ is the inverse function of $Q(\cdot)$. For this scheduler, one can easily check that $R_K = \gamma$, which obviously satisfies the rate constraint in (15.22), and $\tau_k = 0$. This implies that $s_k = 0$ and $z_k^1 = -z_k^2 = -\delta$. Then, it follows from (15.18) that

$$\eta_k(\theta) = \frac{1}{\sigma^2}(1 - \int_{-\delta}^{\delta} x^2\phi(x)dx). \tag{15.23}$$

To evaluate the performance of this scheduler, define

$$\chi(\delta) = 1 - \int_{-\delta}^{\delta} x^2\phi(x)dx \in [0,\ 1]. \tag{15.24}$$

By using the full set of measurements, which means that all measurements are sent to the estimator, it is clear that the LSE, the MLE and the minimum mean square error (MMSE) estimator are all given by

$$\widehat{\theta}_K = \left(\sum_{k=1}^{K} \mathbf{h}_k\mathbf{h}_k^T\right)^{+} \sum_{k=1}^{K} y_k\mathbf{h}_k \tag{15.25}$$

and its associated CRLB can be easily derived as

$$\sigma^2(\sum_{k=1}^{K} \mathbf{h}_k\mathbf{h}_k^T)^{+}.$$

The loss of estimation performance due to the use of this scheduler is thus characterized by

$$[\mathbf{I}_K(\theta)]^{+} - \sigma^2(\sum_{k=1}^{K} \mathbf{h}_k\mathbf{h}_k^T)^{+} = \frac{1-\chi(\delta)}{\chi(\delta)}(\sum_{k=1}^{K} \mathbf{h}_k\mathbf{h}_k^T/\sigma^2)^{+}.$$

To this end, the factor

$$pd(\delta) := \frac{1 - \chi(\delta)}{\chi(\delta)}$$

quantifies the performance degradation due to the reduction of the number of measurement transmissions by deploying the scheduler with the form of (15.2).

Remark 15.4 From Figs. 15.2 and 15.3, we observe that if $\delta > 2$, the increase of the scheduler threshold δ does not significantly decrease the scheduled transmission rate while the estimation performance degrades fast. It is surprising that when $\delta = 1$, the estimation performance is only slightly worse than the optimal estimator with the full set of measurements since the performance degradation factor $pd(1) = 0.248$. In this case, the scheduled transmission rate can be as low as $R_K = 0.3$. This suggests that the use of the scheduler with the form of (15.2) is an efficient method to reduce the communication cost and maintain a good estimation performance. While maintaining $\tau_k = 0$ corresponds to $\widehat{y}_k = \mathbf{h}_k^T \theta$, it is not implementable as it requires the true value of the unknown parameter θ for the design of the scheduler. To overcome this limitation, an adaptive scheduler is to be designed to asymptotically achieve such a performance level. □

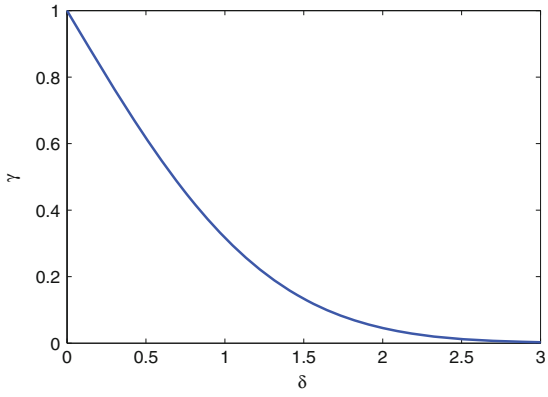

Fig. 15.2 Scheduled transmission rate with respect to the parameter δ

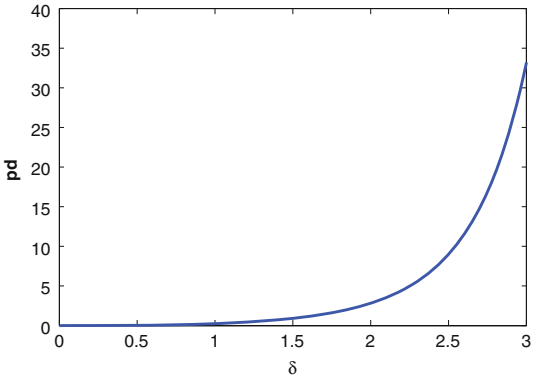

Fig. 15.3 Performance degradation

15.3 Naive Estimation

In the previous section, we illustrate that a good scheduler under any scheduled transmission rate is designed by $\tau_k = 0$ which corresponds to that $\widehat{y}_k = \mathbf{h}_k^T\theta$. In fact, if $\widehat{y}_k = \mathbf{h}_k^T\theta$ is available to the estimator, a very straightforward estimator is constructed as follows:

$$\begin{aligned}\widehat{\theta}_K &= (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}[\gamma_k y_k + (1-\gamma_k)\widehat{y}_k]\mathbf{h}_k \\ &= \theta + (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}\gamma_k v_k\mathbf{h}_k, \end{aligned} \tag{15.26}$$

which is an unbiased estimator since $\mathbb{E}[\gamma_k v_k] = 0$ for all $k \in \{1, \ldots, K\}$. In addition, it is not difficult to compute that

$$\mathbb{E}[(\widehat{\theta}_K - \theta)(\widehat{\theta}_K - \theta)^T] = \sigma^2\chi(\delta)(\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+. \tag{15.27}$$

The above estimator is even better than the standard MLE with the full set of measurements since $\chi(\delta) < 1, \forall\delta < \infty$. The main reason lies in the use of the noise-free "measurement" in the estimator when $\gamma_k = 0$. However, the true parameter θ is actually unknown to the estimator. Thus, the use of $\widehat{y}_k = \mathbf{h}_k^T\theta$ is not implementable for the estimator. Motivated by this observation, a naive estimation algorithm is given by

$$\widehat{\theta}_K = (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}[\gamma_k y_k + (1-\gamma_k)\widehat{y}_k]\mathbf{h}_k, \tag{15.28}$$

where $\widehat{y}_{k+1} = \mathbf{h}_{k+1}\widehat{\theta}_k$ for $k = 0, \ldots, K-1$ and $\widehat{\theta}_0$ is an initial estimate of θ based on the prior information on θ (if any). Albeit simple, the estimation performance of the above estimator is yet to be known. However, the following result is readily established.

