Control and estimation with limited data rate and packet loss

You, Keyou

2011

You, K. (2011). Control and estimation with limited data rate and packet loss. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/46275>

<https://doi.org/10.32657/10356/46275>

Control and Estimation with Limited Data Rate and Packet Loss

You Keyou

School of Electrical & Electronic Engineering

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

2011

Statement of Originality

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

.....

Date

.....

You Keyou

Acknowledgements

I wish to express my heartiest gratitude to my supervisor, Prof. Xie Lihua, for his unconditional support, professional guidance and encouragement during the PhD study. Not only did his ideas and insights help to shape this work, but his aesthetic sense and intuition taught me what good research is truly about. The attitude toward work I learned from him will definitely be my valuable assets for all the years to come. I have always appreciated his best effort to provide me an excellent environment for doing research. To me, he is much more than a thesis supervisor.

It has been a great experience to work with Prof. Fu Minyue from the University of Newcastle, Australia. He is far beyond a collaborator but more like my thesis co-supervisor. I am always impressed by his expertise on grasping decisive ideas for solving research problems, which has significantly broadened my view.

All my fellow students and friends at NTU made my study and life so memorable. In particular, I valued the close discussions with Xiao Nan, Li Tao and Liu Shuai, who helped me to understand quite a lot of critical problems. I am very thankful to Huang Weihua for the pleasant time with him. I also would like to thank all the members of Sensor Network Lab.

The last but not the least, I owe the most to my family. My parents and sisters worked hard and made every attempt for me. They earned little yet managed to bring the best to me, without which it is impossible for me to get a good education.

Abstract

With the rapid advances in information processing, communication and sensing technologies, networked control systems(NCSs) have gained substantial research interest due to their broad applications. The incorporation of resource limited communication networks in the feedback loop induces new challenges in system analysis and synthesis.

One of the main challenges in NCSs is on the quantized feedback control under limited data rate. It is well established that there exists a minimum average data rate above which a discrete LTI system can be stabilized. To exploit the capacity of logarithmic quantization, we show that a finite-level logarithmic quantizer suffices to approach the minimum average data rate for stabilizing a discrete LTI system under two basic network configurations. While in a practical networked system, the issues of packet loss and limited data rate generally co-exist. Data rate theorem for mean square stabilization of an LTI system over lossy digital channels is studied, which reveals the joint effect of quantization and packet loss on the mean square stabilizability. More specifically, the additional data rates required to counter the effect of packet loss are explicitly quantified for single input systems under i.i.d. packet loss, and for scalar systems under Markovian packet loss. Sufficient data rate conditions are also provided for general discrete LTI systems.

Then, we proceed to analyze the effect of limited data rate and packet loss on filtering performance. Under limited data rate constraints, a multi-level quantized innovations Kalman filter is proposed to estimate the state of linear stochastic systems. The quantized filter has the same complexity as the standard Kalman filter and exhibits a comparable estimation performance under quantization with a moderate number of bits. With Markovian packet losses, the problem on mean stability

of the estimation error covariance matrices of Kalman filter is quite challenging. Based on the realization of the packet loss process, two stability notions, namely stability in stopping times and stability in sampling times, are introduced to examine the behavior of the estimation error covariances. Although the two stability notions have different meanings, they are shown to be equivalent. Thus, necessary and sufficient conditions for stability of estimation error covariance matrices are derived for second-order and certain classes of higher-order systems. All stability criteria are given by simple inequalities in terms of the largest open loop pole and the transition probabilities.

Finally, we study the networked multi-agent consensus control problem by considering a group of identical agents, each of which is described by a discrete LTI system. We reveal the joint effect of agent dynamics, network topology and limited data rate on consensusability of linear discrete multi-agent systems under a common control protocol. A remarkable result is that the intrinsic entropy rate of the agent dynamics poses a fundamental limitation on consensusability.

Contents

Acknowledgements	i
Abstract	iii
List of Contents	v
List of Figures	ix
Symbols and Acronyms	xi
1 Introduction	1
1.1 Motivation and Objective	1
1.2 Major Contributions of the Thesis	2
1.3 Organization of the Thesis	5
2 Literature Review	7
2.1 Quantized Control and Estimation	7
2.1.1 Quantized Feedback Control	8
2.1.2 Quantized Estimation	10
2.2 Control and Estimation over Lossy Networks	11
2.3 Consensus Control of Multi-agent Systems	14
Part I Networked Control	17
3 Optimality of Logarithmic Quantization for Stabilization	18
3.1 Problem Formulation	19

3.2	Optimality of Logarithmic Quantization	23
3.2.1	Stabilization Using Quantized Control Feedback	23
3.2.2	Stabilization Using Quantized State Feedback	27
3.3	Summary	32
4	Data Rate Theorem for Stabilization: i.i.d. Packet Loss	34
4.1	Problem Formulation	35
4.2	Main Results	37
4.2.1	Single Input Case	37
4.2.2	Multiple Input Case	44
4.3	Summary	47
5	Data Rate Theorem for Stabilization: Markovian Packet Loss	48
5.1	Problem Formulation	49
5.2	Preliminaries	51
5.3	Scalar Systems	52
5.3.1	Noise Free Systems with Bounded Initial Support	52
5.3.2	General Stochastic Scalar Systems	58
5.4	Vector Systems	67
5.4.1	Real Jordan Form	68
5.4.2	Necessity	68
5.4.3	Sufficiency	70
5.4.4	An Example	74
5.5	Summary	75
6	Stability of Kalman Filtering with Markovian Packet Loss	76
6.1	Problem Formulation	77
6.2	Equivalence of the Two Stability Notions	79
6.3	Second-order Systems	82
6.4	Higher-order systems	83
6.5	Illustrative Example	85
6.6	Proofs	87

6.6.1	Proof of Theorem 6.3	89
6.6.2	Proof of Theorem 6.4	91
6.6.3	Proofs of Results in Section 6.4	94
6.7	Summary	100
7	Quantized Filtering of Linear Stochastic Systems	101
7.1	Problem Formulation	102
7.2	Multi-level Quantized Innovations Kalman Filter	103
7.2.1	Multi-level Quantized Filtering	104
7.2.2	Optimal Quantization Thresholds	106
7.2.3	Convergence Analysis	107
7.3	Robust Quantization	109
7.4	Simulation	110
7.5	Summary	111
Part II	Networked Multi-agent Consensus Control	113
8	Consensusability of Discrete-time Multi-agent Systems	114
8.1	Problem Formulation	115
8.1.1	Communication Graph	115
8.1.2	Consensusability on Graphs	116
8.2	Consensusability with Perfect State Feedback	118
8.2.1	Single Input and Undirected Graphs	118
8.2.2	Performance Analysis of Consensus Protocols	124
8.2.3	Application to Formationability on Graphs	127
8.2.4	Extensions to Multiple Inputs and Digraphs	129
8.3	Special Case: Second-order Consensus	134
8.3.1	Second-order Consensus	134
8.3.2	Illustrative Example	136
8.4	Consensusability with Encoded State Feedback	137
8.4.1	Encoder/decoder Description	139

8.4.2	Consensusability Analysis	140
8.4.3	Illustrative Example	143
8.5	Summary	144
9	Coordination of Multi-agent Systems via Output Feedback	146
9.1	Problem Formulation	147
9.2	Consensusability Analysis	149
9.2.1	Consensus under Static Protocol	149
9.2.2	Consensus under Dynamic Protocol	150
9.3	Special Case: Double-Integrator Systems	153
9.3.1	One-step Relative Position Feedback	153
9.3.2	Two-step Relative Position Feedback	154
9.4	Applications	155
9.4.1	Vehicle Formation on Graphs	156
9.4.2	Formation-based Tracking Problem	157
9.5	Illustrative Examples	160
9.6	Summary	164
10	Conclusion and Recommendation for Future Research	167
10.1	Conclusion	167
10.2	Recommendation for Future Research	168
	Author's Publications	171
	Bibliography	175

List of Figures

1.1	General NCS architecture [47]	2
3.1	Configuration I(left) v.s. Configuration II(right).	21
3.2	Encoder/decoder pair for a digital channel: $g > 0$ is a scaling factor.	22
3.3	Trajectory of scaled state.	31
3.4	Trajectory of the state.	31
4.1	Network configuration.	35
4.2	$R = 2$	44
4.3	$R = 3$	45
5.1	Network configuration	50
5.2	Communication protocol.	57
5.3	Stabilizable and unstabilizable regions for the example in Subsection	
5.4.4.		75
6.1	Network configuration.	77
6.2	A sample path with $q = 0.8$ and $p = 0.1$.	86
6.3	A sample path with $q = 0.6$ and $p = 0.1$.	86
6.4	A sample path with $q = 0.2$ and $p = 0.5$.	87
7.1	Network configuration	103
7.2	$F((z^*)_1^N)$ versus R bits	107
7.3	Comparison of the error variance of the position estimate given by	
	1-LQ-KF and SOI-KF based on 500 samples.	110

7.4	Comparison of the position error variance computed by (7.21) and Monte Carlo method based on Monte Carlo methods with 500 samples.	111
7.5	Comparison of the position error variances by the robust and non-robust 2-bit quantized filters.	112
8.1	Encoder/decoder scheme.	117
8.2	Communication graph	137
8.3	$K = K^*$: the formationability is achieved with the optimal <i>asymptotic convergence factor</i> $r_{asym}^* = 0.7746$	138
8.4	$K = K_1$: the formationability is achieved with an <i>asymptotic convergence factor</i> $r_{asym} = 0.9354$	138
8.5	$K = K_2$: it is impossible to achieve formationability.	139
8.6	A consensus is asymptotically reached.	144
9.1	Stability region: Ω	156
9.2	Consensus is asymptotically reached.	162
9.3	Trajectory of velocity.	163
9.4	Trajectory of position.	163
9.5	A formation is asymptotically produced.	163
9.6	Communication graph	164
9.7	A formation-based leader-follower consensus problem is solved.	164
9.8	Tracking errors are bounded.	165
9.9	$u_0(k)$ is uniformly distributed within $[-1, 1]$: tracking error variances by Monte Carlo method with 500 samples.	165
9.10	$u_0(k)$ is uniformly distributed within $[-5, 5]$: tracking error variances by Monte Carlo method with 500 samples.	165

Symbols and Acronyms

Algebraic Operators

$A^T(A^H)$	Transpose (Conjugate transpose) of matrix A
A^{-1}	Inverse of matrix A
$\text{tr}(A)$	Summation of all the diagonal elements of matrix A
$\det(A)$	Determinant of matrix A
$\rho(A)$	Spectral radius of matrix A
$\bar{\sigma}(A)$	Maximal singular value of matrix A
$\underline{\sigma}(A)$	Minimal singular value of matrix A
$\text{diag}(A_1, \dots, A_n)$	Block diagonal matrix with main block diagonal matrices $A_1, \dots, A_n$ and zero off-diagonal block matrices
$\ \cdot\ $	ℓ^2 norm for vector or the induced norm for matrix
$\ \cdot\ _\infty$	ℓ^∞ norm for vector or the induced norm for matrix
$M > 0$ ($M \geq 0$)	M is positive definite(semi-definite)
$M_1 > M_2$ ($M_1 \geq M_2$)	$M_1 - M_2$ is positive definite(semi-definite)
$A \otimes B$	Kronecker product of A and B

Sets

$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\mathbb{Z}$	Set of integers
$\mathbb{N}$	Set of nonnegative integers

Others

$\lambda_i^u(A)$	An unstable eigenvalue of matrix A
$\mathbf{0}$	Zero vector with a compatible dimension
$\mathbf{1}$	Vector with a compatible dimension and all elements of one
$\lceil \cdot \rceil$	Ceiling function, i.e., $\lceil x \rceil = \min\{l \in \mathbb{Z} l \geq x\}$
$\lfloor \cdot \rfloor$	Floor function, i.e., $\lfloor x \rfloor = \max\{l \in \mathbb{Z} l \leq x\}$
$1_F(w)$	$1_F(w) = 1$ if w belongs to the measurable set F , otherwise 0.
$\mathcal{N}(\mathbf{x}; m, P)$	Gaussian distribution with mean m and covariance P

Acronyms

NCS	- Networked Control System
WSN	- Wireless Sensor Network
LTI	- Linear Time-invariant
MMSE	- Minimum Mean Square Error
LQR	- Linear Quadratic Regulator
LQG	- Linear Quadratic Gaussian
i.i.d.	- Independent and Identically Distributed
w.r.t.	- With respect to

Chapter 1

Introduction

1.1 Motivation and Objective

Networked control systems(NCSs) are feedback control systems wherein the control loops are closed via a real-time network, see Fig. 1.1 for an illustration. The defining feature of an NCS is that information is exchanged between sensors, actuators and controllers via a network. The advantages of NCSs over conventional control systems include low cost of installation, flexibility in system implementation, and the ease of maintenance. Examples of practical significance of NCSs include sensor networks, industrial control networks, multiple vehicle coordination, and Micro-Electro-Mechanical systems, where their aims are to control one or more dynamical systems by deploying a digital communication network for data exchange. The disadvantage is that a real communication network, especially wireless sensor network(WSN), is generally subject to limited network resources and uncertainties such as packet loss, finite data rate, transmission error, data delay, and latency.

The incorporation of a communication network in the feedback loop makes the analysis and design of an NCS complex 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 [82,84]. In addition, there exists a critical packet loss rate above which the mean state estimation error covariance matrices

Figure 1.1: General NCS architecture [47]

of the Kalman filter will diverge [103]. These phenomena strongly imply that the communication network capacity will pose a significant effect on the control/filtering performance. While the conventional control theory is usually established under the assumption that data is transmitted without distortion and the communication theory is relatively indifferent to the specific purpose of the transmitted information, this raises new fundamental problems in investigating NCSs from the perspective of unifying communication, estimation, and control.

These considerations have motivated the research on NCS to better understand the interplay between control/estimation and communication, see the special issues [1,2,15] and the references therein. The primary objective of this thesis is to develop new control tools for analysis and synthesis of an NCS and to quantify the effect of limited data rate and packet loss on the system's performance. Note that the study of other communication effects, e.g., data delay and transmission error, are important as well but not pursued in this thesis.

1.2 Major Contributions of the Thesis

The main results of this thesis are stated informally as follows.

Result 1: (Optimality of logarithmic quantization) The logarithmic quantizer is optimal in the sense of approaching the minimum data rate for stabilizing a linear time-invariant(LTI) system.

One of the most interesting quantizers is the so-called logarithmic quantizer [29,36], which gives the coarsest quantization density for quadratic stabilization of

an unstable single input LTI system. And the existing robust control methods can be applied to design quantized feedback with logarithmic quantization for stabilization and performance control [36]. Though the study of logarithmic quantizer constitutes a vast body of literature, e.g., [16, 19, 29, 36–39, 46, 66, 109], it is unclear whether a logarithmic quantizer can approach the well-known minimum average data rate required for stabilizing an LTI system. This result provides an affirmative answer.

Result 2: (Data rate theorem over lossy channels) Additional bits required to counter the effect of random packet losses on mean square stabilization of linear systems are quantified.

The data rate theorem reveals a striking relationship between the minimum data rate for stabilization of an LTI system and the unstable open loop poles [6, 82, 84, 107]. While in a practical networked system, especially in a resource limited WSN, the issues of packet loss and limited bandwidth in general co-exist. Thus, it is of great importance to find the minimum data rate with the consideration of packet loss, above which a given dynamical system can somehow be stabilized. As analogous to Shannon’s *source coding theorem* [23], this result is an attempt to discover the minimum data rate with random packet loss for mean square stabilization of an LTI system under two probabilistic packet loss models. It contains the existing results on data rate and packet loss rate for stabilization of linear systems as special cases.

Result 3: (An exact characterization of Markovian packet losses on the stability of Kalman filtering) Necessary and sufficient conditions for mean square stability of the estimation error covariance matrices of Kalman filtering with Markovian packet losses are provided for second-order and certain classes of higher-order systems.

Kalman filtering is of tremendous importance in system theory and control due to its various applications ranging from tracking, detection to control. Recently, much attention has been paid to the stability analysis of Kalman filtering with intermittent observations due to the unreliability of the inserted network for propagating data packets, see [47, 100, 132] and the references therein. Because of the temporal correlation of the Markov process, the usual mean square stability condition for the estimation error covariance matrices of vector systems is yet to be

known [51, 100, 119]. By exploiting the system structure, this result explicitly gives 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 losses, in terms of the open loop pole with the largest magnitude and the transition probabilities of the Markov process.

Result 4: (Multi-level quantized innovations Kalman filter) A simple and efficient quantized innovations Kalman filter is proposed to estimate the state of linear stochastic systems.

Due to the introduction of a highly nonlinear quantizer, it is impossible to find an optimal recursive quantized filter. Since the innovation contains new information about the state that is not conveyed by previous observations, it perhaps can be represented by fewer number of bits than measurement. Taking advantage of this observation, we propose a multi-level quantized innovations Kalman filter (MLQ-KF) to estimate the state of linear stochastic systems. The computational cost of the quantized filter is almost identical to that of the Kalman filter. The performance of the optimal MLQ-KF is demonstrated via simulations to be close to that of Kalman filter even under quantization with a moderate number of bits (say 2 bits), which is very useful for applications in sensor networks with the scarcity of bandwidth, computational capacity and energy.

Result 5: (An explicit characterization of agent dynamic, network topology and limited data rate on consensusability of multi-agent systems) The joint effect of agent dynamic, network topology and limited communication data rate on consensusability of linear discrete-time multi-agent systems under a common control protocol is revealed.

Distributed coordination of multiple agents has attracted considerable interest in various scientific communities due to broad applications in many areas [9, 22, 33, 69, 85, 87, 121]. A fundamental problem dealing with the existence of consensus protocols has not been emphasized until the recent work [71]. However, the generalization to the discrete-time case is far from trivial since the feedback information is no longer instantaneously obtained. To achieve consensus for the general agent dynamic, the connectivity of the interaction graph for agents must be sufficiently

strong to dominate the instability of the agent dynamic. Thus, it is crucial to exploit their relationship for achieving consensusability of multi-agent systems under the assumption that the information is exchanged via ideal channels.

1.3 Organization of the Thesis

This thesis is mainly divided into two parts: networked control (Chapters 3-7) and networked multi-agent consensus control (Chapters 8-9). The rest of the thesis is organized as follows.

In Chapter 2, we review the most relevant results in the literature.

In Chapter 3, 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 Chapter 4, data rate theorem for mean square stabilization over lossy channels is developed assuming 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 Chapter 5, we continue to enrich data rate theorem for mean square stabilization over lossy channels. But the packet loss process is 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 Chapter 6, 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 Chapter 7, 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 Chapter 8, the joint effect of agent dynamic, network topology and communication data rate on consensusability of linear discrete-time multi-agent systems is explored. We show that the effect of the *undirected* graph is exactly quantified by its *eigenratio* and the intrinsic entropy rate of each agent poses a fundamental limitation on the graph.

In Chapter 9, an observer-based control protocol using relative output feedback is proposed to achieve consensus of discrete-time multi-agent systems. A necessary and sufficient condition for consensusability under this control protocol is given, which explicitly reveals how the intrinsic entropy rate of the agent dynamic and the *eigenratio* of the *undirected* graph affect consensusability. The theoretical results are illustrated by simulations.

In Chapter 10, some conclusion remarks are drawn and possible future research directions are highlighted.

Chapter 2

Literature Review

Recent years have witnessed a growing research interest in NCSs due to their potentials in broad applications. We feel that a survey of results scattered in the literature is necessary and useful. In this chapter, we will concentrate on the literature that focuses on the interactions between control/estimation and communication of an NCS. Due to the rapid growth of the literature, our review is by no means complete or comprehensive. In Section 2.1, we review quantized control and estimation. In Section 2.2, results on control and estimation over lossy networks are reviewed. The relevant literature on consensus control of multi-agent systems is included in Section 2.3.

2.1 Quantized Control and Estimation

Control using quantized feedback has been an important research area for a long time, even as early as in 1956 [54]. 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 extra additive white noise [114]. 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 WSNs. The change of view on quantization in the control community can be traced back to the

seminal paper [26] 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 is initiated by Wong and Brockett in [115, 116]. Till now, various quantization methods for control and estimation have been developed.

2.1.1 Quantized Feedback Control

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 [26], Brockett and Liberzon [14] 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 study 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., [5, 82, 84, 107, 116, 131], leading to the appealing data rate theorem which states that the stabilization of an LTI system can be achieved with an average data rate R if and only if it satisfies the following inequality:

$$R > \sum_i \log_2 |\lambda_i^u|, \quad (2.1)$$

where λ_i^u denotes an unstable pole of the open-loop system. This implies that if the plant is more unstable, a larger average data rate is required for stabilization. To approach the minimum data rate, a dynamic quantizer is needed.

Coarser quantization means less information flows 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 [29, 36], 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.,

$$\rho = \frac{\prod_i |\lambda_i^u| - 1}{\prod_i |\lambda_i^u| + 1}. \quad (2.2)$$

Similarly, the more unstable the plant, the higher the quantization density (more information flows) is required for quadratic stabilization. By the classical sector bound approach, Fu and Xie [36] 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 implies that the minimum quantization density for MIMO systems is extremely difficult to establish.

However, a static logarithmic quantizer in [29, 36] uses an infinite data rate to represent the quantizer output, which is impractical. In [37], 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 (2.1) to stabilize an LTI system in [123]. See Chapter 3 for details. 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 [112].

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/controller are inter-related but also due to that for a different performance criterion, the optimal quantizer-estimator/controller will be substantially different. In [36], a sector bound approach with the logarithmic quantizer is used to address the 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 [8, 10, 24, 35, 83, 108, 127]. 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 [30, 31], where the number of quantization levels used by the feedback and the convergence time of the closed loop system play a central role.

2.1.2 Quantized Estimation

The key problem in quantized estimation is the joint design of quantizer and the corresponding estimator to minimize the estimation error. The main difficulty 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 [90, 95]. 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 [90, 95]. Motivated by this, an interesting quantizer threshold selection scheme is proposed in [90]. 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 [32] construct an adaptive quantization involving delta modulation with a variable step size. The optimal step size is obtained through an on-line maximum likelihood estimation process, lacking a recursive form. This problem is resolved in [72], where a simple adaptive quantizer and a recursive estimation algorithm are designed to asymptotically approach the MMSE by exploiting the fact that quantizing innovations requires fewer bits than quantizing observations. The advantage of quantizing innovations is also extensively explored in [11, 37, 97, 130].

In fact, abundant quantization schemes have been developed in the context of WSN, e.g., [24, 37, 49, 56, 67, 68, 89, 96, 97, 111, 117, 130]. Luo [68] 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, [89] converts the integration problem into a finite summation using the quantization method. A quantized particle filter is established in [56] 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 [50]. 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 [81, 97], where a simple recursion involving time and measurement updates as in the standard Kalman filter is provided for state estimation. Inspired by [97] 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 [130]. 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 [130]. See Chapter 7 for details. The distinct features of MLQ-KF lie in its simplicity and a comparable performance to the Kalman filter.

2.2 Control and Estimation over Lossy Networks

Due to random fading and congestion, the observation and control packets may be lost while in transit through the 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 [103] that with intermittent observations, Kalman filter is still optimal in the sense of achieving MMSE. By modeling the packet loss process as an independent and identically distributed (i.i.d.) Bernoulli process, Sinopoli *et al* [103] 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 [91] where invertibility on the observable subspace is required. Mo and Sinopoli [79] 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 [79] is that there are counterexamples of second-order systems for which the lower bound given by [103] is not tight.

To capture possible temporal correlations of network conditions, a time homogeneous binary Markov process is adopted to model the packet loss process in [51]. This is usually called the Gilbert-Elliott channel model. Under 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 [51], 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 one. A less conservative sufficient condition for the peak covariance stability under some cases is provided by [119]. 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 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 [122], 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. See Chapter 6 for details.

On the other hand, the optimal linear LQG control over lossy networks has been considered in [43, 52, 100]. The interesting finding is that the separation principle for control and estimation continues to hold under a TCP-like communication protocol while it fails under a UDP-like protocol.

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 [101], 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 could be exactly computed for scalar systems and only has lower and upper bounds for vector systems [101]. Another performance metric called the stochastic boundedness is introduced in [55] 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.

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 [73, 75, 76, 78, 98, 106, 124, 128]. 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, the 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 [75, 76, 106], which unfortunately fails for the moment stabilization [98]. Data rate theorem for mean square stabilization over lossy feedback channels, which is analogous to the Shannon's *source coding theorem*, is established in [73, 78, 124, 128]. See Chapters 4 and 5 for our contributions to this topic.

2.3 Consensus Control of Multi-agent Systems

Distributed coordination of multiple agents has attracted considerable interest in various scientific communities due to broad applications in many areas including formation control [33, 121], distributed sensor networks [21, 22], flocking [85, 87], distributed computation [69], and synchronization of coupled chaotic oscillators [9, 27, 105]. The common property of those applications is that each individual agent lacks global knowledge of the whole system and can only interact with its neighbors to achieve certain global behaviors. Within this framework, communication graph (topology), which determines what information is available for each agent at each time instant, is an important aspect of information flow in distributed coordination. For example, to achieve an average consensus which requires the states of all agents to asymptotically converge to the average of their initial values, the communication graph must contain a spanning tree for a fixed topology [88, 93] while for a switching topology, the union of the communication graphs should contain a spanning tree frequently enough as the system evolves [53, 63, 93].

Research on consensus control can be roughly categorized depending on whether the dynamic of each agent is continuous or discrete. The machinery for studying the continuous-time average consensus is based on the graph Laplacian matrix theory while for the discrete-time case, the Perron matrix theory becomes a convenient tool. Noticeable works focusing on the first-order integrator networks include [33, 61, 62, 88, 93] for the continuous-time consensus and [53, 63, 80, 93] for the

discrete-time consensus. The higher-order consensus problem has been considered in [65, 94, 113]. Though significant progresses have been made toward this topic, all the aforementioned works exclusively concentrate on the design of consensus protocols and convergence analysis. A fundamental problem dealing with the existence of consensus protocols has not been emphasized until the recent work [64, 71], which gives a necessary and sufficient condition for consensusability of continuous-time multi-agent systems with respect to (w.r.t.) a certain type of consensus protocols. By merging ideas from algebraic graph theory and control theory, they characterize the interplay between communication topology and agent dynamics. In particular, the minimal requirement for consensusability is that the dynamic of each identical agent has to be stabilizable and the fixed communication graph must contain a spanning tree for unstable agent dynamics. Related problems on synchronizability of complex networks has been investigated in [9, 17, 63, 105], where sufficient conditions are offered to achieve synchronization of a couple of chaotic oscillators. They illustrate that with an *undirected* communication graph, the network synchronizability is related not only to the second smallest eigenvalue of the graph but also the ratio of the second smallest to the largest eigenvalues of the graph Laplacian matrix, which is called the *eigenratio* of an *undirected* graph in this thesis.

However, the consensusability problem of discrete-time multi-agent systems is considerably harder than its counterpart of continuous-time case [42, 99, 110, 125, 129]. Indeed the consensusability condition derived in [64] basically requires that the complementary sensitivity of each individual agent be positive real, and this condition is fulfilled under feedback control law of LQR. In fact the positive real condition can be satisfied under static output feedback control laws, provided that the agent dynamics are strictly minimum phase and admit the nonsingular high frequency gain [41]. For single input case, we have derived a necessary and sufficient condition for consensusability of linear discrete-time multi-agent systems under a common control protocol, in terms of the agent dynamics and the communication graph [129]. A remarkable feature of this result is that for *undirected* graphs, it clearly reveals the joint impact of the eigenratio of the undirected graph and the intrinsic entropy rate of the agent dynamics on consensusability. For multiple inputs,

sufficient conditions are derived in [125]. See Chapter 8 for details. The convergence speed to consensus is an important design consideration. For *undirected* graphs, the convergence rate to consensus is analyzed and a lower bound of the optimal convergence rate is given in terms of the eigenratio of an undirected graph, which is shown to be attainable for some special cases including multi-agent systems of second-order dynamic [125].

Noting that a real communication channel is constrained by a finite bandwidth, the packet exchanged between each paired agents carries information of a finite number of bits. It is clear that a lower data rate will induce a coarser quantization and thus more information loss. The quantized average consensus problem on the first-order integrator systems has been intensively studied in recent years, see [16, 18, 60] and the references therein. In [129], we also establish that consensusability with perfect state feedback implies consensusability with encoded state feedback, provided that data rate is not less than an explicitly determined lower bound. An observer-based distributed control protocol using relative output is proposed in our work [126]. See Chapter 9 for details.

Part I

Networked Control

Chapter 3

Optimality of Logarithmic Quantization for Stabilization

Perhaps one of the most significant and elegant results on quantized feedback control is the celebrated data rate theorem [84], which states that to stabilize a discrete LTI system, the average data rate of the noiseless channel between the controller and plant must satisfy the strict inequality:

$$R > \sum_i \log_2 |\lambda_i^u| \triangleq R_{\min}, \quad (3.1)$$

where λ_i^u denotes an unstable eigenvalue of the open-loop system matrix. In general, a dynamic quantizer is needed to achieve the minimum data rate.

Motivated by the various advantages of the logarithmic quantizer [29] and to explore its capacity, we ask the question in this chapter: does a logarithmic quantizer require an average data rate higher than the minimum data rate for stabilization? This chapter gives a negative answer. In particular, we derive explicit finite-level logarithmic quantizers and the corresponding controllers to approach the minimum data rate given by (3.1) under two basic network configurations.

The chapter is organized as follows. The problem of interest is formulated in Section 3.1. Section 3.2 constitutes the main part of the chapter where the attainability of the minimum average data rate via logarithmic quantization is proved. Concluding remarks are drawn in Section 3.3. A technical lemma is given in the

appendix of this chapter.

3.1 Problem Formulation

Consider a discrete LTI unstable system

$$x_{k+1} = Ax_k + Bu_k + w_k, \forall k \in \mathbb{N}, \quad (3.2)$$

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., $\|w_k\|_\infty \leq d, \forall k \in \mathbb{N}$ for some positive d . Without loss of generality, assume that $A \in \mathbb{R}^{n \times n}$ has two distinct real Jordan blocks, i.e., $A = \text{diag}(J_1, J_2)$, 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 $|\lambda_1| \neq |\lambda_2|$. Moreover, (A, B) is a controllable pair.

Remark 3.1. *There is no loss of generality to focus on the system of the form (3.2). In fact, consider a system as follows:*

$$\begin{cases} x_{k+1} &= Ax_k + Bu_k + w_k, \\ y_k &= Cx_k + v_k, \end{cases} \quad \forall k \in \mathbb{N}, \quad (3.3)$$

where $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^l$ is the control input, $y_k \in \mathbb{R}^m$ is the output, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^p$ are bounded additive disturbances. (A, B) and (C, A) are stabilizable and detectable pairs, respectively and $\text{rank}(B) = l \leq n$.

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 [20] can be constructed to estimate the state of the system. The estimation error will be uniformly bounded after n steps and indepen-

dent of the initial state. Hence, it is sensible to focus on the state feedback case.

By applying the Wonham decomposition to (3.3) [20], one can convert the multiple inputs system to l single input ones. 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:

$$\bar{A} = \begin{bmatrix} A_1 & A_{12} & \dots & A_{1l} \\ 0 & A_2 & \dots & A_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_l \end{bmatrix}, \bar{B} = \begin{bmatrix} B_1 & B_{12} & \dots & B_{1l} \\ 0 & B_2 & \dots & B_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_l \end{bmatrix},$$

where (A_i, B_i) with $A_i \in \mathbb{R}^{n_i \times n_i}$ and $B_i \in \mathbb{R}^{n_i}$, $i \in \{1, \dots, l\}$, is a controllable pair and $\sum_{i=1}^l n_i = n$. For illustration and brevity, let $l = 2$ and assume that the 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

$$x_{k+1}^1 = A_1 x_k^1 + B_1 u_k^1 + A_{12} x_k^2 + B_{12} u_k^2 + w_k^1; \quad (3.4)$$

$$x_{k+1}^2 = A_2 x_k^2 + B_2 u_k^2 + w_k^2. \quad (3.5)$$

If x_k^2 is stabilized with a data rate greater than $\sum_{\lambda(A_2)} \log_2 |\lambda(A_2)|$, then $\|x_k^2\|_\infty$ will be uniformly bounded and can be treated as a bounded disturbance input to the subsystem (3.4), which can be stabilized similarly with a data rate greater than $\sum_{\lambda(A_1)} \log_2 |\lambda(A_1)|$. As in [84, Section 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 system of form (3.2).

Definition 3.1. [36] A quantizer is called a logarithmic quantizer if it has the form:

$$Q_\infty(v) = \begin{cases} u^{(i)}, & \text{if } \frac{1}{1+\delta} u^{(i)} < v \leq \frac{1}{1-\delta} u^{(i)}, v > 0; \\ 0, & \text{if } v = 0; \\ -Q_\infty(-v), & \text{if } v < 0 \end{cases} \quad (3.6)$$

where $u^{(i)}$, $i \in \mathbb{N}$, are from the set $\mathcal{U} = \{\pm u^{(i)} : u^{(i)} = \rho^i u^{(0)}, i = \pm 1, \pm 2, \dots\} \cup$

$\{\pm u^{(0)}\} \cup \{0\}$, $u^{(0)} > 0$, $\rho \in (0, 1)$ represents the quantizer density and $\delta = \frac{1-\rho}{1+\rho}$.

However, the logarithmic quantizer in (3.6) has an infinite number of quantization levels and needs an infinite number of bits to represent the quantizer output. Define a $(2N + 2)$ -level logarithmic quantizer with density ρ as follows:

$$Q_N(v) = \begin{cases} \rho^i(1 - \delta), & \text{if } \rho^{i+1} < v \leq \rho^i, 0 \leq i \leq N - 1; \\ 0, & \text{if } 0 \leq v \leq \rho^{N-1}; \\ -Q_N(-v), & \text{if } -1 \leq v < 0, \\ 1, & \text{otherwise.} \end{cases} \quad (3.7)$$

In the above, we have chosen $u^{(0)} = \frac{2\rho}{1+\rho}$ in (3.6) and for any $v \notin [-1, 1]$, the alarm level 1 is used to indicate the overflow of the quantizer. Thus, the number of bits required to represent quantizer output is $\lceil \log_2(2N + 2) \rceil$.

Figure 3.1: Configuration I(left) v.s. Configuration II(right).

Two basic network configurations shown in Fig. 3.1 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 (3.2), is to be quantized. In **configuration II**, the vector state measurement is quantized. As in [107], the encoder/decoder pair for the limited data rate communication is described in Fig. 3.2. The first stage of the encoding process consists of designing a scaling factor g^{-1} such that the quantizer input z/g is within the quantization range. The output of the finite-level logarithmic quantizer $Q_N(z/g)$, which takes values from the set $\{\pm \rho^i(1 - \delta) : i = 0, \dots, N - 1\} \cup \{0, 1\}$, is encoded into binary sequence and transmitted via a limited data rate communication channel. The

Figure 3.2: Encoder/decoder pair for a digital channel: $g > 0$ is a scaling factor.

decoder receives the binary sequence and 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 corresponding quantizer 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 (3.2).

Remark 3.2. *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 this 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 3.3. *The two configurations have been widely adopted in literature. For example, [29, 36, 109] focus on **Configuration I** while [14, 84, 107] are restricted to **Configuration II**. The differences are that the quantizer in the encoder of Fig. 3.2 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 (3.2).*

3.2 Optimality of Logarithmic Quantization

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 (3.2) under **configuration I** and **configuration II**, respectively. Thus, the logarithmic quantization is optimal in the sense of using the minimum data rate for stabilization of a linear system.

3.2.1 Stabilization Using Quantized Control Feedback

Theorem 3.1. *Consider the system (3.2) and network **configuration I** of Fig. 3.1, 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 $\mathcal{L}(A) \subset \mathbb{N}$ by

$$\mathcal{L}(A) = \begin{cases} \mathbb{N}, & \text{if } \lambda_1, \lambda_2 \in \mathbb{R}, \\ \{i \in \mathbb{N} | \lambda_1^i \neq (\lambda_1^*)^i\}, & \text{if } \lambda_1 \in \mathbb{C}, \lambda_2 \in \mathbb{R}; \\ \{i \in \mathbb{N} | \lambda_2^i \neq (\lambda_2^*)^i\}, & \text{if } \lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{C}; \\ \{i \in \mathbb{N} | \lambda_j^i \neq (\lambda_j^*)^i, j = 1, 2\}, & \text{otherwise.} \end{cases}$$

Obviously, $\mathcal{L}(A)$ has infinitely many elements. Since (A, B) is a controllable pair, it is readily verified that $(A^m, A^{m-1}B)$ is a controllable pair if $m \in \mathcal{L}(A)$. By applying the control input $u_{mk+t} = 0$, if $1 \leq t \leq m-1$, the down-sampled system of (3.2) with a down-sampling factor m is expressed as

$$x_{m(k+1)} = A^m x_{mk} + A^{m-1} B u_{mk} + d_k, \quad (3.8)$$

where $d_k = \sum_{t=0}^{m-1} A^{m-1-t} w_{mk+t}$. Due to the controllability of $(A^m, A^{m-1}B)$, (3.8) 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 (3.8) into the controllable canonical form:

$$x_{k+1}^s = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \dots & -\alpha_1 \end{bmatrix} x_k^s + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u_k^s + w_k^s. \quad (3.9)$$

Here we define $x_k^s \triangleq Px_{mk}$, $u_k^s \triangleq u_{mk}$ and $w_k^s \triangleq Pd_k$. It is clear from (3.9) 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 (3.2) due to $m < \infty$. 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

$$\begin{cases} u_k^s = gQ_N(f_k^s/g); \\ f_k^s = \begin{cases} 0, & \text{if } k < n; \\ \sum_{j=0}^{n-1} \alpha_{j+1} x_{k-j}^s(n), & \text{if } k \geq n, \end{cases} \end{cases} \quad (3.10)$$

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 $|\alpha_k| \leq \alpha_0 |A|^m, \forall k \in \{1, \dots, n\}$ since $|\lambda_j| \geq 1, \forall j \in \{1, 2\}$.