Lemma 15.3 *If $\mathbb{E}[\widehat{\theta}_0] = \theta$, then the naive estimator is unbiased, i.e., $\mathbb{E}[\widehat{\theta}_K] = \theta$ for all $K \in \mathbb{N}$.*

Proof Denote $\widetilde{\theta}_k = \widehat{\theta}_k - \theta$, the naive estimator is rewritten as

$$\widehat{\theta}_K = \theta + (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}\gamma_k v_k\mathbf{h}_k + (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}(1-\gamma_k)\mathbf{h}_k\mathbf{h}_k^T\widetilde{\theta}_{k-1}.$$

The rest of the proof is obvious by the arguments of mathematical induction. □

Additionally, we have the following result by recalling Theorem 1 in [5].

Lemma 15.4 *The mean square estimation error of the naive estimate is uniformly bounded, i.e.,*

$$\sup_{K\in\mathbb{N}} \mathbb{E}[\|\widehat{\theta}_K - \theta\|^2] < \infty.$$

Unlike [5], it is meaningful to establish the stability of the state estimator of an unstable dynamical system. Here we are focusing on the estimation of static vector parameters, which is of little importance to show the stability of the estimation error. We are more concerned with the asymptotic unbiasedness and achievability of the CRLB of an estimator.

15.4 Iterative ML Estimation

In this section, we analyze the statistical properties of $\widehat{\theta}_K^{ML}$ when the number K of samples tends to infinity. We provide the conditions for the strong consistency and asymptotic normality of the MLE under some mild conditions. Before proceeding, we introduce the concept of *persistent excitation* [6].

Definition 15.2 The sequence of regressors $\{\mathbf{h}_k\}$ is said to be *persistently exciting* if there exists a $\varsigma > 0$ such that

$$\liminf_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbf{h}_k\mathbf{h}_k^T \succeq \varsigma I. \tag{15.29}$$

One of our main results is formally stated as follows.

Theorem 15.1 *Let $\Theta \subset \mathbb{R}^p$ be a compact set containing the true parameter vector θ as an interior point. Suppose that the sequence of regressors $\{\mathbf{h}_k\}$ is* persistently exciting *and $\sup_{k\in\mathbb{N}} \|\mathbf{h}_k\|_\infty < \infty$. Then, it follows that*

(a) $\widehat{\theta}_K^{ML} \xrightarrow{a.s.} \theta$ *as* $K \to \infty$, *where* $\xrightarrow{a.s.}$ *denotes convergence with probability one under* $\mathbb{P}$.

(b) *As K is sufficiently large, it holds that*

$$\sqrt{K}\cdot(\widehat{\theta}_K^{ML} - \theta) \xrightarrow{in\ dist.} \mathcal{N}(0, C_\eta^{-1}), \tag{15.30}$$

where $\xrightarrow{in\ dist.}$ *means convergence in distribution. C_η is nonsingular and given by*

$$C_\eta = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \eta_k(\theta)\mathbf{h}_k\mathbf{h}_k^T. \tag{15.31}$$

The requirement of persistent excitation of the regressors is essential. Note that it is also necessary for the asymptotic convergence of the standard MLE which is based on the full set of measurements.

Remark 15.5 Theorem 15.1 illustrates that under any scheduled transmission rate, the strong consistency is always guaranteed by using a scheduler with the form of (15.2). But a scheduled transmission will inevitably degrade the estimation performance, which is quantified by C_η. In general, a small scheduled transmission rate will correspond to a large scheduling threshold δ. Actually, one can easily verify that

$$\lim_{R_K \to 0} \delta = \infty.$$

Together with (15.19), it follows that

$$\lim_{R_K \to 0} \eta_k(\theta) = 0.$$

Under this extreme case, C_η^{-1} is arbitrarily large, which in turn implies a poor estimation performance.

Corollary 15.1 *If $\mathbf{h}_k$ is a realization of a wide-sense stationary ergodic random process with uniformly bounded ∞-th moment, i.e.,*

$$\sup_{k \in \mathbb{N}} \mathbb{E}[\|\mathbf{h}_k\|_\infty] < \infty$$

and $\widehat{y}_k$ is designed as a time-invariant function of $\mathbf{h}_k$, then it holds that

$$C_\eta = \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \eta_k(\theta) \mathbf{h}_k \mathbf{h}_k^T. \tag{15.32}$$

Proof Since $\widehat{y}_k$ is a time-invariant function of $\mathbf{h}_k$, so is $\eta_k(\theta)$. This implies that the process of $\eta_k(\theta)\mathbf{h}_k\mathbf{h}_k^T$ is a wide-sense stationary ergodic random process [7]. Note that $\eta_k(\theta) \leq 1/\sigma^2$, it is easy to show that

$$\sup_{k \in \mathbb{N}} \mathbb{E}[\|\eta_k(\theta)\mathbf{h}_k\mathbf{h}_k^T\|_\infty] < \infty.$$

By invoking the Birkhoff's ergodic theorem [7], the result follows. □

15.4.1 Adaptive Scheduler

Next, we design an adaptive scheduler to satisfy a given scheduled transmission rate constraint and the effect of the scheduler on the estimation performance is asymptotically quantified.

Theorem 15.2 *Given any scheduled transmission rate $\gamma \in (0, 1]$, consider an adaptive scheduler with $\delta = Q^{-1}(\gamma/2)$ and $\widehat{y}_k = \mathbf{h}_k^T \widehat{\theta}_{k-1}^{ML}$.*

Let $\Theta \subset \mathbb{R}^p$ be a compact set containing the true parameter vector θ as an interior point. If the sequence of regressors $\{\mathbf{h}_k\}$ is persistently exciting *with $\sup_{k\in\mathbb{N}} \|\mathbf{h}_k\|_\infty < \infty$, it holds that*

(a) $\widehat{\theta}_K^{ML} \xrightarrow{a.s.} \theta$ *as* $K \to \infty$.
(b) $\lim_{K\to\infty} R_K = \gamma$;
(c) *As K is sufficiently large, it follows that*

$$\sqrt{K} \cdot (\widehat{\theta}_K^{ML} - \theta) \xrightarrow{in\ dist.} \mathcal{N}(0, C_\delta^{-1}), \tag{15.33}$$

where C_δ is given by

$$C_\delta = \frac{\chi(\delta)}{\sigma^2} \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^T. \tag{15.34}$$

Proof (a) It follows from Theorem 15.1. (b) By the first part, it is clear that $\widehat{\theta}_K^{ML}$ converges to θ in probability as well [7]. For a sufficiently small $\varepsilon > 0$, there exists a sufficiently large k_0 such that

$$\mathbb{P}(|\widehat{y}_k - \mathbf{h}_k^T \theta| < \sigma\varepsilon) > 1 - \varepsilon$$

for all $k \geq k_0$. Note that $Q(-\delta - \tau) - Q(\delta - \tau)$ achieves its maximum at $\tau = 0$, we obtain that

$$\begin{aligned} R_K &= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\widehat{y}_k}[\mathbb{E}[\gamma_k | \widehat{y}_k]] \\ &= 1 - \frac{1}{K} \sum_{k=1}^{K} \int_{\mathbb{R}} [Q(-\delta - \tau_k) - Q(\delta - \tau_k)] p_\theta(\widehat{y}_k) d\widehat{y}_k \\ &\geq 1 - \frac{1}{K} \sum_{k=1}^{K} [1 - 2Q(\delta)] = \gamma, \end{aligned} \tag{15.35}$$

where $\mathbb{E}_{\widehat{y}_k}[\cdot]$ is the mathematical operator with respect to the random variable $\widehat{y}_k$ and $p_\theta(\widehat{y}_k)$ is the pdf of $\widehat{y}_k$ with the abuse of notation.

On the other hand, it is easy to verify that

$$Q(-\delta - \tau) - Q(\delta - \tau) \geq Q(-\delta - \varepsilon) - Q(\delta - \varepsilon)$$

for any $|\tau| \leq \varepsilon$. In addition, one readily establishes that

$$\begin{aligned} &Q(-\delta - \varepsilon) - Q(\delta - \varepsilon) - [Q(-\delta) - Q(\delta)] \\ &\quad = \int_{-\delta-\varepsilon}^{-\delta} \phi(x) dx - \int_{\delta-\varepsilon}^{\delta} \phi(x) dx \geq -\varepsilon\phi(\delta - \varepsilon) \geq -\frac{\varepsilon}{\sqrt{2\pi}}. \end{aligned} \tag{15.36}$$

For $k \geq k_0$, it follows that

$$\begin{aligned}
&\int_{\mathbb{R}} [Q(-\delta - \tau_k) - Q(\delta - \tau_k)] p_\theta(\widehat{y}_k) d\widehat{y}_k \\
&\quad \geq \int_{|\tau_k|<\varepsilon} [Q(-\delta - \tau_k) - Q(\delta - \tau_k)] p_\theta(\widehat{y}_k) d\widehat{y}_k \\
&\quad \geq (1-\varepsilon)[Q(-\delta - \varepsilon) - Q(\delta - \varepsilon)] \\
&\quad \geq (1-\varepsilon)[Q(-\delta) - Q(\delta)] - \frac{\varepsilon(1-\varepsilon)}{\sqrt{2\pi}}. \qquad (15.37)
\end{aligned}$$

For $K > k_0$, we get an upper bound of R_K, i.e.,

$$\begin{aligned}
R_K &\leq 1 - \frac{(1-\varepsilon)(K-k_0)}{K}[Q(-\delta) - Q(\delta)] + \frac{\varepsilon(1-\varepsilon)}{\sqrt{2\pi}} \\
&\to 1 - (1-\varepsilon)[1 - 2Q(\delta)] + \frac{\varepsilon(1-\varepsilon)}{\sqrt{2\pi}} \text{ as } K \to \infty.
\end{aligned}$$

Letting ε go to zero, we obtain that

$$\gamma \leq \liminf_{K\to\infty} R_K \leq \limsup_{K\to\infty} R_K \leq 2Q(\delta) = \gamma. \qquad (15.38)$$

This implies that $\lim_{K\to\infty} R_K = \gamma$.
(c) The asymptotic normality and unbiasedness follows from Theorem 15.1. Only the estimation error covariance matrix is to be evaluated. Toward this end, the joint pdf of Z_K is written by

$$\begin{aligned}
p_\theta(Z_K) &= p_\theta(z_1) \prod_{k=2}^{K} p_\theta(z_k | z_{k-1}, \ldots, z_1) \\
&= p_\theta(z_1) \prod_{k=2}^{K} p_\theta(z_k | \widehat{y}_k). \qquad (15.39)
\end{aligned}$$

Then, the fisher information matrix is computed as

$$I_K(\theta) = \frac{1}{\sigma^2} \sum_{k=1}^{K} \bar{\eta}_k(\theta) \mathbf{h}_k \mathbf{h}_k^T, \qquad (15.40)$$

where

$$\bar{\eta}_k(\theta) = 1 - \mathbb{E}_{\tau_k} \left[\int_{z_k^1}^{z_k^2} (x - s_k)^2 \phi(x) dx \right]. \qquad (15.41)$$

Next, we show that $\lim_{k\to\infty} \bar{\eta}_k(\theta) = \chi(\delta)$. To this purpose, we note that

$$\int_{z_k^1}^{z_k^2} (x - s_k)^2 \phi(x) dx = \int_{z_k^1}^{z_k^2} x^2 \phi(x) dx - \frac{(\phi(z_k^1) - \phi(z_k^2))^2}{Q(z_k^1) - Q(z_k^2)}$$
$$\leq \int_{\mathbb{R}} x^2 \phi(x) dx = 1. \tag{15.42}$$

Since $\lim_{k\to\infty} \widehat{\theta}_k^{ML} = \theta$ with probability one, this implies that $\lim_{k\to\infty} \tau_k = 0$ almost surely. Together with the dominated convergence theorem [7], we obtain that

$$\lim_{k\to\infty} \mathbb{E}_{\tau_k} \left[\int_{z_k^1}^{z_k^2} (x - s_k)^2 \phi(x) dx \right] = \int_{-\delta}^{\delta} x^2 \phi(x) dx. \tag{15.43}$$