Proof of Theorem 3.1: The necessity part has been well established in [84, 107]. Only the sufficiency needs to be elaborated. First, given any $R > \log_2 |A|$, there exists an $\alpha > 1$ satisfying $R \geq \log_2(\alpha |A|)$. Based on Lemma 3.2 and by choosing $\beta_1 = 0$, $\beta_2 = n\alpha_0$, it is possible to select a pair of $m \in \mathcal{L}(A)$ and $N > 0$ such that $\forall \epsilon > 0$,

$$\log_2 \left[1 + \frac{2\log_2(n\alpha_0 |A|^m)}{\log_2 \frac{n\alpha_0 |A|^{m+\epsilon+1}}{n\alpha_0 |A|^{m+\epsilon-1}}} \right] < \log_2(2N + 2) \leq m \log_2 \alpha + \log_2 |A|^m - 1. \quad (3.11)$$

The quantizer level parameter N is determined by (3.11) and the number of bits required to represent each quantizer output is $\lceil \log_2(2N + 2) \rceil$. In addition, it follows

from (3.11) that the average data rate of this protocol satisfies

$$\frac{\lceil \log_2(2N+2) \rceil}{m} \leq \log_2(\alpha|A|) \leq R.$$

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. 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 $|x_k^s(n)| > \Delta$ 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 Δ 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 $|x_k^s(n)| > \Delta$ or $|f_k^s/g| > 1$.

First, assume $|x_k^s(n)| \leq \Delta, \forall k \in \{0, \dots, n-1\}$, which will be relaxed later. Then, for any $k \geq n$, assume that $|x_j^s(n)| \leq \Delta, \forall j \leq k$, it is obvious that $\forall j \in \{n, n+1, \dots, k\}$, $|f_j^s| = |\sum_{t=0}^{n-1} \alpha_{t+1} x_{j-t}^{(n)}| \leq n\alpha_0|A|^m \Delta \triangleq g$. Since $|f_j^s/g| \leq 1$ and $|x_j^s(n)| \leq \Delta, \forall j \leq k$, no alarm level 1 occurs before time k . From (3.9), there exist vectors $c_j \in \mathbb{R}^n, j \in \{0, \dots, n-1\}$, 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

$$x_{k+1}^s(n) = - \sum_{j=0}^{n-1} \alpha_{j+1} x_{k-j}^s(n) + u_k^s + s_k^s, k \geq n. \quad (3.12)$$

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 $\delta = \frac{1}{n\alpha_0|A|^{m+\epsilon}}, \rho = \frac{1-\delta}{1+\delta}$ and $\Delta > 0$ to satisfy that

$$\Delta > \max\left\{\frac{\tilde{s}}{1 - (n\alpha_0|A|^m)^2 \rho^{2N+1}}, \frac{\tilde{s}}{1 - n\alpha_0|A|^m \delta}\right\}. \quad (3.13)$$

In light of (3.11), it is easy to verify that

$$\begin{cases} (n\alpha_0|A|^m)^2 \rho^{2N+1} < 1 \\ n\alpha_0|A|^m \delta < 1. \end{cases} \quad (3.14)$$

Inserting the quantized control in (3.10) into the system (3.12) results in that

$$\begin{aligned}
|x_{k+1}^s(n)| &\leq \begin{cases} |f_k^s| + \tilde{s}, & \text{if } |f_k^s/g| \leq \rho^N \\ \delta|f_k^s| + \tilde{s}, & \text{if } \rho^N < |f_k^s/g| \leq 1 \end{cases} \\
&\leq \begin{cases} n\alpha_0|A|^m\rho^N\Delta + \tilde{s}, & \text{if } |f_k^s/g| \leq \rho^N \\ n\alpha_0|A|^m\delta\Delta + \tilde{s}, & \text{if } \rho^N < |f_k^s/g| \leq 1 \end{cases} \\
&\leq \Delta \text{ due to the selection of } \Delta \text{ in (3.13) and (3.14).}
\end{aligned}$$

Inductively, $|x_k^s(n)| \leq \Delta, \forall k \in \mathbb{N}$. Next, assume that the assumption that $|x_j^s(n)| \leq \Delta, \forall j \in \{0, \dots, n-1\}$ 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

$$\gamma = \sum_{j=1}^n \alpha_j + 1 \quad (3.15)$$

to dynamically update the scaling factor. Specifically, set $\Delta_k = \Delta$ and update the scaling factor as follows:

$$\Delta_{k+j+1} = \begin{cases} \gamma\Delta_{k+j}, & \text{alarm level 1 occurs,} \\ \Delta_{k+j}, & \text{otherwise,} \end{cases}$$

which is simultaneously processed on the both sides of the channel. Set $u_{k+j}^s = 0$, the increasing speed of Δ_{k+j} is thus faster than that of $x_{k+j}^s(n)$ by (3.9) and (3.15). Δ_{k+j} will eventually capture $x_{k+j}^s(n)$ or $\lim_{j \rightarrow \infty} \frac{x_{k+j}^s(n)}{\Delta_{k+j}} = 0$. Let the scaling factor be $g_{k+j} = n\alpha_0|A|^m\Delta_{k+j}$, then $\lim_{j \rightarrow \infty} f_{k+j}^s/g_{k+j} = 0$, implying that there exists a finite $k_0 \geq n-1$ such that the signals received within the time period $\{k+k_0-n+1, \dots, k+k_0\}$ do not give rise to the alarm level 1, which suggests that $|x_{k+j}^s(n)| \leq \Delta_{k+k_0}, \forall j \in \{k_0-n+1, \dots, k_0\}$. Then, repeating the above proof as the bounded case at time $k+k_0+n-1$ yields that $|x_{k+j}^s(n)| \leq \Delta_{k+k_0}, \forall j \geq k_0+n-1$.

Finally, it follows that $\limsup_{k \rightarrow \infty} |x_k^s(n)| < \infty$, which eventually leads to that $\limsup_{k \rightarrow \infty} \|x_k\|_\infty < \infty$. ■

The following corollary gives the corresponding result for asymptotic stabilization, i.e., $\lim_{k \rightarrow \infty} \|x_k\|_\infty = 0$.

Corollary 3.1. *Consider the system (3.2) with $w_k = 0$ and network **configuration I** of Fig. 3.1, 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 sufficiency needs to be established. Define a scaling factor

$$\eta \triangleq \max\{(n\alpha_0|A|^m)^2\rho^{2N+1}, n\alpha_0|A|^m\delta\}, \quad (3.16)$$

which is strictly less than one by (3.14), i.e., $\eta < 1$. Let $\Delta_{k+1} = \eta\Delta_k$ with an arbitrary $\Delta_0 > 0$, which is assumed to be agreed by both the quantizer and the decoder. Assume that $|x_k^s(n)| \leq \Delta_0, \forall k \in \{0, \dots, n-1\}$, the control input and quantizer are given in Theorem 3.1 with Δ replaced by Δ_k . Then, it is straightforward that $|x_{k+1}^s(n)| \leq \eta\Delta_k = \Delta_{k+1}$. Thus, $x_k^s(n)$ can be driven exponentially to zero since $\lim_{k \rightarrow \infty} |x_k^s(n)| \leq \Delta_0 \lim_{j \rightarrow \infty} \eta^j = 0$. Due to $m < \infty$, it follows that $\lim_{k \rightarrow \infty} \|x_k\|_\infty = 0$. The removal of the boundedness assumption for the initial state is similar to what we have done in Theorem 3.1. ■

3.2.2 Stabilization Using 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 subsection shows that the logarithmic quantizer can still approach the minimum average data rate.

Theorem 3.2. *Consider the system (3.2) and network **configuration II** of Fig. 3.1, 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 (3.2), where the

downsampling factor $m \geq 2n$ is to be determined later. More precisely, indexing the scalar components of the state of (3.2) by an additional superscript $h \in \{1, \dots, n\}$. At time $t = mk + h - 1$, the h -th element of x_k^s will be quantized by $Q_{N_1}(x_k^s(h)/\Delta)$ if $h \leq n_1$ and $Q_{N_2}(x_k^s(h)/\Delta)$ otherwise, where the quantization level parameter N_i and Δ are determined by the available data rate. Neglecting the transmission time implies that the quantized x_k^s can reach the controller before time $mk + n$. Since $m \geq 2n$, the control law within one cycle $\{mk, \dots, m(k+1) - 1\}$ can be proposed as follows:

$$\left\{ \begin{array}{l} \begin{bmatrix} u_{mk+m-1} \\ \vdots \\ u_{mk+m-n} \end{bmatrix} = -\Delta \mathcal{C}^T (\mathcal{C} \mathcal{C}^T)^{-1} A^m Q\left(\frac{x_k^s}{\Delta}\right), \\ u_{mk+t} = 0, \forall t \in \{0, \dots, m-n-1\}, \end{array} \right. \quad (3.17)$$

where the controllability matrix $\mathcal{C}$ is defined as $\mathcal{C} \triangleq [B, AB, \dots, A^{n-1}B]$ and the product quantizer $Q(\cdot)$ is composed by

$$Q(\cdot) = \underbrace{[Q_{N_1}(\cdot), \dots, Q_{N_1}(\cdot)]}_{n_1} \underbrace{[Q_{N_2}(\cdot), \dots, Q_{N_2}(\cdot)]}_{n_2}^T.$$

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

$$\|J_i^m\|_\infty \leq \zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^m, \forall i \in \{1, 2\}. \quad (3.18)$$

Proof. It is known that there exists a $\zeta > 0$, independent of J_i , n_i , and m_i , such that $\|J_i\| \leq \zeta m_i^{n_i-1} |\lambda_i|^m$, where $\|\cdot\|$ is the spectral norm induced from the Euclidean norm [84]. Together with the fact that $\|J_i\|_\infty \leq \sqrt{n_i} \|J_i\|$, (3.18) is immediately inferred. $\blacksquare$

Proof of Theorem 3.2: 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 3.2, for any $\epsilon > 0$, we can choose a pair of integers $m \geq 2n$ and N_i satisfying that

$\forall i \in \{1, 2\}$,

$$\log_2 \left[1 + \frac{2 \log_2 \zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^m}{\log_2 \frac{\zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^{m+\epsilon+1}}{\zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^{m+\epsilon-1}}} \right] < \log_2(2N_i+2) \leq m \log_2 \alpha + \log_2 |\lambda_i|^m - 1. \quad (3.19)$$

The quantization level parameter N_i is selected based on (3.19). The average data rate of this protocol is computed by

$$\frac{n_1 \lceil \log_2(2N_1+2) \rceil + n_2 \lceil \log_2(2N_2+2) \rceil}{m} \leq \log_2(\alpha |A|) \leq R,$$

implying 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 $\delta_i = \frac{1}{\zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^{m+\epsilon}}$ and $\rho_i = \frac{1-\delta_i}{1+\delta_i} = \frac{\zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^{m+\epsilon-1}}{\zeta \sqrt{n_i} m^{n_i-1} |\lambda_i|^{m+\epsilon+1}}$, which gives that

$$\begin{cases} (\zeta \sqrt{n_i} m^{n_i-1})^2 \rho_i^{2N_i+1} < 1, \\ \delta_i \zeta \sqrt{n_i} m^{n_i-1} < 1. \end{cases} \quad (3.20)$$

Define the uniform upper bound of d_k in (3.8) by $D \triangleq d \sum_{t=0}^{m-1} \|A\|_\infty^{m-1-t}$, then $\|d_k\|_\infty \leq D, \forall k \in \mathbb{N}$. Similarly, the initial state x_0 is assumed to be bounded by $\Delta > 0$, where Δ is selected to satisfy

$$\Delta \geq \max_{i \in \{1, 2\}} \left\{ \frac{D}{1 - (\zeta \sqrt{n_i} m^{n_i-1})^2 \rho_i^{2N_i+1}}, \frac{D}{1 - \delta_i \zeta \sqrt{n_i} m^{n_i-1}} \right\}. \quad (3.21)$$

Inserting the control law in (3.17) into (3.2) yields that

$$\begin{aligned} x_{k+1}^s &= A^m x_k^s + \sum_{t=0}^{m-1} A^{m-1-t} (B u_{mk+t} + w_{mk+t}) \\ &= A^m x_k^s + \sum_{t=m-n}^{m-1} A^{m-1-t} B u_{mk+t} + d_k \\ &= A^m [x_k^s - \Delta Q(\frac{x_k^s}{\Delta})] + d_k. \end{aligned} \quad (3.22)$$

Assuming that $\|x_k^s\|_\infty \leq \Delta$, there is no alarm level 1 for the scaled state $x_k^s(h)$ by the scaling factor Δ . 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 (3.22), we obtain that:

$$\begin{aligned}
\|(x_{k+1}^s)^{(i)}\|_\infty &= \|J_i^m[(x_k^s)^{(i)} - \Delta Q^{(i)}(\frac{(x_k^s)^{(i)}}{\Delta})] + d_k^{(i)}\|_\infty \\
&\leq \|J_i^m\|_\infty \|(x_k^s)^{(i)} - \Delta Q^{(i)}(\frac{(x_k^s)^{(i)}}{\Delta})\|_\infty + D \\
&\leq \begin{cases} \|J_i^m\|_\infty \|(x_k^s)^{(i)}\|_\infty + D, & \text{if } \|(x_k^s)^{(i)}\|_\infty \leq \rho_i^{N_i-1} \\ \|J_i^m\|_\infty \delta_i \|(x_k^s)^{(i)}\|_\infty + D, & \text{if } \rho_i^{N_i-1} < \|(x_k^s)^{(i)}\|_\infty \leq 1 \end{cases} \\
&\leq \begin{cases} \zeta \sqrt{n_i} m^{n_i-1} \rho_i^{N_i-1} \Delta + D, & \text{if } \|(x_k^s)^{(i)}\|_\infty \leq \rho_i^{N_i-1} \\ \zeta \sqrt{n_i} m^{n_i-1} \delta_i \Delta + D, & \text{if } \rho_i^{N_i-1} < \|(x_k^s)^{(i)}\|_\infty \leq 1. \end{cases} \\
&\leq \Delta \text{ by (3.20) and (3.21).}
\end{aligned}$$

Inductively, $\|x_k^s\|_\infty \leq \Delta, \forall k \in \mathbb{N}$. This follows that $\limsup_{k \rightarrow \infty} \|x_k\|_\infty < \infty$ due to that $m < \infty$. The removal of the boundedness assumption of $\|x_0^s\|_\infty \leq \Delta$ is similar to the case of **configuration I** and is omitted. ■

Remark 3.4. *It is worth mentioning that the parallel result in Corollary 3.1 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.*

Example 3.1. *In this section, an example is attached to illustrate the proposed quantizer and corresponding control law to stabilize an unstable scalar system, i.e.,*

$$x_{k+1} = 2x_k + u_k + w_k, \forall k \in \mathbb{N}. \quad (3.23)$$

The initial state is chosen as $x_0 = 10$. The additive disturbance noise is bounded with $|w_k| \leq 0.5$. The data rate is assumed to be $R = \log_2(2|\lambda|) = 2$. We choose $\epsilon = 0.2$ and $\tau = 5, N = 60$ to satisfy (3.19):

$$6.8146 = \log_2(1 + \frac{2 \log_2 2^5}{\log_2 \frac{2^5 + \epsilon + 1}{2^5 + \epsilon - 1}}) < \log_2(2 \times 60 - 1) = 6.8948 < 5 \log_2 2 - 2 + \log_2 2^5 = 8. \quad (3.24)$$

The quantizer parameter $\delta = \frac{1}{\epsilon+2^5} = 0.0311$ and the quantizer density $\rho = \frac{1-\delta}{1+\delta} = 0.9398$. By (3.13), a large Δ is chosen as

$$\Delta = \frac{5 \times 0.5 \times 2^5}{1 - 2^5 \times \delta} = 1.2880 \times 10^5.$$

Using the quantizer in (3.7) and control law in (3.17), the scaled state trajectory of (3.22), which is defined as $\frac{z_{j\tau}}{\Delta}$ is shown in Fig. 3.3. It is clear that $|z_{j\tau}| \leq \Delta$, implying the stability of z_k .

Figure 3.3: Trajectory of scaled state.

To investigate the asymptotic stabilization, we remove the disturbance, i.e., $w_k = 0$ and let $\Delta_0 = 20$, $\kappa(\tau, \lambda) = |\lambda|^\tau$ and $\eta = \max\{|\lambda|^\tau \rho^{N-1}, |\lambda|^\tau \delta\} = 0.9938$ by (3.16). The state trajectory using the control law offered in Corollary 3.1 is shown in Fig. 3.4, where the state eventually converges to zero.

Figure 3.4: Trajectory of the state.

3.3 Summary

We have addressed the attainability of the logarithmic quantizer in the sense of approaching the minimum average data rate 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, our proposed control law and quantizer may produce a poor transient response.

Appendix: A Technical Lemma

Define $\kappa(m, \lambda) = (\beta_1 m^{n-1} + \beta_2)|\lambda|^m$, $\forall n \geq 1, \beta_1, \beta_2 \geq 0$ and $\beta_1 + \beta_2 > 0$, we have the following result.

Lemma 3.2. *$\forall \alpha > 1$, $\forall \epsilon > 0$ and $|\lambda| \geq 1$, there exist positive integers m and N such that*

$$\log_2 \left[1 + \frac{2 \log_2 \kappa(m, \lambda)}{\log_2 \frac{\kappa(m, \lambda) + \epsilon + 1}{\kappa(m, \lambda) + \epsilon - 1}} \right] < \log_2(2N + 2) \leq m \log_2 \alpha + \log_2 |\lambda|^m - 1. \quad (3.25)$$

Proof. It is trivial if $|\lambda| = 1$ and $\beta_1 = 0$. Assume $|\lambda| > 1$ or $\beta_1 > 0$, then $\kappa(m, \lambda) \rightarrow \infty$ as $m \rightarrow \infty$. Jointly with the fact that $\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e$ yields

$$\log_2 \frac{\kappa(m, \lambda) + \epsilon + 1}{\kappa(m, \lambda) + \epsilon - 1} \cong 2\kappa^{-1}(m, \lambda) \log_2 e, \quad (3.26)$$

if 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 $\ln \kappa(m, \lambda) \geq 1$, we have

$$\begin{aligned} (1 + \kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} &= (1 + (\beta_1 m^{n-1} + \beta_2)|\lambda|^m \ln \kappa(m, \lambda))^{1/m} \\ &\geq |\lambda| \beta_1^{1/m} (m^{1/m})^{n-1} (\ln \kappa(m, \lambda))^{1/m} \geq |\lambda| \beta_1^{1/m} \rightarrow |\lambda| \text{ as } m \rightarrow \infty \end{aligned}$$

due to that $\lim_{m \rightarrow \infty} x^{1/m} = 1, \forall x > 0$. On the other hand, choosing a large m such that $\kappa(m, \lambda) \ln \kappa(m, \lambda) \geq 1$, $\beta_1 m^n \geq \beta_2$ and $2\beta_1 m^n \geq \ln(2\beta_1 m^n)$ gives the following inequalities:

$$\begin{aligned} (1 + \kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} &\leq (2\kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} \\ &\leq |\lambda|(4\beta_1)^{1/m}(m^{1/m})^n [m \ln |\lambda| + \ln(2\beta_1 m^n)]^{1/m} \\ &\leq |\lambda|(4\beta_1)^{1/m}(m^{1/m})^n [m^n \ln |\lambda| + 2\beta_1 m^n]^{1/m} \\ &= |\lambda|[4\beta_1(\ln |\lambda| + 2\beta_1)]^{1/m}(m^{1/m})^{2n} \rightarrow |\lambda| \text{ as } m \rightarrow \infty, \end{aligned}$$

due to that $\lim_{m \rightarrow \infty} m^{1/m} = 1$.