Thus, it is straightforward that $\lim_{k\to\infty} \bar{\eta}_k(\theta) = \chi(\delta)$, which in turn implies that

$$\lim_{K\to\infty} \frac{I_K(\theta)}{K} = \frac{\chi(\delta)}{\sigma^2} \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^T. \tag{15.44}$$

By the asymptotic efficiency of the MLE [4], it finally holds that

$$C_\delta = \lim_{K\to\infty} \frac{I_K(\theta)}{K}. \tag{15.45}$$

The proof is completed. □

15.5 Proof of Theorem 15.1

Lemma 15.5 *The gradient of the log-likelihood function $\ell_K(\theta)$ is expressed by*

$$\mathbf{g}_K(\theta) = \frac{1}{\sigma^2} \sum_{k=1}^{K} (\bar{y}_k(\theta) - \mathbf{h}_k^T \theta) \mathbf{h}_k. \tag{15.46}$$

Proof It is straightforward. □

Let $\xi_k(\theta) = \log p_\theta(z_k)$, $\zeta_K(\theta) = \ell_K(\theta)/K = \frac{1}{K} \sum_{k=1}^{K} \xi_k(\theta)$, $\zeta_K'(\theta) = \frac{\partial}{\partial\theta} \zeta_K(\theta)$ and $\zeta_K''(\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T} \zeta_K(\theta)$. Define the following functions by

$$\bar{\zeta}(\theta) = \lim_{K\to\infty} \mathbb{E}[\zeta_K(\theta)],$$
$$\bar{\zeta}''(\theta) = \lim_{K\to\infty} \mathbb{E}[\zeta_K''(\theta)]. \tag{15.47}$$

We also define $\xi_k'(\theta) = \frac{\partial}{\partial\theta}\xi_k(\theta)$ and $\xi_k''(\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\xi_k(\theta)$.

Lemma 15.6 *Under the conditions of Theorem 15.1, the functions $\zeta_K(\theta)$, $\zeta_K'(\theta)$ and $\zeta_K''(\theta)$ are continuous. Moreover, $\zeta_K(\theta) \to \bar{\zeta}_K(\theta)$, $\zeta_K'(\theta) \to 0$, and $\zeta_K''(\theta) \to \bar{\zeta}_K''(\theta)$ with probability one and uniformly on $\theta \in \Theta$ as $K \to \infty$.*

Proof We shall use the Rajchman's strong law of large numbers [8, Theorem5.1.2] to prove the convergence with probability one. Due to the statistical independence of $\xi_k(\theta)$ and the fact that $\sup_k \mathbb{E}[\|\xi_k(\theta)\|^2] < \infty$, it follows that $\zeta_K(\theta) \to \bar{\zeta}_K(\theta)$ with probability one.

Since $\mathbb{E}[(y_k - \bar{y}_k(\theta))^2] \le 2\mathbb{E}[y_k^2] \le 4(\|\mathbf{h}_k^T\theta\|^2 + \mathbb{E}[v_k^2]) \le 4(\|\theta\|^2 \sup_k \|\mathbf{h}_k\|^2 + \sigma^2) < \infty$, it follows from Lemma 15.5 that

$$\begin{aligned}\mathbb{E}[\|\xi_k'(\theta)\|^2] &\le \frac{\|\mathbf{h}_k\|^2}{\sigma^2}\mathbb{E}[\|\bar{y}_k(\theta) - \mathbf{h}_k^T\theta\|^2] \\ &\le \frac{2\|\mathbf{h}_k\|^2}{\sigma^2}(\mathbb{E}[(y_k - \mathbf{h}_k^T\theta)^2] + \mathbb{E}[(y_k - \bar{y}_k(\theta))^2]) < \infty. \end{aligned} \tag{15.48}$$

To verify that $\sup_k \mathbb{E}[\|\xi_k''(\theta)\|^2] < \infty$, it is sufficient to show that $\sup_k \mathbb{E}[\eta_k^2(\theta)] < \infty$, which obviously holds as $\eta_k(\theta) \le 1/\sigma^2$. One can easily test that $\mathbb{E}[\xi_k'(\theta)] = 0$. Thus, $\zeta_K'(\theta) \to 0$ and $\zeta_K''(\theta) \to \bar{\zeta}_K''(\theta)$ with probability one as $K \to \infty$.

Next, we show that the convergence is also uniform with respect to $\theta \in \Theta$. By Theorem 7.17 in [9], it is sufficient to establish the uniform convergence of $\zeta_K''(\theta)$. Toward this, we first verify the strong stochastic equi-continuity of $\zeta_K''(\theta)$. One can verify from (15.12) that $\partial\beta_k(\theta)/\partial\theta$ is continuous with respect to θ. Due to that Θ is a compact set, we obtain that

$$\sup_{k\in\mathbb{N}} \sup_{\phi,\theta\in\Theta} |\beta_k(\theta) - \beta_k(\phi)| \cdot \|\phi - \theta\|^{-1} < \infty, \quad \forall \phi \ne \theta,$$

which in turn implies that

$$\sup_{K\in\mathbb{N}} \sup_{\phi,\theta\in\Theta} \|\zeta_K''(\theta) - \zeta_K''(\phi)\| \cdot \|\phi - \theta\|^{-1} < \infty, \quad \forall \phi \ne \theta. \tag{15.49}$$

By Theorem 21.10 in [10], we know that $\zeta_K''(\theta)$ is strong stochastically equi-continuity. Together with Theorem 21.8 in [10], it follows that $\zeta_K''(\theta)$ converges uniformly with respect to θ in Θ as $K \to \infty$. □

Lemma 15.7 *Under the conditions of Theorem 15.1, the function $\bar{\zeta}(\phi)$ attains its maximum value if and only if $\phi = \theta \in \Theta$.*

Proof By Lemma 15.1, it follows that $\zeta_K(\phi)$ is concave in $\phi \in \mathbb{R}^p$. This implies that $\bar{\zeta}(\phi)$ is concave as well [3]. In addition, we notice that

$$\begin{aligned}\bar{\zeta}(\phi) - \bar{\zeta}(\theta) &= \lim_{K\to\infty} \mathbb{E}_\theta[\zeta_K(\phi) - \zeta_K(\theta)] \\ &= \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_\theta\left[\log \frac{p_\phi(z_k)}{p_\theta(z_k)}\right] \\ &= -\lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} D_{KL}(p_\theta(z_k)|p_\phi(z_k)) \\ &\leq 0, \end{aligned} \tag{15.50}$$

where $D_{KL}(\cdot|\cdot)$ denotes the Kullback Leibler distance [11] between two pdfs. □

Lemma 15.8 *Under the conditions of Theorem 15.1, it holds that*

$$\sqrt{K}\zeta_K'(\theta) \xrightarrow{\text{in dist.}} \mathcal{N}(0, C_\eta^{-1}). \tag{15.51}$$

Proof For any $y \in \mathbb{R}^p$, let $\vartheta_k = y^T \xi_k'(\theta)$. We show that ϑ_k satisfies the Lyaponov condition [7] and then apply the Lindeberg-Feller central limit theorem [7] to establish the result.