Case 2: $\beta_1 = 0, \beta_2 > 0$ and $|\lambda| > 1$.

Let $m \geq 1$, it immediately follows that

$$\begin{aligned} (1 + \kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} &\geq |\lambda|\beta_2^{1/m}(\ln \beta_2 + m \ln |\lambda|)^{1/m} \\ &\geq |\lambda|\beta_2^{1/m}(\ln \beta_2 + \ln |\lambda|)^{1/m} \rightarrow |\lambda| \text{ as } m \rightarrow \infty. \end{aligned}$$

Also, for a sufficiently large m , e.g., $\kappa(m, \lambda) \ln \kappa(m, \lambda) \geq 1$ and $m \geq \ln \beta_2 / \ln |\lambda|$, one can establish that

$$\begin{aligned} (1 + \kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} &\leq (2\beta_2 |\lambda|^m \ln(\beta_2 |\lambda|^m))^{1/m} \\ &\leq |\lambda|(4\beta_2 \ln |\lambda|)^{1/m} m^{1/m} \rightarrow |\lambda| \text{ as } m \rightarrow \infty. \end{aligned}$$

Consequently, we derive the limit $\lim_{m \rightarrow \infty} (1 + \kappa(m, \lambda) \ln \kappa(m, \lambda))^{1/m} = |\lambda|$. In the light of (3.26), it is clear that $\lim_{m \rightarrow \infty} \left[1 + \frac{2 \log_2 \kappa(m, \lambda)}{\log_2 \frac{\kappa(m, \lambda) + \epsilon + 1}{\kappa(m, \lambda) + \epsilon - 1}} \right]^{\frac{1}{m}} = |\lambda|$, which further implies that for a sufficiently large m ,

$$\log_2 \left[1 + \frac{2 \log_2 \kappa(m, \lambda)}{\log_2 \frac{\kappa(m, \lambda) + \epsilon + 1}{\kappa(m, \lambda) + \epsilon - 1}} \right] \cong \log_2 |\lambda|^m.$$

Since $\alpha > 1$, $m \log_2 \alpha \rightarrow \infty$ as $m \rightarrow \infty$, the difference between the left hand side and the right hand side of (3.25) tends to infinity if $m \rightarrow \infty$. Thus, it is always possible to select m and N to satisfy (3.25). $\blacksquare$

Chapter 4

Data Rate Theorem for Stabilization: i.i.d. Packet Loss

In Chapter 3, the logarithmic quantizer was shown to be optimal in the sense of approaching the minimum data rate under the assumption that there is no packet loss for the communication channel. Clearly, the loss of packet also induces information loss. The fundamental question is how the finite data rate and packet loss jointly affect the stabilizability of a linear system? This chapter is an attempt to answer this question. We model the packet loss process of the channel as an independent and identically distributed (i.i.d.) process. Then, 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 clearly 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.

This chapter is organized as follows. The problem of interest is formulated in Section 4.1. The main results are presented in Section 4.2, where we start at the investigation of the minimum data rate for single input systems and then proceed to multiple input systems. Concluding remarks are made in Section 4.3.

4.1 Problem Formulation

Consider the following discrete LTI system

$$x_{k+1} = Ax_k + Bu_k, \quad (4.1)$$

where $x_k \in \mathbb{R}^n$ is the measurable state for feedback and $u_k \in \mathbb{R}^l$ is the control input.

To make the problem interesting, $A \in \mathbb{R}^{n \times n}$ is assumed to be unstable.

Figure 4.1: Network configuration.

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 . However, 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 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

$$q_N(x) = \begin{cases} \frac{|2^{N-1}x|+0.5}{2^{N-1}}, & \text{if } -1 \leq x < 1; \\ 1 - \frac{0.5}{2^{N-1}}, & \text{if } x = 1. \end{cases} \quad (4.2)$$

The quantization intervals are labeled from left to right by $I_{2^N}(0), \dots, I_{2^N}(2^N - 1)$. Thus, if $|x| \leq M$ with M known, the quantization error induced by the above N -bit uniform quantizer is bounded as $|x - Mq_N(\frac{x}{M})| \leq \frac{M}{2^N}$.

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 present quantized signal s_k at time k is generated by allowing the encoder to access all the past and present un-quantized control input $f_0, \dots, f_k$, the past quantized symbols $s_0, \dots, s_{k-1}$ and packet reception status $\gamma_0, \dots, \gamma_{k-1}$, i.e., by defining $f_0^k = \{f_0, \dots, f_k\}$ and similarly s_0^k and γ_0^k , $s_k = \mathcal{E}_k(f_0^k, s_0^{k-1}, \gamma_0^{k-1})$, where $\mathcal{E}_k(\cdot)$ is the coder mapping at time k . Likewise, at time k , the decoder generates the control input u_k by $u_k = \mathcal{D}_k((s\gamma)_0^k, \gamma_0^k)$, where $(s\gamma)_0^k = \{s_0\gamma_0, \dots, s_k\gamma_k\}$ and $\mathcal{D}_k(\cdot)$ 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 a quantized feedback if for any finite initial state x_0 , there is a control policy relying on the quantized data such that $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k\|^2] = 0$, where the mathematical expectation operator $\mathbb{E}[\cdot]$ is taken w.r.t. the random packet loss process $\{\gamma_k\}_{k \geq 0}$.*

The problem of interest 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 Remark 3.1, 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 Main Results

In this section, we derive the minimum data rate for single input vector systems via the quantized control. Our result shows that the corresponding minimum data rate is uniquely determined by the unstable eigenvalues of the open loop matrix and the packet loss rate.

4.2.1 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. 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 and only if*

(a) *The packet loss rate is less than the threshold given in [103], namely,*

$$p < \frac{1}{\prod_{|\lambda_i(A)| \geq 1} |\lambda_i(A)|^2}. \quad (4.3)$$

(b) *The data rate R satisfies that*

$$R > R_{\inf} = \sum_{|\lambda_i(A)| \geq 1} \log_2 |\lambda_i(A)| + \frac{1}{2} \log_2 \frac{1-p}{1-p \prod_{|\lambda_i(A)| \geq 1} |\lambda_i(A)|^2}, \quad (4.4)$$

where $\lambda_i(A)$ denotes an eigenvalue of A .

Remark 4.2. *Due to the existence of packet loss, additional bits are required to counter the packet loss effect on stabilizability of the system. They are explicitly quantified by the second term of the right hand side of (4.4) 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 $0 < p < \frac{1}{\prod_{|\lambda_i(A)| \geq 1} |\lambda_i(A)|^2}$.

For the perfect channel case, i.e., $p = 0$, (4.3) is obviously satisfied and (4.4) recovers the well-known minimum data rate in [84, 107]. On the other hand, when the channel bandwidth is infinite, i.e., $R = \infty$, the data rate inequality of (4.4) is automatically satisfied and the necessary and sufficient condition for asymptotic stabilization in the mean square sense is fully characterized in (4.3). It is interesting to note that the same conclusion can be found in [28, 100] 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.

Proof of Theorem 1: 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.

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

$$x_{k+1} = \begin{bmatrix} 0 & 1 \\ -a_2 & -a_1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k, \quad (4.5)$$

where $\det(\lambda I - A) = \lambda^2 + a_1\lambda + a_2$. Clearly, $\det(A) = a_2$. Denote the i -th element of x_k by $x_k^{(i)}$, it follows from (4.5) that if the system is asymptotically stabilized in the mean square sense, then

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

Thus, the asymptotic stabilization of the system (4.5) in the mean square sense is

equivalent to that of the sub-system:

$$x_{k+1}^{(2)} = -a_1 x_k^{(2)} - a_2 x_{k-1}^{(2)} + u_k, \quad (4.6)$$

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

Observe that after the k^{th} 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 (4.6), the decoder knows that $x_{k+1}^{(2)}$ is in the set $\chi_{k+1}^- = -a_1 \chi_k - a_2 \chi_{k-1}$ before the reception of the quantized control signal s_k , where for any two Lebesgue measurable sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$, we define the sum of sets by $a\mathcal{X} + b\mathcal{Y} \triangleq \{ax + by | x \in \mathcal{X}, y \in \mathcal{Y}\}$. In addition, denote the Lebesgue measure of a measurable set $\mathcal{X} \subset \mathbb{R}^n$ by $\mu(\mathcal{X})$. Applying the Brunn-Minkowski inequality [23], we obtain that

$$\mu(\chi_{k+1}^-) \geq \mu(-a_1 \chi_k) + \mu(-a_2 \chi_{k-1}) = |a_1| \mu(\chi_k) + |a_2| \mu(\chi_{k-1}). \quad (4.7)$$

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 (4.6) and (4.7), it holds that

$$\mu(\chi_{k+1}) \geq \frac{1}{2^{R\gamma_{k+1}}} [|a_1| \mu(\chi_k) + |a_2| \mu(\chi_{k-1})] \quad (4.8)$$

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 . Noticing that $L_j \geq 0, \forall j \in \mathbb{N}$ and taking square on both sides of (4.8), the following inequalities are in force.

$$\begin{aligned} \mathbb{E}[L_{k+1}^2] &\geq \mathbb{E}\left[\frac{|a_1|^2}{2^{2R\gamma_{k+1}}}\right] \mathbb{E}[L_k^2] + \mathbb{E}\left[\frac{|a_2|^2}{2^{2R\gamma_{k+1}}}\right] \mathbb{E}[L_{k-1}^2] + 2\mathbb{E}\left[\frac{|a_2||a_1|}{2^{2R\gamma_{k+1}}}\right] \mathbb{E}[L_k L_{k-1}] \\ &\geq |a_2|^2 \mathbb{E}\left[\frac{1}{2^{2R\gamma_{k+1}}}\right] \mathbb{E}[L_{k-1}^2] = |a_2|^2 [p + (1-p)2^{-2R}] \mathbb{E}[L_{k-1}^2], \end{aligned} \quad (4.9)$$

where we have used the fact that the binary valued random variable γ_{k+1} is indepen-

dent 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 $\gamma_j, j \leq k$ [4]. Also it clearly holds that $2\{\sup |x|, x \in \mathcal{X}\} \geq \mu(\mathcal{X})$ for any Lebesgue measurable subset $\mathcal{X} \subset \mathbb{R}$, the asymptotic stabilization in the mean square sense of $x_k^{(2)}$ implies that $\liminf_{k \rightarrow \infty} \mathbb{E}[L_k^2] = 0, \forall L_0 \geq 0$. Together with the recursion of (4.9), 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.

Sufficiency: Define the subset $\mathcal{L}(A) \subseteq \mathbb{N}$ by

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

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

Divide the times $j \in \mathbb{N}$ into blocks $\{km, \dots, (k+1)m-1\}$, $k \in \mathbb{N}$, of uniform duration, where $m \in \mathcal{L}(A)$ 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_{km} . To this end, at the start of the time block $\{km, \dots, (k+1)m-1\}$, i.e., at time km , the controller generates a control law f_{km} and set $f_{km+t} = 0, t \in \{1, \dots, m-1\}$. The idea is to use the time period from km to $(k+1)m-1$ to sequentially transmit the quantized f_{km} . Accordingly, the decoder applies the decoded control input $u_{km+t} = 0, t \in \{0, 1, \dots, m-2\}$ and $u_{(k+1)m-1}$, which is the decoder's estimate of f_{km} . In this case, the down-sampled system of (4.1) with the down-sampling factor m is expressed as

$$x_{(k+1)m} = A^m x_{km} + B u_{(k+1)m-1}. \quad (4.10)$$

Due to the controllability of (A^m, B) , (4.10) can be transformed into a controllable canonical form. In particular, there exists a nonsingular matrix Q such that

the controllable canonical form of (4.10) is expressed by

$$x_{(k+1)m} = \begin{bmatrix} 0 & 1 \\ -\alpha_2 & -\alpha_1 \end{bmatrix} x_{km} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{(k+1)m-1} \quad (4.11)$$

where we abuse notation by defining $x_{km} \triangleq Qx_{km}$ and $|\alpha_2| = |\det(A)|^m$. 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

$$|\alpha_k| \leq \alpha_0 |\alpha_2|, \forall k \in \{1, 2\}. \quad (4.12)$$

As in the proof of the necessity part, if we can stabilize the last element of the state vector, denoted by $x_{km}^{(2)}$, in the mean square sense, x_{km} will be stabilized as well, which further implies the stabilization of the original system (4.1) due to $m < \infty$. Given any data rate R that is strictly greater than the minimum data rate in (4.4), i.e., $R > R_{\inf}$ and a packet loss rate p satisfying (4.3), the control signal f_{km} and the correspondingly decoded signal $u_{(k+1)m-1}$ are proposed as follows:

$$\begin{cases} u_{(k+1)m-1} = \tilde{f}_{km} q_{N_k}(f_{km}/\tilde{f}_{km}); \\ f_{km} = \begin{cases} 0, & \text{if } k < 2; \\ \alpha_1 x_{km}^{(2)} + \alpha_2 x_{(k-1)m}^{(2)}, & \text{if } k \geq 2, \end{cases} \end{cases} \quad (4.13)$$

where the random variable

$$N_k = R \sum_{j=km}^{(k+1)m-1} \gamma_j \quad (4.14)$$

represents the cumulative number of quantization levels during the time block $\{km, \dots, (k+1)m-1\}$ and $\tilde{f}_{km}$ is the scaling factor to capture f_{km} , i.e., $|f_{km}| \leq \tilde{f}_{km}$. Since f_{km} and the scaling factor $\tilde{f}_{km}$ are produced at time km , we denote $f_k \triangleq f_{km}$ and $\tilde{f}_k \triangleq \tilde{f}_{km}$ in the rest of the paper for notational simplicity.

Before proceeding to the proof of stabilizability, the recursive implementation of the quantizer is first described. At the start of each time block $\{km, \dots, (k+1)m-1\}$, the scaled $f_k/\tilde{f}_k$ will first be quantized by the uniform quantizer in (4.2) with bit number R [106]. The quantized message s_{km} will then be sent to the decoder. If the packet is received by the decoder, i.e. $\gamma_{km} = 1$, 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_{km} . Otherwise, the packet is dropped ($\gamma_{km} = 0$), the encoder and decoder agree that $f_k/\tilde{f}_k \in [-1, 1] \triangleq I_1$. Thus, after the first transmission at time km , the encoder and decoder agree on the fact that $f_k/\tilde{f}_k \in I_{2^{R\gamma_{km}}}(\cdot)$.

The remaining $m-1$ transmissions in the time block $\{km, \dots, (k+1)m-1\}$ are devoted to reducing the size of the subinterval $I_{2^{R\gamma_{km}}}(\cdot)$. Specifically, at time $km+1$, the encoder and decoder equally divide $I_{2^{R\gamma_{km}}}(\cdot)$ into 2^R subintervals, sequentially generating the partitions of $I_{2^{R\gamma_{km}}}(\cdot)$ of the quantizer $q_{2^{R\gamma_{km}+2^R}}(\cdot)$. Similarly, after the second transmission of s_{km+1} , the encoder and decoder agree on the fact that $f_k/\tilde{f}_k \in I_{2^{R\gamma_{km}+2^{R\gamma_{km}+1}}}(\cdot)$. Continuing the above process until the end of the time block $\{km, \dots, (k+1)m-1\}$, 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 (4.14). Using the above protocol, we may get an accurate estimate of f_{km} and thus a better control input $u_{(k+1)m-1}$.

Similarly, we also respectively rewrite $x_{km}^{(2)}, u_{(k+1)m-1}$ as $x_k^{(2)}, u_k$ in the sequel. Furthermore, for ease of presentation, assume that $|x_0^{(2)}| \leq \Delta_0$ and $|x_1^{(2)}| \leq \Delta_1$ and the upper bounds $\Delta_i, i \in \{0, 1\}$ are known by the encoder and decoder. This assumption can be removed by using the approach in [123]. In light of (4.11), it immediately holds that

$$x_{k+1}^{(2)} = -\alpha_1 x_k^{(2)} - \alpha_2 x_{k-1}^{(2)} + u_k, \quad k \geq 2. \quad (4.15)$$

Define the upper bound of $x_k^{(2)}$ by Δ_k , i.e., $|x_k^{(2)}| \leq \Delta_k$ and $\Delta'_k = \max\{\Delta_k, \Delta_{k-1}\}$ with $\Delta_k = 0$ if $k < 0$. The synchronization of the time-varying factors Δ_k and Δ'_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 $\Delta_i, i \in \{0, 1\}$ is transparent to the decoder and encoder by assumption.

Since $|\sum_{j=0}^1 \alpha_{j+1} x_{k-j}^{(2)}| \leq \sum_{j=0}^1 |\alpha_{j+1}| \cdot |x_{k-j}^{(2)}| \leq 2\alpha_0 |\alpha_2| \Delta'_k \triangleq \tilde{f}_k$ by (4.12), inserting the control law in (4.13) to the system (4.15) obtains that

$$|x_{k+1}^{(2)}| = \tilde{f}_k \left| \frac{f_k}{\tilde{f}_k} - q_{N_k} \left(\frac{f_k}{\tilde{f}_k} \right) \right| \leq \frac{\tilde{f}_k}{N_k} = \frac{2\alpha_0 |\alpha_2|}{N_k} \Delta'_k \triangleq \Delta_{k+1}. \quad (4.16)$$

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

Case1: If $\Delta_{k+1} > \Delta_k$, then $\Delta'_{k+1} = \Delta_{k+1} = \frac{2\alpha_0|\alpha_2|}{N_k}\Delta'_k$.