Let $\varepsilon > 0$, it follows that

$$\begin{aligned}\mathbb{E}[|\bar{y}_k(\theta) - \mathbf{h}_k^T\theta|^{2+\varepsilon}] &= \mathbb{E}[|\gamma_k v_k + \sigma(1-\gamma_k)s_k(\theta)|^{2+\varepsilon}] \\ &\leq 3^{1+\varepsilon}(\mathbb{E}[|v_k|^{2+\varepsilon}] + \sigma^{2+\varepsilon}\mathbb{E}[|s_k(\theta)|^{2+\varepsilon}]) \leq M < \infty,\end{aligned}$$

where the first inequality follows from the C_r-inequality, see e.g. Lemma 5 of [12] and the existence of a upper bound M is clear. Due to that $\sup_{k\in\mathbb{N}} \|h_k\| < \infty$, there exists a $M_1 > 0$ such that

$$\mathbb{E}[|\vartheta_k|^{2+\varepsilon}] = \mathbb{E}[|\bar{y}_k(\theta) - \mathbf{h}_k^T\theta|^{2+\varepsilon} \cdot (\mathbf{h}_k^T y)^{2+\varepsilon}] \leq M_1.$$

This implies that

$$1/K \cdot \sum_{k=1}^{K} \mathbb{E}[|\vartheta_k|^{2+\varepsilon}] \leq M_1.$$

Let

$$P_K = 1/K \cdot \sum_{k=1}^{K} \mathbb{E}[\vartheta_k^2] = y^T y/K \cdot \sum_{k=1}^{K} \eta_k(\theta)\mathbf{h}_k^T\mathbf{h}_k,$$

which converges to $y^T y \cdot C_\eta > 0$ due the persistent excitation condition of $\mathbf{h}_k$ and $\eta_k(\theta) > 0$ in (15.19). Hence, we finally obtain that

$$\lim_{K\to\infty} P_K^{-1-\varepsilon/2}(1/K \cdot \sum_{k=1}^{K} \mathbb{E}[|\vartheta_k|^{2+\varepsilon}]) = 0. \tag{15.52}$$

By invoking the Lindeberg-Feller central limit theorem, the proof is completed. □

Proof of Theorem 15.1 (a) Based on the above lemmas, we have established that with probability one, $\zeta_K(\theta)$ converges uniformly to $\bar{\zeta}(\theta)$, which attains its maximum at the true parameter θ_0 only. Together with the compactness of Θ, it follows from Property 24.2 in [13] that $\widehat{\theta}_K^{ML} \xrightarrow{\text{a.s.}} \theta$. (b) From Lemma 15.6, $\zeta_K''(\theta)$ converge uniformly to $\bar{\zeta}_K''(\theta)$. From (15.19) and persistent excitind condition, C_η is nonsingular. Together with Lemma 15.8 and θ is an interior point of the compact set Θ, the result of (b) follows from Property 24.16 in [13]. □

15.6 EM-Based Estimation

Till now, we have conducted the asymptotic analysis of the MLE as the number of measurements tends to infinity. Observe that the MLE is obtained by solving a convex optimization, which is computationally demanding and lacks a recursive form. The problem of interest is how to design a recursive estimation algorithm to asymptotically achieve the CRLB. Toward this end, an EM-based estimation algorithm is to be designed, which has been exploited for the quantized identification problem in [14]. Suppose that the output $Y_K := \{y_1, \ldots, y_K\}$ was known, it would be easier to maximize $\log p_\theta(Y_K)$. Since Y_K is unavailable, we replace $\log p_\theta(Z_K)$ by the average of $\log p_\theta(Z_K, Y_K)$ and obtain the following EM method to solve the MLE problem (15.5) via the following iterative procedure.

$$\widehat{\theta}_K^{(t)} = \arg\max_{\theta\in\mathbb{R}^p} Q_K(\theta, \widehat{\theta}_K^{(t-1)}), \quad t \in \mathbb{N}; \tag{15.53}$$

$$Q_K(\theta, \widehat{\theta}) = \int \log p_\theta(Z_K, Y_K) p_{\widehat{\theta}}(Y_K|Z_K) dY_K. \tag{15.54}$$

To obtain an on-line estimator, we compute one iteration at each sample time. Then, it follows from (15.53) that

$$\widehat{\theta}_K^{EM} = \arg\max_{\theta\in\mathbb{R}^p} Q_K(\theta, \widehat{\theta}_{K-1}^{EM}). \tag{15.55}$$

It turns out that, under quite general conditions, the sequence of numbers $\ell_K(\widehat{\theta}_K^{EM})$, $K = 1, 2, \ldots$ is monotonically increasing, and therefore the parameter estimation sequence $\widehat{\theta}_K^{EM}$ converges to a local maximum of the log-likelihood function $\ell_K(\theta)$ [15]. By the concave property in Lemma 15.1, $\widehat{\theta}_K^{EM}$ asymptotically coincides with $\widehat{\theta}_K^{ML}$. Note that the asymptotic analysis of $\widehat{\theta}_K^{ML}$ has been conducted in the previous section. To solve (15.55), we establish the following result.