Case2: Otherwise, if $\Delta_{k+1} \leq \Delta_k$, it is obvious that $\Delta'_{k+1} = \Delta_k \leq \Delta'_k$. Under this situation, repeating the derivation of (4.16) yields that $|x_{k+2}^{(2)}| \leq \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\Delta'_{k+1} \leq \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\Delta'_k \triangleq \Delta_{k+2}$. Furthermore, if $\Delta_{k+2} > \Delta_{k+1}$, $\Delta'_{k+2} = \Delta_{k+2} = \frac{2\alpha_0|\alpha_2|}{N_{k+1}}\Delta'_k$. Otherwise, $\Delta_{k+2} \leq \Delta_{k+1}$, we conclude that $\Delta'_{k+2} = \Delta_{k+1} = \frac{2\alpha_0|\alpha_2|}{N_k}\Delta'_k$.

Thus, at any circumstance, there must exist $i_k \in \{1, 2\}, l_k \in \{0, 1\}$ such that $\Delta'_{k+i_k} = \frac{2\alpha_0|\alpha_2|}{N_{k+l_k}}\Delta'_k$. Next, select a subsequence from $\{\Delta'_k\}$ as follows: $k_1 = 1, k_{j+1} = k_j + i_{k_j}$. Consequently, we have the recursion $\Delta'_{k_{j+1}} = \frac{2\alpha_0|\alpha_2|}{N_{k_j+l_{k_j}}}\Delta'_{k_j}, \forall j \in \mathbb{N}$. Taking square and then expectation on both sides of the equality yields that

$$\begin{aligned} \mathbb{E}[(\Delta'_{k_{j+1}})^2] &= (2\alpha_0)^2(|\alpha_2|)^2\mathbb{E}\left[\frac{1}{N_{k_j+l_{k_j}}^2}\right]\mathbb{E}[(\Delta'_{k_j})^2] \\ &= (2\alpha_0)^2[|\det A|^2(p + (1-p)2^{-2R})]^m\mathbb{E}[(\Delta'_{k_j})^2] \\ &\triangleq \eta\mathbb{E}[(\Delta'_{k_j})^2]. \end{aligned} \quad (4.17)$$

Here the first equality is due to that $N_{k_j+l_{k_j}}$ is independent of Δ'_{k_j} since the packet loss process is assumed to be an i.i.d. process while the second equality uses the fact that $\mathbb{E}[2^{-2R}\sum_{j=km}^{(k+1)m-1}\gamma_j] = (\mathbb{E}[2^{-2R}\gamma_1])^m$ and $|\alpha_2| = |\det A|^m$.

The remaining problem is to select m from $\mathcal{L}(A)$ to make $\eta = (2\alpha_0)^2[|\det A|^2(p + (1-p)2^{-2R})]^m < 1$. In particular, in light of (4.4), one can test that $|\det A|^2(p + (1-p)2^{-2R}) < 1$. Thus, it is possible to choose a sufficient large $m \in \mathcal{L}(A)$, i.e.,

$$m > -\frac{2\log_2(2\alpha_0)}{\log_2[|\det A|^2(p + (1-p)2^{-2R})]}, \quad (4.18)$$

such that $\eta < 1$, which immediately infers that $\lim_{j \rightarrow \infty} \mathbb{E}[(\Delta'_{k_j})^2] = 0$ by (4.17).

Observe that $\Delta_v \leq \min\{\Delta'_{v-1}, \Delta'_v, \Delta'_{v+1}\}, \forall v \in \mathbb{N}$ and $k_{j+1} - k_j = i_{k_j} \leq 2$, there must exist a $k_{j_v} \in \{v-1, v, v+1\}$ such that $\Delta_v \leq \Delta'_{k_{j_v}}$. Together with the fact that $v \rightarrow \infty$ implies $k_{j_v} \rightarrow \infty$, we obtain $\lim_{v \rightarrow \infty} \mathbb{E}[\Delta_v^2] \leq \lim_{v \rightarrow \infty} \mathbb{E}[(\Delta'_{k_{j_v}})^2] =$

Figure 4.2: $R = 2$

$$\lim_{k_{jv} \rightarrow \infty} \mathbb{E}[(\Delta'_{k_{jv}})^2] = 0.$$

Then, it follows that $\lim_{k \rightarrow \infty} \mathbb{E}[(x_k^{(2)})^2] \leq \lim_{k \rightarrow \infty} \mathbb{E}[\Delta_k^2] = 0$, which further implies that $\lim_{k \rightarrow \infty} \mathbb{E}[||x_k||^2] = 0$.

Remark 4.3. *Similar conditions as in Theorem 4.1 have been obtained for scalar systems in [73, 78]. 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., $\lim_{k \rightarrow \infty} \mathbb{E}[||x_k||^m] = 0$. Nevertheless, the idea is essentially the same as the case of mean square stabilization ($m = 2$).*

Example 4.1. *We use a simple example to verify our result. Consider following system:*

$$A = \begin{bmatrix} 1.5 & 0.5 \\ 1 & -0.5 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

It is clear that A has two eigenvalues, e.g., $\lambda_1 = 1.7247$ and $\lambda_2 = -0.7247$. Let the packet dropout rate $p = 0.3$. By Theorem 4.1, $R > 2.14$. By using a data rate $R = 2$, Fig. 4.2 shows divergence of the close-loop system. However, if $R = 3$ and the controller is designed by the technique in the proof of Theorem 4.1, Fig. 4.3 illustrates that the system is mean square stabilized.

4.2.2 Multiple Input Case

Consider system (4.1) with multiple inputs, i.e. $u_k \in \mathbb{R}^l$ with $l > 1$, and satisfying Assumption 4.1. Without loss of generality, assume that the control input matrix

Figure 4.3: $R = 3$

$B \in \mathbb{R}^{n \times l}$ has full column rank, namely, $\text{rank}(B) = l$.

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 subsection, 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., $R = \sum_{j=1}^l R_j$. The following lemma is needed.

Lemma 4.1. *Suppose that the sequence $\{z_k\} \subset \mathbb{R}$ is recursively computed by the formula $z_{k+1} = (1 - a_k)z_k + b_k$, $\forall k \in \mathbb{N}$ and $a_k \in [0, 1)$, $\sum_{k=0}^{\infty} a_k = \infty$, $|z_0| < \infty$. Then if $\lim_{k \rightarrow \infty} \frac{b_k}{a_k}$ exists, we have that $\lim_{k \rightarrow \infty} z_k = \lim_{k \rightarrow \infty} \frac{b_k}{a_k}$.*

Proof. Define $\phi(k+1, j) = (1 - a_k)\phi(k, j)$, $\phi(j, j) = 1, \forall k \geq j$, it is easy to verify that $\sum_{j=0}^k \phi(k+1, j+1)a_j = 1 - \prod_{j=0}^k (1 - a_j) \rightarrow 1$ as $k \rightarrow \infty$. 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 $\phi(k, j) \rightarrow 0$ as $k \rightarrow \infty, \forall j \in \mathbb{N}$, the result is established by the Toeplitz Lemma [4, pp.235-236]. ■

We now apply the Wonham decomposition [20] 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

$$\bar{A} = \begin{bmatrix} A_1 & A_{12} & \cdots & A_{1l} \\ 0 & A_2 & \cdots & A_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_l \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} b_1 & b_{12} & \cdots & b_{1l} \\ 0 & b_2 & \cdots & b_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_l \end{bmatrix}, \quad (4.19)$$

where the matrices $A_i \in \mathbb{R}^{n_i \times n_i}$ and $b_i \in \mathbb{R}^{n_i}, i \in \{1, \dots, l\}$ satisfy $\sum_{i=1}^l n_i = n$. 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*

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

$$p < \frac{1}{\max_{i \in \{1, \dots, l\}} \{\prod_j |\lambda_j(A_i)|^2\}}. \quad (4.20)$$

(2) *The data rate R satisfies that*

$$R > \sum_{|\lambda_i(A)| \geq 1} \log_2 |\lambda_i(A)| + \frac{1}{2} \sum_{i=1}^l \log_2 \frac{1-p}{1-p \prod_j |\lambda_j(A_i)|^2}. \quad (4.21)$$

Proof. Due to the page limitation, only the essential steps of the proof are provided.

For brevity, assume $l = 2$ in (4.19) 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 m , which is determined by the given data rate and its ability to preserve the controllability of $(\bar{A}^m, \bar{B})$, to down-sample the system, we get $x_{(k+1)m} = \bar{A}^m x_{km} + \bar{B}u_{(k+1)m-1}$. By denoting $\bar{A}^m = \begin{bmatrix} A_1^m & \bar{A}_{12} \\ 0 & A_2^m \end{bmatrix}$ and partitioning the state vector $x_{km} \triangleq [(x_k^{(1)})^T, (x_k^{(2)})^T]^T$ in conformity with $\bar{A}^m$,

the subsystems are described by

$$x_{k+1}^{(1)} = A_1^m x_k^{(1)} + b_1 u_k^{(1)} + \bar{A}_{12} x_{km}^{(2)} + b_{12} u_k^{(2)}, \quad (4.22)$$

$$x_{k+1}^{(2)} = A_2^m x_k^{(2)} + b_2 u_k^{(2)}, \quad (4.23)$$

where the control vector is written as $u_{(k+1)m-1} \triangleq [u_k^{(1)}, u_k^{(2)}]^T$.

Given any data rate R and the packet loss rate $p \in (0, 1)$ respectively satisfying (4.20) and (4.21), we separate R into $R = R_1 + R_2$ such that $R_i > \log_2 |\det(A_i)| + \frac{1}{2} \log_2 \frac{1-p}{1-p|\det A_i|^2}$, $i \in \{1, 2\}$. Repeating the process in the sufficiency proof of Theorem 4.1 on the subsystem (4.23), we can stabilize $x_k^{(2)}$ in the mean square sense using the data rate R_2 , i.e. $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k^{(2)}\|^2] = 0$. As for the subsystem (4.22), 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 (4.17): $\mathbb{E}[(\Delta'_{k_{j+1}})^2] \leq \eta \mathbb{E}[(\Delta'_{k_j})^2] + c \mathbb{E}[\|x_{k_j}^{(2)}\|^2]$, where $c \in \mathbb{R}$ is a constant, $\eta < 1$ and Δ'_k has the same meaning as in the proof of the sufficiency of Theorem 4.1. Together with Lemma 4.1 in the appendix, it gives that $\lim_{j \rightarrow \infty} \mathbb{E}[(\Delta'_{k_j})^2] = 0$, which can further imply $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k^{(1)}\|^2] = 0$. Due to $m < \infty$, it finally holds that $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k\|^2] = 0$. ■

4.3 Summary

Motivated by control problems over lossy communication channels, this chapter has derived the necessary and sufficient condition for the asymptotic stabilization of single input discrete LTI systems in the mean square sense over a lossy communication channel 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.

Chapter 5

Data Rate Theorem for Stabilization: Markovian Packet Loss

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 contains existing results on data rate and packet loss rate for stabilization of linear scalar systems as special cases and provides a means for better bandwidth utilization by jointly considering bits per sample and an effective sampling. Necessary and sufficient conditions 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.

This chapter is organized as follows. The problem formulation is described in Section 5.1. Some important preliminary results are provided in Section 5.2. The study of the minimum data rate for scalar systems is carried out in Section 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. Necessary and sufficient conditions for vector systems are respectively studied in Section 5.4. Conclusion remarks are drawn in Section 5.5.

5.1 Problem Formulation

Consider the following stochastic linear time-invariant system

$$x_{k+1} = Ax_k + Bu_k + w_k, \quad k = 0, 1, \dots \quad (5.1)$$

where $x_k \in \mathbb{R}^n$ is the measurable state, $u_k \in \mathbb{R}^l$ is the input, and $w_k \in \mathbb{R}^n$ is the stochastic disturbance of zero mean. 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 $\sup_{k \in \mathbb{N}} \mathbb{E}[\|w_k\|^{2+\delta}] < \varpi$. Also, x_0 and $w_k, \forall k \in \mathbb{N}$ are mutually independent and have probability densities. Moreover, the differential entropy [23] of x_0 , denoted by $h(x_0)$, exists and $\exists \Delta > 0$ such that $\inf_{k \in \mathbb{N}} e^{2h(w_k)} \geq \Delta$. It is worth mentioning that the above conditions have been considered in [84] and contain the general additive white Gaussian noise assumption as a special case. To make the problem interesting, we focus on unstable systems and that (A, B) is a stabilizable pair.

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 acknowledgment can be completed within one sampling interval. As in [51] and [120], 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 [103] 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. 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

Figure 5.1: Network configuration

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

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 1, 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 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

Denote $(\Omega, \mathcal{F}, \mathbb{P})$ the common probability space for all random variables in the chapter and let $\mathcal{F}_k \triangleq \sigma(\gamma_0^k) \subset \mathcal{F}$ be an increasing sequence of σ -fields (filtration) generated by random variables $\{\gamma_0, \dots, \gamma_k\}$. 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 $\gamma_0 = 1$ [120]. Then, $t_0 = 1^1$ and $t_k, k \geq 1$ is precisely defined by

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

By the ergodic property of the Markov chain $\{\gamma_k\}_{k \geq 0}$, t_k is finite *a.e.* for any $k \in \mathbb{N}$ [51]. 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. [120] *The sojourn times $\{t_k^*\}_{k > 0}$ are independent and identically*

¹The encoder can only know whether the current packet is lost or received at the next sampling time. For example, if the packet generated at time k is successfully delivered, the decoder will receive the packet and send a reception acknowledge ($\gamma_k = 1$) to the encoder. Since this process is assumed to be completed within one sampling interval, the encoder will know the packet reception status at time $k + 1$.

distributed. Furthermore, the distribution of t_1^* is explicitly expressed as

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

We shall exploit this fact to establish our results. Unlike the technique employed in [73, 78, 124], we develop a new framework by down sampling the system of (5.1) with the sampling interval equal to t_k^* . Here $\{t_k^*\}$ can be interpreted as a “communication logic” to trigger the transmission of the packet.

5.3 Scalar Systems

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

$$x_{k+1} = \lambda x_k + bu_k + w_k, \quad (5.6)$$

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:

$$x_{k+1} = \lambda x_k + bu_k, \quad (5.7)$$

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

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,*

$$q > 1 - 1/|\lambda|^2; \quad (5.8)$$

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

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

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

Remark 5.1.

(1) *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 mean square sense.*

(2) *Neglecting quantization effects, i.e., $R = \infty$, the inequality of (5.10) is automatically satisfied. It is interesting to note that our condition recovers the result in [45, 51].*

(3) *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 [73, 78, 124], see Theorem 4.1.

- (4) In light of (5.10), the larger the magnitude of the unstable mode, the more bits are needed to compensate the effect of packet loss. As mentioned in Section 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 $p \rightarrow 0$ and $q \rightarrow 1$, the minimum data rate in (5.7) converges to the well-known data rate theorem for stabilization of a linear system [82, 107, 123].

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

Lemma 5.2. [82] 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 $c_\omega(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ with the number of quantization levels not greater than $\omega \in \mathbb{N}$. Then, $\forall \theta \in (\frac{1}{3}, 1), \forall \omega \in \mathbb{N}, \mathbb{E}[\|x - c_\omega(x)\|^2] \geq \beta \omega^{-2} \|P_x\|_\theta^{2\theta/(1-\theta)}$, where $\beta > 0$ is a parameter determined by θ and $\|P_x\|_\theta^{2\theta/(1-\theta)}$ is Rényi differential entropy power of order θ [23], i.e., $\|P_x\|_\theta^{2\theta/(1-\theta)} = (\int_{\mathbb{R}} \|P_x(x)\|^\theta dx)^{\frac{2}{1-\theta}}$.

Proof [of Theorem 5.1] **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 [82], 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

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\lambda^k x_0 - c_{\omega_k}^*(x_0)\|^2] = 0. \quad (5.11)$$

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

$$\begin{aligned} \mathbb{E}[\|\lambda^k x_0 - c_{\Gamma_k}^*(x_0)\|^2] &= \mathbb{E}[\mathbb{E}[\|\lambda^k x_0 - c_{\Gamma_k}^*(x_0)\|^2 | \Gamma_k]] \\ &\geq \mathbb{E}[\beta \Gamma_k^{-2} \|P_{\lambda^k x_0}\|_\theta^{2\theta/(1-\theta)}] = \beta \mathbb{E}[\Gamma_k^{-2}] \left(\int_{\mathbb{R}} \|P_{\lambda^k x_0}(x)\|^\theta dx \right)^{\frac{2}{1-\theta}} \\ &= \beta \mathbb{E}[\Gamma_k^{-2}] \left(\int_{\mathbb{R}} \|P_{x_0}(\lambda^{-k} x) \lambda^{-k}\|^\theta dx \right)^{\frac{2}{1-\theta}} \\ &= \beta \mathbb{E}[\Gamma_k^{-2}] (\lambda^k \int_{\mathbb{R}} \|P_{x_0}(y) \lambda^{-k}\|^\theta dy)^{\frac{2}{1-\theta}} = \beta \|P_{x_0}\|_\theta^{2\theta/(1-\theta)} \mathbb{E}[\frac{\lambda^{2k}}{\Gamma_k^2}], \end{aligned} \quad (5.12)$$

where the inequality is due to Lemma 5.2 and the change of integration variable $y = \lambda^{-k}x$ is performed in the second last equality. Let $\xi_k = \lambda^k/\Gamma_k, \forall k \in \mathbb{N}$, it leads to that:

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

By (5.11), it follows from (5.12) that $\lim_{k \rightarrow \infty} \mathbb{E}[\frac{\lambda^{2k}}{\Gamma_k^2}] = 0$ 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

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

By Theorem 4 of [120], 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 stopping times $\{t_k - 1\}$, i.e., $\lim_{k \rightarrow \infty} \mathbb{E}[\|\xi_k\|^2] = 0 \Leftrightarrow \lim_{k \rightarrow \infty} \mathbb{E}[\|\xi_{t_k-1}\|^2] = 0$. 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 $\mathbb{E}[\|\xi_{t_k-1}\|^2] = \mathbb{E}[\frac{|\lambda|^{2 \sum_{j=0}^{k-1} t_{j+1}^*}}{2^{2Rk}} |\xi_0|^2] = \frac{1}{2^{2R}} (\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}])^k$. Consequently, a necessary condition to make $\lim_{k \rightarrow \infty} \mathbb{E}[\|\xi_k\|^2] = 0$ is that $\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}] < 1$ by the equivalence property. By Lemma 5.1, the proof of the necessity is completed by the following arguments:

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

Sufficiency: In control strategies to be developed in this subsection, a uniform quantizer in (4.2) 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:

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

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

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

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

$$l_1 = \frac{|\lambda|}{2^R} l_0; \quad (5.18)$$

$$l_{j+1} = |\lambda| l_j, j \in \{t_k, t_k + 1, \dots, t_{k+1} - 2\}; \quad (5.19)$$

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 $\frac{x_{t_k} - \hat{x}_{t_k}}{l_{t_k}}$, by a R -bit uniform quantizer. The quantizer output $\sigma_{t_k} \triangleq q_R\left(\frac{x_{t_k} - \hat{x}_{t_k}}{l_{t_k}}\right)$ 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 acknowledgement ($\gamma_{t_k} = 0$) 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 acknowledgement from the decoder

Figure 5.2: Communication protocol.