Lemma 15.9 *The function of $Q(\cdot,\cdot)$ is given by*

$$Q_K(\theta,\widehat{\theta}) = -\frac{1}{2\sigma^2}\sum_{k=1}^{K}(\bar{y}_k(\widehat{\theta}) - \mathbf{h}_k^T\theta)^2 + f_K(\widehat{\theta}), \tag{15.56}$$

where $f_K(\widehat{\theta})$ only depends on $\widehat{\theta}$ and

$$\begin{aligned}\bar{y}_k(\theta) &= \mathbb{E}[y_k|z_k] = \int y_k p_\theta(y_k|z_k)dy_k \\ &= \gamma_k y_k + (1-\gamma_k)(\mathbf{h}_k^T\theta + \sigma s_k(\theta)).\end{aligned} \tag{15.57}$$

Proof It is similar to the proof of Lemma 5 in [16]. Details are omitted. □

Combining the above, the maximization in (15.55) is explicitly solved by

$$\begin{aligned}\widehat{\theta}_K^{EM} &= \arg\min_{\theta\in\mathbb{R}^p}\sum_{k=1}^{K}(\bar{y}_k(\widehat{\theta}_{K-1}^{EM}) - \mathbf{h}_k^T\theta)^2 \\ &= (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}\bar{y}_k(\widehat{\theta}_{K-1}^{EM})\mathbf{h}_k.\end{aligned} \tag{15.58}$$

It is obvious that the above estimate closely resembles the standard MLE with the full set of measurements. In comparison with the naive estimator (15.28), it follows from (15.57) that the EM-based estimator has a correction term which results in an unbiased estimator.

Lemma 15.10 *$\widehat{\theta}_K^{EM}$ is an unbiased estimator, i.e., $\mathbb{E}[\widehat{\theta}_K^{EM}] = \theta$.*

Proof Since $y_k = \mathbf{h}_k^T\theta + v_k$, it follows that

$$\widehat{\theta}_K^{EM} = \theta + (\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^T)^+\sum_{k=1}^{K}(\gamma_k v_k + \sigma(1-\gamma_k)s_k(\widehat{\theta}_{K-1}^{EM})). \tag{15.59}$$

Next, we further have that

$$\begin{aligned}&\mathbb{E}[\gamma_k v_k + \sigma(1-\gamma_k)s_k(\widehat{\theta}_{K-1}^{EM})] \\ &= \mathbb{E}_{\widehat{\theta}_{K-1}^{EM}}[\mathbb{E}[\gamma_k v_k + \sigma(1-\gamma_k)s_k(\theta)|\widehat{\theta}_{K-1}^{EM} = \theta]] \\ &= \sigma^2\mathbb{E}_{\widehat{\theta}_{K-1}^{EM}}[\mathbb{E}[\alpha_k(\theta)|\widehat{\theta}_{K-1}^{EM} = \theta]] \\ &= 0,\end{aligned} \tag{15.60}$$

where the last equality follows from (15.17). Hence, it is trivial that $\mathbb{E}[\widehat{\theta}_{K-1}^{EM}] = \theta$. □

By following the steps in [17], the estimator in (15.58) can be written in a recursive form by

$$\begin{aligned}\widehat{\theta}_{k+1} &= \widehat{\theta}_k + P_k \mathbf{h}_k (\sigma^2 + \mathbf{h}_k^T P_k \mathbf{h}_k)^{-1} (\bar{y}_k(\widehat{\theta}_k) - \mathbf{h}_k^T \widehat{\theta}_k);\\ P_{k+1} &= P_k - P_k \mathbf{h}_k (\sigma^2 + \mathbf{h}_k^T P_k \mathbf{h}_k)^{-1} \mathbf{h}_k^T P_k. \end{aligned} \tag{15.61}$$

Remark 15.6 Without scheduling, the gradient of the log-likelihood function is given by

$$\frac{1}{\sigma^2} \sum_{k=1}^{K} (y_k - \mathbf{h}_k^T \theta) \mathbf{h}_k,$$

from which the MLE can be explicitly obtained. Here we only need to replace y_k by $\bar{y}_k(\theta)$ when y_k is unavailable. Note that $\bar{y}_k(\theta)$ is a nonlinear function of θ. By (15.5), it is generically impossible to find a closed form of the MLE by letting

$$\frac{\partial \ell_K(\theta)}{\partial \theta} = 0.$$

Since

$$\bar{y}_k(\theta) = \arg\min_{x \in \mathbb{R}} \mathbb{E}[(y_k - x)^2 | z_k], \tag{15.62}$$

then $\bar{y}_k(\theta)$ minimizes the mean quadratic approximation error of y_k. By (15.61), the estimator $\widehat{\theta}_k$ is obtained by simply replacing y_k by $\bar{y}_k(\widehat{\theta}_k)$ in the standard LSE algorithm using the full set of measurements. From this perspective, our estimator shares a similar computational complexity as the standard LSE. Moreover, we will prove shortly that its performance comes comparable to the standard LSE, even under a moderate scheduled transmission rate.

Theorem 15.3 *Let $\Theta \subset \mathbb{R}^p$ be a compact set containing the true parameter vector θ as an interior point. Suppose that the sequence of regressors $\{\mathbf{h}_k\}$ is* persistently exciting *with $\sup_{k \in \mathbb{N}} \|\mathbf{h}_k\|_\infty < \infty$. Then, it follows that*

(a) $\widehat{\theta}_K^{EM} \xrightarrow{a.s.} \theta$ *as* $K \to \infty$.
(b) *As K is sufficiently large, it holds that*

$$\sqrt{K} \cdot (\widehat{\theta}_K^{EM} - \theta) \xrightarrow{\text{in dist.}} \mathcal{N}(0, C_\eta^{-1}), \tag{15.63}$$

where C_η is given in (15.31).

Proof It follows from Lemma 15.1 and Theorem 15.1. □

15.6.1 Design of $\widehat{y}_k$

For any sequence of $\widehat{y}_k$, we have established the strong consistency and asymptotic normality of the EM based estimator in the previous section.

In this section, the design of $\widehat{y}_k$ will be discussed. By doing this, we adaptively design $\widehat{y}_k$ to achieve the CRLB with $\eta_k(\theta)$ given in (15.23). The striking feature of the EM based estimation algorithm is due to its simplicity, which allow us to easily implement on-line. By adaptively designing $\widehat{y}_k$ as follows

$$\widehat{y}_k = \mathbf{h}_k^T \widehat{\theta}_{k-1}^{EM}, \tag{15.64}$$

we have the following result.