$(\gamma_{t_{k+1}-1} = 1)$ 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 $\tilde{x}_j = x_j - \hat{x}_j, \forall j \in \mathbb{N}$, the dynamical equation governing the error evolution is given by

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

It can be shown that the quantizer does not overflow at all times, i.e., $|\tilde{x}_j| \leq l_j, \forall j \in \mathbb{N}$. In fact, it obviously holds for $j = 0, 1$ since $|x_0| \leq l_0$ and $|\tilde{x}_1| \leq \frac{|\lambda|}{2^R} l_0 = l_1$. Assume that $\exists k \geq 0, |\tilde{x}_{t_k}| \leq l_{t_k}$, then $|\tilde{x}_{t_{k+1}}| = |\lambda|^{t_{k+1}^*} l_{t_k} |\frac{\tilde{x}_{t_k}}{l_{t_k}} - q(\frac{\tilde{x}_{t_k}}{l_{t_k}})| \leq \frac{|\lambda|^{t_{k+1}^*}}{2^R} l_{t_k} = l_{t_{k+1}}$. Further, for any $j \in \{t_k, t_k + 1, \dots, t_{k+1} - 2\}$, $|\tilde{x}_j| = |\lambda|^{j-t_k} |\tilde{x}_{t_k}| \leq |\lambda|^{j-t_k} l_{t_k} = l_j$. Thus, by induction, we have that $|\tilde{x}_j| \leq l_j, \forall j \in \mathbb{N}$. 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 $\lim_{k \rightarrow \infty} \mathbb{E}[l_{t_k}^2] = \lim_{k \rightarrow \infty} \mathbb{E}[\prod_{j=0}^{k-1} \frac{|\lambda|^{2t_{j+1}^*}}{2^{2R}}] l_1^2 = \lim_{k \rightarrow \infty} (\mathbb{E}[\frac{|\lambda|^{2t_1^*}}{2^{2R}}])^k l_1^2 = l_1^2 \lim_{k \rightarrow \infty} \eta^k = 0$, 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

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

$$= \mathbb{E}[t_1^* |\lambda|^{2t_1^*}] l_1^2 \sum_{k=0}^{\infty} \eta^k = \frac{\mathbb{E}[t_1^* |\lambda|^{2t_1^*}] l_1^2}{1 - \eta} < \infty, \quad (5.22)$$

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 $x_0, t_1^*, \dots, t_k^*$ and l_{t_k} is adapted to $\sigma(x_0, t_1^*, \dots, t_k^*)$ in (5.21).

Together with that $|\tilde{x}_k| \leq l_k, \forall k \in \mathbb{N}$, it immediately follows from (5.22) that $\lim_{k \rightarrow \infty} \mathbb{E}[|\tilde{x}_k|^2] \leq \lim_{k \rightarrow \infty} \mathbb{E}[l_k^2] = 0$. By iteration of (5.7) and substituting the control signal $u_k = \mu \hat{x}_k$ into (5.7), it yields that $x_{k+1} = \lambda x_k + b\mu \hat{x}_k = (\lambda + b\mu)x_k + b\mu \tilde{x}_k = (\lambda + b\mu)^{k+1} x_0 + \sum_{j=0}^k (\lambda + b\mu)^{k-j} b\mu \tilde{x}_j$. Jointly with the inequality $(\mathbb{E}[|x+y|^2])^{1/2} \leq (\mathbb{E}[|x|^2])^{1/2} + (\mathbb{E}[|y|^2])^{1/2}$, we finally obtain that $\lim_{k \rightarrow \infty} \mathbb{E}[|x|_{k+1}^2] \leq [\lim_{k \rightarrow \infty} \sum_{j=0}^k |\lambda + b\mu|^{k-j} |b\mu| (\mathbb{E}[|\tilde{x}_j|^2])^{1/2}]^2 = 0$ by the Toeplitz lemma [4, Sec 6.1.2] and the result that $\lim_{k \rightarrow \infty} (\mathbb{E}[|\tilde{x}_k|^2])^{1/2} = 0$. ■

5.3.2 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,*

$$q > 1 - 1/|\lambda|^2; \quad (5.23)$$

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

$$R > \log_2 |\lambda| + \frac{1}{2} \log_2 [1 + \frac{p(|\lambda|^2 - 1)}{1 - (1 - q)|\lambda|^2}]. \quad (5.24)$$

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 [78], 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 [23] that Φ_k is a lower bound of the second moment of the state x_k , i.e., $\mathbb{E}[|x_k|^2] \geq \Phi_k, \forall k \in \mathbb{N}$. The following lemma is needed to prove 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 [78], the assertion is straightforward. ■

Proof of Theorem 5.2

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

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

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 [23]. The first inequality follows from the entropy power inequality [23] 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 is the result of Lemma 5.3. Since $S_{k-1} \leftrightarrow (S_{k-2}, \gamma_{k-1}) \leftrightarrow$

γ_k forms a Markov chain, we can similarly derive that

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

Inserting (5.26) into (5.25) leads to that

$$\begin{aligned}
& \mathbb{E}_{S_k} [e^{2h(x_{k+1}|S_k=s_k)}] \\
& \geq \mathbb{E}_{\gamma_{k-1}, \gamma_k} \left[\frac{|\lambda|^4}{2^{2R(\gamma_{k-1}+\gamma_k)}} \mathbb{E}_{S_{k-1}|\gamma_{k-1}, \gamma_k} [e^{2h(x_{k-1}|S_{k-2}=s_{k-2})}] \right] + (\mathbb{E}_{\gamma_k} [\frac{|\lambda|^2}{2^{2R\gamma_k}}] + 1) \Delta \\
& \geq \mathbb{E}_{\gamma_1, \dots, \gamma_k} \left[\frac{|\lambda|^{2k}}{2^{2R(\gamma_1+\dots+\gamma_k)}} \mathbb{E}_{S_1|\gamma_1, \dots, \gamma_k} [e^{2h(x_1|S_0=s_0)}] \right] + \Delta \left[\sum_{j=2}^k \mathbb{E}_{\gamma_j, \dots, \gamma_k} [\frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\dots+\gamma_k)}}] + 1 \right] \\
& \geq \mathbb{E} \left[\frac{|\lambda|^{2(k+1)}}{2^{2R(\gamma_0+\gamma_1+\dots+\gamma_k)}} \right] e^{2h(x_0)} + \Delta \left[\sum_{j=1}^k \mathbb{E} \left[\frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\dots+\gamma_k)}} \right] + 1 \right],
\end{aligned} \tag{5.27}$$

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:

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

Then, ξ_{k+1} is written as $\xi_{k+1} = \frac{|\lambda|^{2(k+1)} e^{2h(x_0)}}{2^{2R(\gamma_0+\gamma_1+\dots+\gamma_k)}} + \Delta \left[\sum_{j=1}^k \frac{|\lambda|^{2(k-j+1)}}{2^{2R(\gamma_j+\dots+\gamma_k)}} + 1 \right]$. Associated with the system (5.28), we introduce the following notation:

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

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

$$\begin{aligned}
\alpha_{k+1}^j &= \sum_{i=0}^1 \mathbb{E} \left[\left(\frac{|\lambda|^2}{2^{2Ri}} \xi_k + \Delta \right) 1_{\{\gamma_{k+1}=j \cap \gamma_k=i\}} \right] \\
&= \sum_{i=0}^1 \mathbb{E} \left[\mathbb{E} \left[\left(\frac{|\lambda|^2}{2^{2Ri}} \xi_k + \Delta \right) 1_{\{\gamma_{k+1}=j \cap \gamma_k=i\}} \middle| \mathcal{F}_k \right] \right]
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^1 \frac{|\lambda|^2}{2^{2Ri}} \mathbb{E}[\xi_k 1_{\{\gamma_k=i\}}] \mathbb{E}[1_{\{\gamma_{k+1}=j \cap \gamma_k=i\}} | \mathcal{F}_k] + \Delta \mathbb{E}[1_{\{\gamma_{k+1}=j \cap \gamma_k=i\}}] \\
&= \sum_{i=0}^1 \frac{|\lambda|^2 p_{ij}}{2^{2Ri}} \alpha_k^i + \Delta \sum_{i=0}^1 p_{ij} \pi_k^i,
\end{aligned} \tag{5.30}$$

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

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

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

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

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

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

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, $\pi_k^j \rightarrow \frac{p^{1-j} q^j}{p+q}, \forall j \in \{0, 1\}$ as $k \rightarrow \infty$. Then, $\exists \psi^j > 0, \psi_k^j \rightarrow \psi^j > 0, \forall j \in \{0, 1\}$ as $k \rightarrow \infty$. In view of (5.31), it follows that the spectral radius of $\mathcal{A}$ is strictly less than one since otherwise, $\lim_{k \rightarrow \infty} \|\alpha_k\|_1 = \infty$. Thus, letting $\Delta = 0$ in (5.31), we obtain $\lim_{k \rightarrow \infty} \|\alpha_k\|_1 = 0$. Again by (5.27), it yields that

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

As in the proof of the necessity of Theorem 5.1, a necessary condition for (5.32) to hold is that $\mathbb{E}\left[\frac{|\lambda|^{2t_1^*}}{2^{2R}}\right] < 1$. The rest of proof follows from that of Theorem 5.1.

Sufficiency: We adopt the adaptive quantizer developed in [84] 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 $I_R(0), \dots, I_R(2^R - 1)$. 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} , $i \in \{2, \dots, R-1\}$ 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 $\sigma \in \{1, \dots, 2^R - 2\}$, 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 $x \in I_R(\sigma)$, $\sigma \in \{1, \dots, 2^R - 2\}$, 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 $I_i(\sigma)$, $\sigma \in \{1, \dots, 2^i - 2\}$ 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 [84]. 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 $M_\delta[x, r] \geq \mathbb{E}[|x|^2]$ [84]. A fundamental property of the above quantizer is given below.

Lemma 5.4. [84, 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 $\sigma \in \{0, \dots, 2^R - 1\}$ 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 $\forall n \in \mathbb{N}$,

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

Proof. It is a standard result in elementary inequality theory and can be easily verified. ■

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 $\{km, \dots, (k+1)m - 1\}, \forall k \in \mathbb{N}$ 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:

$$\begin{aligned} \hat{x}_0 &= \hat{x}_1 = 0; \\ \hat{x}_{j+1} &= (\lambda + b\mu)\hat{x}_j, j \in \{t_{km}, t_{km} + 1, \dots, t_{(k+1)m} - 2\}; \\ \hat{x}_{t_{(k+1)m}} &= \lambda^{t_{(k+1)m} - t_{km}} [\hat{x}_{t_{km}} + l_k Q_{mR}(\frac{x_{t_{km}} - \hat{x}_{t_{km}}}{l_k})] + \sum_{j=t_{km}}^{t_{(k+1)m} - 1} \lambda^{t_{(k+1)m} - j - 1} b\mu \hat{x}_j, \end{aligned}$$

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 $Q_{mR}(\cdot)$ works as follows. At random times $t_{km}, \forall k \in \mathbb{N}$, the encoder quantizes the normalized innovation, denoted as $\frac{x_{t_{km}} - \hat{x}_{t_{km}}}{l_k}$, 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_{km} , the decoder will receive the packet during the time interval $[t_{km+1} - 1, t_{km+1})$ 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 $\frac{x_{t_{km}} - \hat{x}_{t_{km}}}{l_k} \in I_R(\cdot)$ at time t_{km+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 $(t_{km+2} - 1, t_{km+2}]$, the encoder and decoder agree on that $\frac{x_{t_{km}} - \hat{x}_{t_{km}}}{l_k} \in I_{R+R}(\cdot)$. Continuing the same steps and after receiving the m -th packet reception acknowledgement in the time interval $[t_{(k+1)m} - 1, t_{(k+1)m})$, the encoder and decoder agree on the fact that $\frac{x_{t_{km}} - \hat{x}_{t_{km}}}{l_k}$ belongs to one of the subintervals $I_{mR}(\cdot)$. Since $|\lambda|^{2+\delta}(1-q) < 1$, select an $\epsilon > 1$ such that $|\lambda|^{(2+\delta)\epsilon}(1-q) < 1$. Then, it follows from Lemma 5.1 that $\mathbb{E}[|\lambda|^{(2+\delta)\epsilon t_1^*}] < \infty$. Using C_r -inequality in Lemma 5.5, we obtain that $\forall r \geq 0$,

$$\begin{aligned} \mathbb{E}[(t_m - t_0)^r] &= \mathbb{E}\left[\left(\sum_{j=1}^m t_j^*\right)^r\right] \leq (1 + m^{r-1})\mathbb{E}\left[\sum_{j=1}^m (t_j^*)^r\right] \\ &= m(1 + m^{r-1})\mathbb{E}[(t_1^*)^r] < \infty. \end{aligned} \quad (5.33)$$

Now, define $f(x) = \frac{|\lambda|^{(2+\delta)x}}{x-1-\delta}$ and choose $\epsilon' > 1$ to make $1/\epsilon' + \frac{1}{\epsilon} = 1$. By using the Hölder inequality it follows that

$$\begin{aligned} \mathbb{E}[(t_m - t_0)f(t_m - t_0)] &= \mathbb{E}[(t_m - t_0)^{2+\delta}|\lambda|^{(2+\delta)(t_m - t_0)}] \\ &\leq (\mathbb{E}[(t_m - t_0)^{(2+\delta)\epsilon'}])^{1/\epsilon'} (\mathbb{E}[|\lambda|^{(2+\delta)\epsilon t_1^*}])^{m/\epsilon} < \infty. \end{aligned} \quad (5.34)$$

Let $g_{t_{km}} \triangleq \sum_{j=t_{km}}^{t_{(k+1)m}-1} \lambda^{t_{(k+1)m}-j-1} w_j$. Then, the $(2+\delta)$ -th absolute moment of $g_{t_{km}}$ is uniformly bounded. Precisely,

$$\mathbb{E}[|g_{t_{km}}|^{2+\delta}] \leq \mathbb{E}[f(t_{(k+1)m} - t_{km}) \sum_{j=t_{km}}^{t_{(k+1)m}-1} |w_j|^{2+\delta}] \quad (5.35)$$

$$\leq \varpi \mathbb{E}[(t_{(k+1)m} - t_{km})f(t_{(k+1)m} - t_{km})] \quad (5.36)$$

$$= \varpi \mathbb{E}[(t_m - t_0)f(t_m - t_0)] \triangleq \alpha^{2+\delta}. \quad (5.37)$$

In the above, we applied the C_r inequality of Lemma 5.5 in (5.35) and (5.36) was

obtained in view of the assumption that the channel variation is independent of the noise process while (5.37) is due to that $t_{(k+1)m} - t_{km} = t_{km+1}^* + \dots + t_{(k+1)m}^*$ has the same distribution as that of $t_m - t_0 = t_1^* + \dots + t_m^*$ by Lemma 5.1. By using the Hölder inequality [4], it can be shown that $\mathbb{E}[|g_{t_{km}}|^2] \leq (\mathbb{E}[|g_{t_{km}}|^{2+\delta}])^{\frac{2}{2+\delta}} \leq \alpha^2$. The scaling coefficient $\{l_k\}$ with $l_0 = \alpha$ is updated as follows

$$l_{k+1} = \max\{\alpha, l_k |\lambda|^{t_{(k+1)m} - t_{km}} \kappa_{mR}(\sigma_{t_{km}})\}. \quad (5.38)$$

The proof of [84] is extended to our case for the proof of the stability of the error dynamics in random times t_{km} . Define the estimation error by $\tilde{x}_j = x_j - \hat{x}_j$, the error dynamics is governed by

$$\begin{aligned} \tilde{x}_{t_{(k+1)m}} &= \lambda^{t_{(k+1)m} - t_{km}} [\tilde{x}_{t_{km}} - l_k Q_{mR}(\frac{\tilde{x}_{t_{km}}}{l_k})] + g_{t_{km}}, \\ \tilde{x}_{j+1} &= \lambda \tilde{x}_j + w_j, j \in \{t_{km}, t_{km} + 1, \dots, t_{(k+1)m} - 2\}. \end{aligned} \quad (5.39)$$

Define $\theta_k = \mathbb{E}[l_k^2 + |\tilde{x}_{t_{km}}|^{2+\delta} l_k^{-\delta}]$ and let $\phi \triangleq 2^{1+\delta} > 1$. In view of the error dynamics in (5.39), one can easily derive the following result:

$$|\tilde{x}_{t_{(k+1)m}}|^{2+\delta} \leq \phi (|\lambda|^{t_{(k+1)m} - t_{km}} |^{2+\delta} |\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta} + |g_{t_{km}}|^{2+\delta}). \quad (5.40)$$

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

$$\begin{aligned} \mathbb{E}[|\tilde{x}_{t_{(k+1)m}}|^{2+\delta} l_{k+1}^{-\delta}] &\leq \phi \mathbb{E}[\frac{\lambda^{2(t_{(k+1)m} - t_{km})} |\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta}}{[l_k \kappa_{mR}(\sigma_{t_{km}})]^\delta} + \frac{g_{t_{km}}^{2+\delta}}{l_{k+1}^\delta}] \\ &= \phi ([\mathbb{E}[\lambda^{2t_1^*}]]^m \mathbb{E}[\frac{|\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta}}{[l_k \kappa_{mR}(\sigma_{t_{km}})]^\delta}] + \mathbb{E}[\frac{g_{t_{km}}^{2+\delta}}{l_{k+1}^\delta}]) \\ &\leq \phi (\vartheta^m \mathbb{E}[\frac{|\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta}}{[l_k \kappa_{mR}(\sigma_{t_{km}})]^\delta}] + \mathbb{E}[\frac{g_{t_{km}}^{2+\delta}}{\alpha^\delta}]) \\ &\leq \phi (\vartheta^m \mathbb{E}[\frac{|\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta}}{[l_k \kappa_{mR}(\sigma_{t_{km}})]^\delta}] + \alpha^2), \end{aligned} \quad (5.41)$$

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

Also, it follows from (5.38) that $\mathbb{E}[l_{k+1}^2] \leq \alpha^2 + \vartheta^m \mathbb{E}[|l_k \kappa_{mR}(\sigma_{t_{km}})|^2]$. By summing the above two inequalities we obtain that

$$\begin{aligned} \theta_{k+1} &\leq \phi(2\alpha^2 + \vartheta^m \mathbb{E}[\frac{|\tilde{x}_{t_{km}} - l_k Q_{mR}(\tilde{x}_{t_{km}}/l_k)|^{2+\delta}}{[l_k \kappa_{mR}(\sigma_{t_{km}})]^\delta} + |l_k \kappa_{mR}(\sigma_{t_{km}})|^2]) \\ &\leq 2\phi\alpha^2 + \phi\vartheta^m \frac{\varsigma}{2^{2mR}} M_\delta[\tilde{x}_{t_{km}}, l_k] \end{aligned} \quad (5.42)$$

$$\leq 2\phi\alpha^2 + \phi\varsigma(\frac{\vartheta}{2^{2R}})^m \theta_k, \quad (5.43)$$

where (5.42) 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 $m > 0$ such that $\nu \triangleq \phi\varsigma(\frac{\vartheta}{2^{2R}})^m < 1$. In addition, it immediately follows from (5.43) that $\sup_k \theta_k \leq \frac{2\phi\alpha^2}{1-\nu}$. For any given $j > 1$, we get that

$$\begin{aligned} \mathbb{E}[|\tilde{x}_j|^2] &= \sum_{k=0}^{\infty} \mathbb{E}[|\tilde{x}_j|^2] 1_{\{t_{km} \leq j < t_{(k+1)m}\}} \\ &\leq \sum_{k=0}^{\infty} \mathbb{E}[|\lambda^{j-t_{km}} \tilde{x}_{t_{km}}|^2] 1_{\{t_{km} \leq j < t_{(k+1)m}\}} + \alpha^2 \\ &\leq \vartheta^m \sup_k \theta_k + \alpha^2 \leq \frac{2\vartheta^m \phi \alpha^2}{1-\nu} + \alpha^2. \end{aligned} \quad (5.44)$$

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

Remark 5.3. *It should be noted that in [78], 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 [78, 84], a similar adaptive quantizer is adopted to*

establish the sufficiency. What makes the current problem more challenging is that the down sampling interval, denoted by $t_{(k+1)m} - t_{km}$, is stochastic and unbounded. While in [78, 84], 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. Further, 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.*

5.4 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 [73, 78].

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 [78]. Note that the condition on data rate is a sum of the data rates from all used quantizers.

5.4.1 Real Jordan Form

As in [78, 84], we adopt a real Jordan form for the system under investigation which is briefly reviewed below. There is no loss of generality to assume that all the eigenvalues of A lie outside or on the unit circle [107].

Let $\lambda_1, \dots, \lambda_d \in \mathbb{C}$ 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 [48] has the block diagonal structure $J = \text{diag}(J_1, \dots, J_d) \in \mathbb{R}^{n \times n}$ 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:

$$x_{k+1} = Jx_k + Bu_k, \quad (5.45)$$

where the state vector $x_k = [x_k^{(1)T}, \dots, x_k^{(d)T}]^T \in \mathbb{R}^n$ 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 $P_{x_0}(\cdot)$.

5.4.2 Necessity

Theorem 5.3. *Consider the system (5.45) 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} , $j \in \{1, \dots, n_i\}$ denote the dimension of an invariant real subspace of J_i , $i \in \{1, \dots, d\}$. Then, a necessary condition for the asymptotic MS-stabilization of the networked system is that $\forall r_i \in \mathcal{D}_i \triangleq \{d_{i1}, \dots, d_{in_i}\}$ 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,

$$q > 1 - \frac{1}{(\prod_{i=1}^d |\lambda_i|^{2r_i})^{1/\Sigma_r}}; \quad (5.46)$$

(b) The data rate satisfies the following strict inequality

$$R > \frac{\Sigma_r}{2} \log_2 \mathbb{E}[(\prod_{i=1}^d |\lambda_i|^{2r_i})^{t_1^*/\Sigma_r}]. \quad (5.47)$$

Remark 5.6.

(1) For a lossless digital channel, i.e., $p \rightarrow 0$ and $q \rightarrow 1$, it is obvious that the sojourn time t_1^* is always equal to one, (5.46) is automatically enforced, and the inequality (5.47) reduces to $R > \sum_{i=1}^d \mu_i \log_2 |\lambda_i|$ by selecting $r_i = \mu_i, \forall i \in \{1, \dots, d\}$. This is the well-known minimum data rate condition for stabilizing an unstable linear system [84, 107].

(2) If it is specialized to scalar systems, (5.47) becomes $R > \frac{1}{2} \log_2 \mathbb{E}[|\lambda_1|^{2t_1^*}]$, which is the same as the rate condition in (5.10).

(3) When the packet loss is i.i.d., our result recovers the one derived in [78].

Proof of Theorem 5.3:

Together with the results in Section 5.3, the proof of [78] is extended here to establish the stability of error dynamics in random time t_{km} . By the definition of $\mathcal{D}_i$, for any $r_i \in \mathcal{D}_i$, the block J_i has an invariant real subspace, denoted by $\mathcal{H}_i$, of dimension r_i . Denote the indices of the nonempty subspaces by $\{e_1, \dots, e_{d_r}\}$, e.g., $\mathcal{H}_{e_i} \neq \emptyset$ and the corresponding state variables w.r.t. $\mathcal{H}_{e_i}$ by $x_k^{(e_i)}$. Consider the subspace $\mathcal{H}$ formed by taking the Cartesian product of all the nonempty invariant real subspaces, i.e., $\mathcal{H} = \prod_{i=1}^{d_r} \mathcal{H}_{e_i}$, the dimension of $\mathcal{H}$ is computed as $\Sigma_r = \sum_{i=1}^{d_r} r_i$. Stack the unstable state variables $x_k^{(e_i)}$ to get a new vector state $x_k^{\mathcal{H}} = [x_k^{(e_1)T}, \dots, x_k^{(e_{d_r})T}]^T \triangleq P x_k$, where P is some transformation matrix. Thus, the new vector state $x_k^{\mathcal{H}}$ evolves as follows:

$$x_{k+1}^{\mathcal{H}} = J^{\mathcal{H}} x_k^{\mathcal{H}} + P B u_k, \quad (5.48)$$

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

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

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

$$\lim_{k \rightarrow \infty} \mathbb{E}\left[\frac{[\det(J^{\mathcal{H}})]^{2(k+1)/\Sigma_r}}{2^{2R(\gamma_0+\dots+\gamma_k)/\Sigma_r}}\right] = 0. \quad (5.49)$$

As in Theorem 5.1, a necessary condition for (5.49) is that

$$\mathbb{E}\left[\frac{[\det(J^{\mathcal{H}})]^{2t^*/\Sigma_r}}{2^{2R/\Sigma_r}}\right] < 1. \quad (5.50)$$

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

5.4.3 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 (5.45) 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,*

$$q > 1 - \frac{1}{\max_{i \in \{1, \dots, d\}} |\lambda_i|^2}; \quad (5.51)$$

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

$$R > \frac{\mu_i}{2\mathbf{a}_i(R)} \log_2(\mathbb{E}[|\lambda_i|^{2t^*}]), \forall i \in \{1, \dots, d\}, \quad (5.52)$$

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

$$\begin{cases} 0 \leq \mathbf{a}_j(R) \leq 1; \\ \sum_{j=1}^d \mathbf{a}_j(R) \leq 1, \quad \forall j \in \{1, \dots, d\}; \\ \frac{R}{\mu_j} \mathbf{a}_j(R) \in \mathbb{N}. \end{cases} \quad (5.53)$$

Remark 5.7.

(1) For a Markovian lossy channel with infinite bandwidth, the data rate condition of (5.52) is automatically satisfied. The probability condition of the channel recovering from packet loss in (5.51) reduces to that of [45]. That is, the sufficient condition is also necessary in this case.

(2) The sufficient condition is optimal when the magnitude of strictly unstable eigenvalues is the same. For example, assume $\exists d_1 \leq d$ such that $|\lambda_1| = \dots = |\lambda_{d_1}| > 1$ and $|\lambda_j| = 1, \forall j \in \{d_1 + 1, \dots, d\}$ for the vector system of (5.45), the transition probability and rate condition in (5.51) and (5.52) 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 (5.46) and (5.47) by choosing $r_i = \mu_i, \forall i \in \{1, \dots, d_1\}$ and $r_i = 0, \forall i \in \{d_1 + 1, \dots, d\}$. In particular, if all the unstable eigenvalues have the same magnitude, e.g., $|\lambda_1| = \dots = |\lambda_d|$, the sufficient condition is necessary as well.

Proof of Theorem 5.4:

For any data rate R satisfying (5.52) and (5.53), the uniform quantizer of (4.2) will be adopted. Similar to the proof of Theorem 5.2, divide the integers $j \in \mathbb{N}$ into cycles $\{km, \dots, (k+1)m-1\}, \forall k \in \mathbb{N}$ with length $m \in \mathbb{N}$, 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:

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

where a stabilizing control gain K is chosen to satisfy that the spectra radius of $J + BK$ is strictly less than one. Denote the ℓ -th component of the state vector corresponding to the i -th unstable mode by $x_k^{(i,\ell)}$, where $i \in \{1, \dots, d\}$ and $\ell \in \{1, \dots, \mu_i\}$. The vector σ_0 is composed by $\sigma_0 = [\sigma_0^{(1,1)}, \dots, \sigma_0^{(1,\mu_1)}, \sigma_0^{(2,1)}, \dots, \sigma_0^{(d,\mu_d)}]^T$, with $\sigma_0^{(i,\ell)} = q_{R\mathbf{a}_i(R)/\mu_i}(\frac{x_0^{(i,\ell)}}{l_0})$ while $\sigma_{t_{km}} = [\sigma_{t_{km}}^{(1,1)}, \dots, \sigma_{t_{km}}^{(1,\mu_1)}, \sigma_{t_{km}}^{(2,1)}, \dots, \sigma_{t_{km}}^{(d,\mu_d)}]^T$, with $\sigma_{t_{km}}^{(i,\ell)} = q_{mR\mathbf{a}_i(R)/\mu_i}(\frac{x_{t_{km}}^{(i,\ell)} - \hat{x}_{t_{km}}^{(i,\ell)}}{L_k^i})$, where $q_{R\mathbf{a}_i(R)/\mu_i}(\cdot)$ is a uniform quantizer of (4.2) using $R\mathbf{a}_i(R)/\mu_i$ bits of precision to represent quantizer output and similarly for $q_{mR\mathbf{a}_i(R)/\mu_i}(\cdot)$. Note that by (5.53), 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 $L_k = \text{diag}(L_k^1 I_{\mu_1}, \dots, L_k^d I_{\mu_d})$. Moreover, L_k^i is simultaneously updated on both sides of the channel via

$$\begin{aligned}L_0^i &= \frac{\zeta \sqrt{\mu_i} |\lambda_i|}{2^{R\mathbf{a}_i(R)/\mu_i}} l_0; \\ L_{k+1}^i &= \frac{\zeta \sqrt{\mu_i} (t_{(k+1)m} - t_{km})^{\mu_i - 1} |\lambda_i|^{t_{(k+1)m} - t_{km}}}{2^{mR\mathbf{a}_i(R)/\mu_i}} L_k^i.\end{aligned}\tag{5.54}$$

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

$$\begin{aligned}\tilde{x}_0 &= x_0, \tilde{x}_1 = J(x_0 - l_0 \sigma_0); \\ \tilde{x}_{j+1} &= J \tilde{x}_j, j \in \{t_{km}, t_{km} + 1, \dots, t_{(k+1)m} - 2\}; \\ \tilde{x}_{t_{(k+1)m}} &= J^{t_{(k+1)m} - t_{km}} (\tilde{x}_{t_{km}} - L_k \sigma_{t_{km}}).\end{aligned}\tag{5.55}$$

By Lemma 3.18, the quantizer will not overflow, i.e., $|x_{t_{km}}^{(i,\ell)} - \hat{x}_{t_{km}}^{(i,\ell)}| \leq L_k^i, \forall k \in \mathbb{N}$. In fact it obviously holds for $k = 0$. Assume $\exists k \geq 1, |\tilde{x}_{t_{km}}^{(i,\ell)}| \leq L_k^i, \forall i \in \{1, \dots, d\}, \forall \ell \in$

$\{1, \dots, \mu_i\}$, then $\|\tilde{x}_{t_{km}}^{(i)}\|_\infty \leq L_k^i$. By the error dynamics in (5.55), the update recursion for L_k^i in (5.54) and Lemma 3.18, it can be established that $|\tilde{x}_{t_{(k+1)m}}^{(i,\ell)}| \leq \|\tilde{x}_{t_{(k+1)m}}^{(i)}\|_\infty \leq \|J_i^{t_{(k+1)m}-t_{km}}\|_\infty \|\tilde{x}_{t_{km}}^{(i)} - L_k^i \sigma_{t_{km}}^{(i)}\|_\infty \leq L_{k+1}^i$. In addition, (5.51) implies that $\forall i \in \{1, \dots, d\}, |\lambda_i|^2(1-q) < 1$. That is, there exists a $g > 1$ such that

$$|\lambda_i|^{2g}(1-q) < 1. \quad (5.56)$$

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 $\lim_{j \rightarrow \infty} g_j = 1$, it is easy to show that $(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^{g_j} \leq (|\lambda_i|^{2t_1^*})^{g_j} \leq (|\lambda_i|^{2t_1^*})^g$. By the dominated convergence theorem [4], it follows that $\lim_{j \rightarrow \infty} \mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^{g_j}] = \mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^c] < 1$ by (5.52). Thus, there exists a $c \in (1, g]$ such that the following inequality holds:

$$\mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^c] < 1. \quad (5.57)$$

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

$$\begin{aligned} \mathbb{E}[(L_{k+1}^i)^2] &= \zeta^2 \mu_i \mathbb{E}[\frac{(t_{(k+1)m} - t_{km})^{2(\mu_i-1)} |\lambda_i|^{2(t_{(k+1)m}-t_{km})}}{2^{2mR\mathbf{a}_i(R)/\mu_i}}] \mathbb{E}[(L_k^i)^2] \\ &\leq \zeta^2 \mu_i (\mathbb{E}[(t_{(k+1)m} - t_{km})^{2c'(\mu_i-1)}])^{1/c'} (\mathbb{E}[(\frac{|\lambda_i|^{2(t_{(k+1)m}-t_{km})}}{2^{2mR\mathbf{a}_i(R)/\mu_i}})^c])^{1/c} \mathbb{E}[(L_k^i)^2] \\ &\leq C_1 (\mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^c])^{m/c} \mathbb{E}[(L_k^i)^2] \\ &\triangleq \eta_i \mathbb{E}[(L_k^i)^2]. \end{aligned} \quad (5.58)$$

Here the first equality is obtained from (5.54) and the fact that $t_{(k+1)m} - t_{km}$ is independent of L_k^i by Lemma 5.1. The Hölder inequality was applied in the first inequality. The constant $C_1 \triangleq \zeta^2 \mu_i (\mathbb{E}[(t_{(k+1)m} - t_{km})^{2c'(\mu_i-1)}])^{1/c'}$ in the second inequality is finite by (5.33). In light of (5.57), there exists an $m > 0$ such that

$$\eta_i = C_1 (\mathbb{E}[(\frac{|\lambda_i|^{2t_1^*}}{2^{2R\mathbf{a}_i(R)/\mu_i}})^c])^{m/c} < 1. \quad (5.59)$$

Using the above inequality and Lemma 3.18, we can further derive that

$$\begin{aligned} \mathbb{E}[\sum_{k=1}^{\infty} \|\tilde{x}_k^{(i)}\|_{\infty}^2] &\leq \zeta^2 \mu_i \sum_{k=0}^{\infty} \mathbb{E}[\sum_{j=t_{km}}^{t_{(k+1)m}-1} (j - t_{km})^{2(\mu_i-1)} |\lambda_i|^{2(j-t_{km})} \|\tilde{x}_{t_{km}}^{(i)}\|_{\infty}^2] \\ &\leq \zeta^2 \mu_i \mathbb{E}[(t_m - t_0)^{2\mu_i-1} |\lambda_i|^{2(t_m-t_0)}] \sum_{k=0}^{\infty} \mathbb{E}[\|\tilde{x}_{t_{km}}^{(i)}\|_{\infty}^2] \end{aligned} \quad (5.60)$$

$$\leq C_2 \sum_{k=0}^{\infty} \mathbb{E}[(L_k^i)^2] \leq C_2 (L_0^i)^2 \sum_{k=0}^{\infty} \eta_i^k. \quad (5.61)$$

In the above, the second inequality follows from Lemma 3.18. The inequality in (5.60) is due to that $\forall j \in \{t_{km}, t_{km} + 1, \dots, t_{(k+1)m}\}$, $(j - t_{km})^{2(\mu_i-1)} |\lambda_i|^{2(j-t_{km})} \leq (t_{(k+1)m} - t_{km})^{2(\mu_i-1)} |\lambda_i|^{2(t_{(k+1)m}-t_{km})}$ and $t_{(k+1)m} - t_{km}$ is of the same distribution as $t_m - t_0$. Choosing $g' > 1$ such that $1/g' + \frac{1}{g} = 1$ and using Hölder inequality, then $C_2 = \zeta^2 \mu_i \mathbb{E}[(t_m - t_0)^{2\mu_i-1} |\lambda_i|^{2(t_m-t_0)}] \leq \zeta^2 \mu_i (\mathbb{E}[(t_m - t_0)^{g'(2\mu_i-1)}])^{1/g'} (\mathbb{E}[|\lambda_i|^{2gt_1^*}])^{m/g} < \infty$ by (5.33) and (5.56). By (5.59) and (5.61), we obtain that $\mathbb{E}[\sum_{k=1}^{\infty} \|\tilde{x}_k^{(i)}\|_{\infty}^2] < \infty$. This immediately implies that $\lim_{k \rightarrow \infty} \mathbb{E}[\|\tilde{x}_k^{(i)}\|_{\infty}^2] = 0$, which further concludes that $\lim_{k \rightarrow \infty} \mathbb{E}[\|\tilde{x}_k^{(i)}\|^2] \leq \mu_i \lim_{k \rightarrow \infty} \mathbb{E}[\|\tilde{x}_k^{(i)}\|_{\infty}^2] = 0$. Consequently, we have shown that $\lim_{k \rightarrow \infty} \mathbb{E}[\|\tilde{x}_k\|^2] = 0$. With the designed stabilizing controller, it is trivial that $\lim_{k \rightarrow \infty} \mathbb{E}[\|x_k\|^2] = 0$. The rest of the proof follows from a similar line as in [78] and details are omitted. $\blacksquare$

5.4.4 An Example

In this subsection, 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 Theorem 5.3 and Theorem 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.

Figure 5.3: Stabilizable and unstabilizable regions for the example in Subsection 5.4.4.

5.5 Summary

Packet loss and data rate constraint are two important issues of networked control systems. In this chapter, we have investigated the necessary and sufficient data rate conditions for the mean square stabilization of networked unstable linear systems where the communication channel is subject to Markovian packet loss. The temporal correlations of the packet loss process posed significant challenges for 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 and is explicitly given in terms of the magnitude of the unstable mode and the transition probabilities of the Markov chain. The result completely 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 and sufficient conditions were respectively derived and shown to be optimal for some special cases. Our approach can also be directly used to NCSs where transmission times are driven by an i.i.d. stochastic process with the consideration of reducing the size of the data transferred at each transmission.

Chapter 6

Stability of Kalman Filtering with Markovian Packet Loss

This chapter studies the stability of Kalman filtering over a network subject to Markovian packet loss. For second-order and certain classes of higher-order systems, necessary and sufficient conditions for stability of the mean estimation error covariance matrices. All stability criteria are expressed by simple inequalities in terms of the largest eigenvalue of the open loop matrix and transition probabilities of the Markov process. Their implications and relationships with related results in the literature are discussed.