Theorem 15.4 *Given any scheduled transmission rate $\gamma \in (0, 1]$, consider an adaptive scheduler with $\delta = Q^{-1}(\gamma/2)$ and $\widehat{y}_k = \mathbf{h}_k^T \widehat{\theta}_{k-1}^{EM}$. Let $\Theta \subset \mathbb{R}^p$ be a compact set containing the true parameter vector θ as an interior point. If the sequence of regressors $\{\mathbf{h}_k\}$ is* persistently exciting *with $\sup_{k\in\mathbb{N}} \|h_k\|_\infty < \infty$ and*

$$\Phi_h = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^T,$$

then

1. $\widehat{\theta}_K^{EM} \overset{a.s.}{\to} \theta$ *as* $K \to \infty$.
2. $\lim_{K\to\infty} R_K = \gamma$;
3. *As K is sufficiently large, it holds that*

$$\sqrt{K} \cdot (\widehat{\theta}_K^{EM} - \theta) \overset{in\ dist.}{\longrightarrow} \mathcal{N}(0, C_\delta^{-1}), \tag{15.65}$$

where C_δ is given in (15.34).

Proof It follows from Lemma 15.1 and Theorem 15.2. □

15.7 Numerical Example

Consider a linear system as follows:

$$y_k = \mathbf{h}_k^T \theta + v_k, \tag{15.66}$$

where the true parameter vector $\theta = [0.4, 1]^T$ and v_k is a white Gaussian noise with zero mean and unit variance. The regressor $\{\mathbf{h}_k\}$ is generated from a white Gaussian vector, i.e., $\mathbf{h}_k \sim \mathcal{N}(0, \rho^2 * I_2)$. The signal to noise ratio (SNR) is computed as

$$10 \cdot \log_{10}(\mathbb{E}[y_k^2]/E[v_k^2]) = 10 \cdot \log_{10}(\rho^2 \theta^T \theta + 1) \text{ dB}.$$

We are concerned with an adaptive scheduler to asymptotically achieve a scheduled transmission rate $\gamma = 0.6$ as the number of measurement tends to infinity, where the scheduler threshold is solved by $\delta = 0.525$. Using this scheduler threshold and $\rho = 2$, we also apply a scheduler by randomly selecting $\widehat{y}_k$ from a Gaussian distribution $\mathcal{N}(0, 3^2)$. From Fig. 15.4, it is shown that both the EM based algorithms asymptotically converge to the true parameter values as the number of measurement tends to infinity, which is consistent with Theorem 15.3.

To evaluate the scheduled transmission rate of the above two schedulers, we use the Monte Carlo method with 5,000 samples. From Fig. 15.5, it is shown that the scheduled transmission rate of the adaptive scheduler designed in Theorem 15.4 asymptotically converges to the prescribed rate $\gamma = 0.6$. In Fig. 15.6, the MSE with the adaptive scheduler is the estimated

$$K \cdot \mathrm{tr}(\mathbb{E}[(\widehat{\theta}_K^{EM} - \theta)(\widehat{\theta}_K^{EM} - \theta)^T])$$

by using the Monte Carlo method with 5,000 samples. Note that we have converted its unit into dB. The MSE without scheduler corresponds to that of the standard LSE. One can observe that under the scheduled transmission rate $\gamma = 0.6$, the estimation

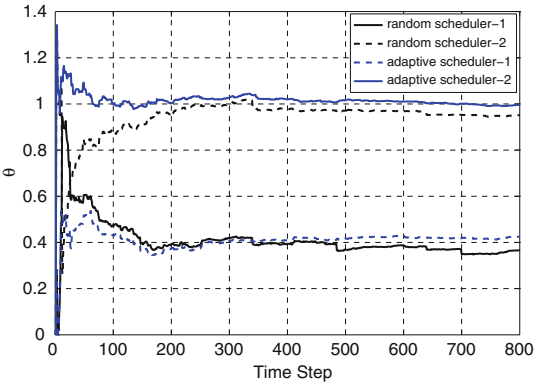

Fig. 15.4 Asymptotic convergence

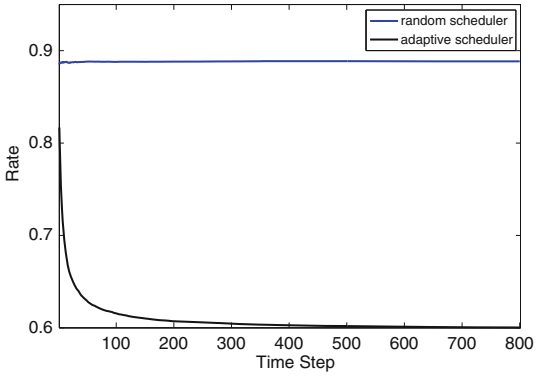

Fig. 15.5 Scheduled transmission rate

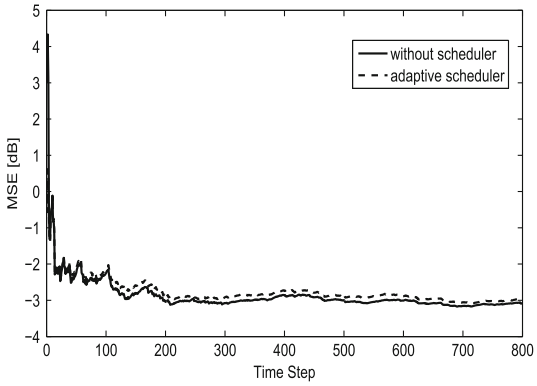

Fig. 15.6 Mean square estimation error

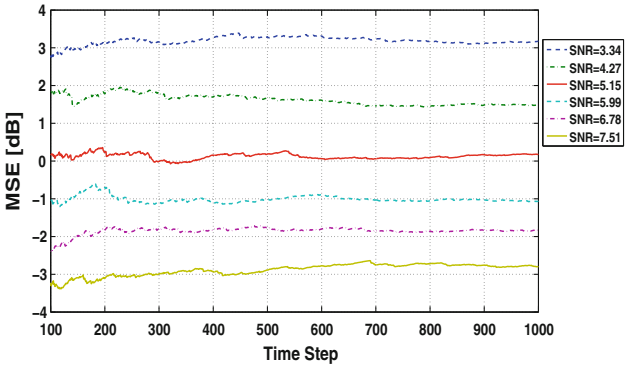

Fig. 15.7 Mean square estimation error

performance of the EM based algorithm comes close to the standard LSE, which is based on the full set of measurements.

Next, we check how the SNR affects the asymptotic estimation performance of the adaptive scheduler under the scheduled transmission rate $\gamma = 0.6$ by altering the variance of the regressor $\{\mathbf{h}_k\}$ from $\rho = 1$ to $\rho = 2$ with step size 0.2. Figure 15.7 confirms that the higher the SNR, the better the estimation performance. This is consistent with Theorem 15.2.

15.8 Summary

Motivated by the limited energy and communication capabilities of sensor node, we have proposed a scheduler for the parameter estimation problem of a linear system to reduce the number of measurement communications. The challenge in this estimation framework lies in the nonlinearity of the scheduler. To quantify the effect of the

scheduler, we have conducted a rigorous stochastic analysis in this work and provided conditions on the strong consistency and asymptotic normality of the estimator with respect to the number of sensor measurements. The proposed estimation algorithm bears a close structure of the standard MLE with the estimation performance comparable to the standard MLE even under a moderate scheduled transmission rate. This nice property is expected to be of great importance in the resource limited networks. Simulation has been included to support our theoretic results.

References

1. A. Ribeiro, G. Giannakis, Bandwidth-constrained distributed estimation for wireless sensor networks-part II: unknown pdf. IEEE Trans. Signal Process. **54**(7), 2784–2796 (2006)
2. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)
3. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
4. S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice-Hall, Englewood Cliffs, 1993)
5. G. Battistelli, A. Benavoli, L. Chisci, Data-driven communication for state estimation with sensor networks. Automatica **48**(5), 926–935 (2012)
6. M. Green, J. Moore, Persistence of excitation in linear systems. Syst. Control Lett. **7**(5), 351–360 (1986)
7. R. Ash, C. Doléans-Dade, *Probability and Measure Theory* (Academic Press, San Diego, 2000)
8. K. Chung, *A Course in Probability Theory* (Academic Press, Burlington, 2001)
9. W. Rudin, *Principles of Mathematical Analysis* (McGraw-Hill, New York, 1976)
10. J. Davidson, *Stochastic Limit Theory: An Introduction for Econometricians* (Oxford University Press, Oxford, 1994)
11. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006)
12. K. You, L. Xie, Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses. IEEE Trans. Autom. Control **56**(4), 772–785 (2011)
13. C. Gourieroux, A. Monfort, *Statistics and Econometric Models* (Cambridge University Press, Cambridge, 1995)
14. D. Marelli, K. You, M. Fu, Identification of ARMA models using intermittent and quantized output observations. Automatica **49**(2), 360–369 (2013)
15. A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the Em algorithm. J. R. Stat. Soc. B **39**(1), 1–38 (1977)
16. B. Godoy, G. Goodwin, J. Agüero, D. Marelli, T. Wigren, On identification of FIR systems having quantized output data. Automatica **47**(9), 1905–1915 (2011)
17. L. Ljung, *System Identification: Theory for the User* (Prentice Hall PTR, Upper Saddle River, 1999)

Appendix A
On Matrices

This appendix collects several results on matrix analysis that are repeatedly used throughout the book. The proofs are omitted and can be found in, e.g., [1, 2].

Lemma A.1 (Schur's complement [1]) *Suppose that a Hermitian matrix is partitioned as*

$$\begin{bmatrix} A & B \\ B^H & C \end{bmatrix},$$

where A and C are square. This matrix is positive definite if and only if A is positive definite and $C > B^H A^{-1} B$. *Furthermore, this condition is equivalent to having* $\rho(B^H A^{-1} B C^{-1}) < 1$.

Lemma A.2 (Hadamard's inequality [1]) *If* $A \in \mathbb{C}^{n\times n}$ *is positive semidefinite, then*

$$\det(A) \leq \prod_{i=1}^{n} [A]_{ii}.$$

Furthermore, when A is positive definite, then equality holds if and only if A is diagonal.

Lemma A.3 (Matrix inversion lemma [1]) *Suppose* $Y = X + MQN$, *and* X, Y, Q *are nonsingular. It holds that*

$$Y^{-1} = X^{-1} - X^{-1}M(Q^{-1} + NX^{-1}M)^{-1}NX^{-1}.$$

Lemma A.4 (Properties of trace, determinant and Kronecker product [1, 2]) *For* X, Y, M, N *with compatible dimensions, the following properties hold.*

(1) $\mathrm{tr}(M+N) = \mathrm{tr}(M) + \mathrm{tr}(N)$, $\mathrm{tr}(MN) = \mathrm{tr}(NM)$, $\det(MN) = \det(M)\det(N)$.
(2) *If* $M \geq 0, N > 0$, *then*

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3

$$\left(\min_i \lambda_i(N)\right) \mathrm{tr}(M) \leq \mathrm{tr}(MN) \leq \left(\max_i \lambda_i(N)\right) \mathrm{tr}(M),$$

where $\lambda_i(N)$ *is the ith eigenvalue of* N.

(3) $\det(I + MN) = \det(I + NM)$.

(4) *The matrix equation* $Y = MXN$ *can be written into* $\mathrm{vec}(Y) = (N^T \otimes M)\mathrm{vec}(X)$.

Lemma A.5 (Property of Hadamard product [2]) *If* $A \geq B \geq 0$ *and* $C \geq D \geq 0$, *then* $A \odot C \geq B \odot D$.

Lemma A.6 (Shur product theorem [2]) *If* A, B *are positive semidefinite, then so is* $A \odot B$. *If, in addition,* B *is positive definite and* A *has no diagonal entry equals to* 0, *then* $A \odot B$ *is positive definite. In particular, if* A, B *are positive definite, then so is* $A \odot B$.

References

1. R. Horn, C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge 1985)
2. R. Horn, C. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991)

Index

K. You et al., *Analysis and Design of Networked Control Systems*,
Communications and Control Engineering, DOI 10.1007/978-1-4471-6615-3