The organization of the chapter is as follows. The problem under consideration is precisely formulated in Section 6.1, where two stability notions are introduced. In Section 6.2, 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 stability of the mean estimation error covariance matrices of second-order systems is provided in Section 6.3. The necessary condition presented in Section 6.2 is proved to be sufficient for certain classes of higher-order systems in Section 6.4. Numerical examples are presented in Section 6.5. Most of proofs are moved to Section 6.6. Concluding remarks are drawn in Section 6.7.

Figure 6.1: Network configuration.

6.1 Problem Formulation

Consider a discrete-time stochastic linear system:

$$\begin{cases} x_{k+1} = Ax_k + w_k; \\ y_k = Cx_k + v_k, \end{cases} \quad (6.1)$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ are vector state and measurement. $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are white Gaussian noises with zero means and covariance matrices $Q > 0$ and $R > 0$, respectively. C is of full row rank, i.e., $\text{rank}(C) = m \leq n$. The initial state x_0 is a random Gaussian vector of mean $\hat{x}_0$ and the covariance matrix $P_0 > 0$. Moreover, w_k , v_k and x_0 are mutually independent.

We consider a network environment where the raw measurements of the system are transmitted to an estimator via an unreliable communication channel, see Fig. 6.1. Due to random fading and/or congestion of the communication channel, packets may be lost while in transit through the channel. Different from [124, 128], the present work ignores other effects such as quantization, transmission errors and data 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 studied in [103] due to possible temporal correlation of network conditions. Furthermore, assume that $\{\gamma_k\}_{k \geq 0}$ does not contain any information of the system. 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 transition probability matrix is Markov process is given in (5.2), i.e.,

$$\Pi^+ = (\mathbb{P}\{\gamma_{k+1} = j | \gamma_k = i\})_{i,j \in \mathbb{S}} = \begin{bmatrix} 1-q & q \\ p & 1-p \end{bmatrix}. \quad (6.2)$$

Denote $(\Omega, \mathcal{F}, \mathbb{P})$ the common probability space for all random variables in the chap-

ter, where Ω is the space of elementary events, $\mathcal{F}$ is the underlying σ -field on Ω , and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. Let $\mathcal{F}_k \triangleq \sigma(y_i\gamma_i, \gamma_i, i \leq k) \subset \mathcal{F}$ 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 w.r.t. $\mathbb{P}$. Similar to Chapter 5, define a sequence of stopping times $\{t_k\}_{k \geq 0}$ adapted to the Markov process $\{\gamma_k\}_{k \geq 0}$ as follows:

$$\begin{aligned} t_0 &= 0, \quad t_1 = \inf\{k | k \geq 1, \gamma_k = 1\}, \\ t_2 &= \inf\{k | k > t_1, \gamma_k = 1\}, \dots, \\ t_k &= \inf\{k | k > t_{k-1}, \gamma_k = 1\} \end{aligned} \tag{6.3}$$

and sojourn time $t_k^* \triangleq t_k - t_{k-1} > 0$.

Denote the state estimate and one-step prediction corresponding to the minimum mean square error estimator by $\hat{x}_{k|k} = \mathbb{E}[x_k | \mathcal{F}_k]$ and $\hat{x}_{k+1|k} = \mathbb{E}[x_{k+1} | \mathcal{F}_k]$, respectively. The associated estimation error covariance matrices are defined by $P_{k|k} = \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^H | \mathcal{F}_k]$ and $P_{k+1|k} = \mathbb{E}[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^H | \mathcal{F}_k]$, where A^H is the conjugate transpose of A . By [103], it is known that the Kalman filter is still optimal. That is, the following recursions are in force:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \gamma_k K_k (y_k - C\hat{x}_{k|k-1}); \tag{6.4}$$

$$P_{k|k} = P_{k|k-1} - \gamma_k K_k C P_{k|k-1}, \tag{6.5}$$

where $K_k = P_{k|k-1} C^H (C P_{k|k-1} C^H + R)^{-1}$. In addition, the time update equations continue to hold: $\hat{x}_{k+1|k} = A\hat{x}_{k|k}$, $P_{k+1|k} = A P_{k|k} A^H + Q$ 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 $M_k = P_{t_k+1}$. To analyze the behavior of the estimation error covariance matrices, we introduce two types of stability.

Definition 6.1. *We say that the mean state estimation error covariance matrices are stable in sampling times if $\sup_{k \in \mathbb{N}} \mathbb{E}[P_k] < \infty^1$ while they are stable in stopping*

¹This notation means that there is a positive definite $\bar{P}$ such that $\mathbb{E}[P_k] < \bar{P}$ for any $k \in \mathbb{N}$.

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 Section 6.2. 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 [51] 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.

In consideration of Theorems 3 and 8 of [79], there is no loss of generality to assume that:

Assumption 6.1. P_0, Q, R are all identity matrices with compatible dimensions.

Assumption 6.2. All the eigenvalues of A lie outside the unit circle.

For the eigenvalues on the unit circle, we can do an arbitrarily small perturbation on A , e.g., $\tilde{A} = (1 + \epsilon)A$, such that the eigenvalue of $\tilde{A}$ is strictly greater than one. Note that our sufficient condition are given in the strictly inequality, which allows us to do such perturbation on A . See more details in [79].

Assumption 6.3. (C, A) is observable.

6.2 Equivalence of the Two Stability Notions

It is generically difficult to directly study the notion of stability in sampling times [51]. However, the two stability notions will be shown to be equivalent in this section.

For any $i \in \mathbb{S}$, denote $\mathbb{E}^i[\cdot]$ the mathematical expectation operator conditioned on the event that $\{\gamma_0 = i\}$.

Similar meaning applies to the notation $\sup_{k \in \mathbb{N}} \mathbb{E}[M_k] < \infty$.

Lemma 6.1. *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) “ $\Leftarrow$ ” It is obvious since $\mathbb{E}[P_k] \leq \mathbb{E}^1[P_k] + \mathbb{E}^0[P_k]$. “ $\Rightarrow$ ” Let $\mathbb{P}\{\gamma_0 = 1\} = \mathbb{P}\{\gamma_0 = 0\} = 1/2$. Note that $P_k \geq 0$, then $\mathbb{E}[P_k] \geq \mathbb{E}^1[P_k]/2$ and $\mathbb{E}[P_k] \geq \mathbb{E}^0[P_k]/2$.

(b) Similar to (a). ■

Theorem 6.1. *Consider the system (6.1) satisfying Assumptions 6.1-6.3 and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix (6.2). 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 $k \geq 1$. Since t_k is a stopping time, $\mathcal{F}_{t_k} \triangleq \sigma(y_i \gamma_i, \gamma_i, i \leq t_k)$ is a well defined σ -field. Noting that $P_k \geq Q = I$ for all $k \in \mathbb{N}$ and $M_k = P_{t_k+1}$, it immediately follows from the property of conditional expectation that

$$\mathbb{E}[M_{k+1}] = \mathbb{E}[\mathbb{E}[M_{k+1} | \mathcal{F}_{t_k}]] = \mathbb{E}[g^{t_k+1}(M_k)] \geq \mathbb{E}[g^{t_k+1}(Q)] \geq \mathbb{E}\left[\sum_{j=0}^{t_k^*} A^j (A^j)^H\right], \quad (6.6)$$

where the first inequality is due to that $g(\cdot)$ is a monotonically increasing function. Let $J = \text{diag}(J_1, \dots, J_d) \in \mathbb{C}^{n \times n}$ be the Jordan canonical form of A , where $J_i \in \mathbb{C}^{n_i \times n_i}$ corresponds 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

$$\sum_{j=0}^{t_{k+1}^*} A^j (A^j)^H = U \sum_{j=0}^{t_{k+1}^*} J^j U^{-1} U^{-H} (J^j)^H U^H \geq \lambda_{\min}(U^{-1} U^{-H}) U \sum_{j=0}^{t_{k+1}^*} J^j (J^j)^H U^H, \quad (6.7)$$

where $\lambda_{\min}(U^{-1} U^{-H}) > 0$ is the smallest eigenvalue of $U^{-1} U^{-H}$. In view of (6.6), (6.7) and Lemma 6.1, it is clear that $\sup_{k \in \mathbb{N}} \mathbb{E}[M_{k+1}] < \infty$ implies that

$$\sup_{k \in \mathbb{N}} \mathbb{E}^1\left[\sum_{j=0}^{t_{k+1}^*} J^j (J^j)^H\right] < \infty. \quad (6.8)$$

Its (n_i, n_i) -th element is computed as $\mathbb{E}^1[\sum_{j=0}^{t_1^*} |\lambda_i|^{2j}] = \frac{|\lambda_i|^2 \mathbb{E}^1[|\lambda_i|^{2t_1^*}] - 1}{|\lambda_i|^2 - 1}$. By (6.8) and the equivalence property of norms on a finite-dimensional vector space, it follows that $\frac{|\lambda_i|^2 \mathbb{E}^1[|\lambda_i|^{2t_1^*}] - 1}{|\lambda_i|^2 - 1} < \infty$. 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 6.2. *Consider the system (6.1) satisfying Assumptions 6.1-6.3 and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix (6.2). 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 $P_k - P_k C^H (C P_k C^H + R)^{-1} C P_k \geq 0$ [119], it yields that for any $k > 3$,

$$P_{k+1} \geq (1 - \gamma_k) A P_k A^H + Q \geq \sum_{j=1}^k \left(\prod_{i=j}^k (1 - \gamma_i) \right) A^{k-j} (A^{k-j})^H, \quad (6.9)$$

where the second inequality is due to that $Q = I$ by Assumption 6.1. Denote $\pi_j^i = \mathbb{P}\{\gamma_j = i\}, i \in \{0, 1\}$ and $\pi_j = [\pi_j^0, \pi_j^1]$. By (6.2), 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., $\lim_{j \rightarrow \infty} \pi_j^i = \pi^i, i \in \mathbb{S}$. By (6.2), we further obtain that $\pi^0 = \frac{p}{p+q} > 0$. Then, it follows that $\underline{\pi}^0 \triangleq \inf_{j \geq 1} \pi_j^0 > 0$, which further implies that for all $j \geq 2$,

$$\mathbb{E}[\prod_{i=j}^k (1 - \gamma_i)] \geq \mathbb{E}[\prod_{i=j}^k (1 - \gamma_i) | \gamma_{j-1} = 0] \mathbb{P}(\gamma_{j-1} = 0) \geq \underline{\pi}^0 (1 - q)^{k-j}. \quad (6.10)$$

In view of (6.9), 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 6.1, it yields that $\rho(A)^2(1 - q) < 1$. ■

Remark 6.1. *Let $\bar{q} = \max\{q, 1 - p\}$, [119] 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 6.2 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 Section 6.6.

Theorem 6.3. *Consider the system (6.1) satisfying Assumptions 6.1-6.3 and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix (6.2). 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.

6.3 Second-order Systems

Consider second-order systems with the following structure:

Assumption 6.4. $A = \text{diag}(\lambda_1, \lambda_2)$ and $\text{rank}(C) = 1$, where $\lambda_2 = \lambda_1 \exp(\frac{2\pi r}{d} \mathbf{i})$, $\mathbf{i}^2 = -1$, $d > r \geq 1$ and $r, d \in \mathbb{N}$ are irreducible.

Under Assumption 6.4, 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 6.4.

Theorem 6.4. *Consider the system (6.1) satisfying Assumptions 6.1-6.3 and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix (6.2). Then,*

- (a) *if (C, A) satisfies Assumption 6.4, 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 Section 6.6. By Theorem 6.3, the results in Theorem 6.4 apply to the notion of stability in sampling times as well. Some remarks are included below.

Remark 6.2. *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 6.3. *If the conjugate complex eigenvalues satisfy that $\lambda_2 = \lambda_1 \exp(2\pi\varphi\mathbf{i})$, where φ is an irrational number, Assumption 6.4 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 Section 6.4, 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 6.4. *In [51], 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. For vector systems, they give a conservative sufficient condition for the peak covariance stability and do not consider the usual stability.*

Remark 6.5. *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 Assumption 6.4 in Theorem 6.4 is reduced to that $q > 1 - \rho(A)^{-\frac{2d}{d-1}}$, which recovers the result in [79]. Note that under i.i.d. packet losses, a lower bound for the critical packet loss rate given in [103] is interpreted as $q > 1 - \rho(A)^{-2}$, which is obviously not tight for systems satisfying Assumption 6.4.*

6.4 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 [79, 91, 103]. 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 6.1 is also sufficient under certain classes of higher-order systems, whose proofs are given in Section 6.6. To this aim, some definitions introduced in [79] are adopted.

Definition 6.2. *The pair (C, A) is one step observable if C is of full column rank.*

Definition 6.3. *Assume that (C, A) is in diagonal form, i.e., $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ and $C = [C_1, \dots, C_n]$. An equi-block of the system is defined as the subsystem corresponding to the block $(C_{\mathcal{I}}, A_{\mathcal{I}})$, where $\mathcal{I} = \{i_1, \dots, i_l\} \subset \{1, \dots, n\}$ is an index set such that $|\lambda_{i_1}| = \dots = |\lambda_{i_l}|$ and $A_{\mathcal{I}} = \text{diag}(\lambda_{i_1}, \dots, \lambda_{i_l})$, $C_{\mathcal{I}} = [C_{i_1}, \dots, C_{i_l}]$.*

Definition 6.4. *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 6.5. *(C, A) is a non-degenerate pair.*

Theorem 6.5. *Consider the system (6.1) satisfying Assumptions 6.1-6.3, 6.5 and the packet loss process of the measurements governed by a Markov process with transition probability matrix (6.2). 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 [79] 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 [103] is tight for non-degenerate systems. While in Theorem 6.5, we give a necessary and sufficient condition for stability of non-degenerate systems under Markovian packet losses. Next, the necessary condition in Theorem 6.1 is proved to be sufficient for another class of higher-order systems with the following structure.

Assumption 6.6. $A^{-1} = \text{diag}(J_1, \dots, J_m)$ 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 6.6. Consider the system (6.1) satisfying Assumptions 6.1-6.3, 6.6 and the packet loss process of the measurements governed by a time-homogeneous Markov process with transition probability matrix (6.2). 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 6.6. Note that except for the case that A has n eigenvalues and each of them is with a distinct magnitude, Assumptions 6.5 and 6.6 define two disjoint classes of higher-order systems.

6.5 Illustrative Example

Example 1: Let a second-order system be specified by

$$A = \begin{bmatrix} 1.5 & 0 \\ 0 & -1.5 \end{bmatrix} \text{ and } C = [1 \quad 1]. \quad (6.11)$$

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 6.4. Two sample paths with different recovery rates are shown in Fig. 6.2 and Fig. 6.3, 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 6.4 is violated.

Example 2: The results on higher-order systems in Section 6.4 are applied to target tracking over a packet loss network. The dynamic of target is expressed by [102]

$$x_{k+1} = \begin{bmatrix} 1 & h & h^2 \\ 0 & 1 & h \\ 0 & 0 & 1 \end{bmatrix} x_k + w_k, \quad (6.12)$$

where h is the sampling period and x_k denotes the target state at time kh , including

Figure 6.2: A sample path with $q = 0.8$ and $p = 0.1$. Figure 6.3: A sample path with $q = 0.6$ and $p = 0.1$.

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

$$Q = 2\alpha\sigma_m^2 \begin{bmatrix} h^5/20 & h^4/8 & h^3/6 \\ h^4/8 & h^3/3 & h^2/2 \\ h^3/6 & h^2/2 & h \end{bmatrix}, \quad (6.13)$$

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:

$$y_k = [1 \ 0 \ 0]x_k + v_k, \quad (6.14)$$

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

Figure 6.4: A sample path with $q = 0.2$ and $p = 0.5$.

and covariance as follows [102]:

$$P_0 = \begin{bmatrix} R & R/h & 0 \\ R/h & 2R/h^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In this example, set $h = 0.1s$, $\alpha = 0.1$, $\sigma_m^2 = 1$ and $R = 0.01$. Although here A is marginally unstable, a scaling on A can be made as in Theorem 8 of [79]. Jointly with Theorem 6.6, 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. 6.4, which illustrates that the tracking task is fulfilled.

6.6 Proofs

Lemma 6.2. [104] *For any $A \in \mathbb{R}^{n \times n}$ and $\epsilon > 0$, it holds that*

$$\|A^k\| \leq N\eta^k, \forall k \geq 0, \quad (6.15)$$

where $N = \sqrt{n}(1 + \frac{2}{\epsilon})^{n-1}$ and $\eta = \rho(A) + \epsilon\|A\|$.

If A is invertible, define $\phi(k, i) = A^{t_i - t_k}$ if $k > i$ and $\phi(k, i) = I$ if $k \leq i$. Let

$$\Theta_k = \sum_{i=0}^k \gamma_i (A^{i-k})^H C^H C A^{i-k} + (A^{-k})^H A^{-k}, \quad (6.16)$$

$$\Lambda_k = \sum_{j=0}^k \phi^H(k, j) C^H C \phi(k, j) + \phi^H(k, 0) \phi(k, 0), \quad (6.17)$$

$$\Xi_k = \sum_{j=0}^k \phi^H(j, 0) C^H C \phi(j, 0) + \phi^H(k, 0) \phi(k, 0), \quad (6.18)$$

$$\Xi = \sum_{j=0}^{\infty} \phi^H(j, 0) C^H C \phi(j, 0). \quad (6.19)$$

Lemma 6.3. *Under Assumptions 6.1-6.3, there exist strictly positive constant numbers α and β such that for any $k \in \mathbb{N}$,*

$$\alpha A \Lambda_k^{-1} A^H \leq M_k \leq \beta A \Lambda_k^{-1} A^H. \quad (6.20)$$

Proof. By revising Lemma 2 in [79] 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 (6.2), 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 6.4. *Under Assumptions 6.1-6.3, there exist strictly positive constant numbers $\tilde{\alpha}$ and $\tilde{\beta}$ such that*

$$\tilde{\alpha} A \mathbb{E}^1[\Xi^{-1}] A^H \leq \sup_{k \in \mathbb{N}} \mathbb{E}^1[M_k] \leq \tilde{\beta} A \mathbb{E}^1[\Xi^{-1}] A^H. \quad (6.21)$$

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., $(t_k^*, t_k^* + t_{k-1}^*, \dots, t_k^* + \dots + t_1^*) \stackrel{d}{=} (t_1^*, t_1^* + t_2^*, \dots, t_1^* + \dots + t_k^*)$, where $\stackrel{d}{=}$ 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 (6.17) and (6.18). Jointly with Lemma 6.3, it follows that

$$\mathbb{E}^1[M_k] \leq \beta A \mathbb{E}^1[\Xi_k^{-1}] A^H. \quad (6.22)$$

Under Assumption 6.2, it is possible to select a positive $\epsilon < \frac{1 - \rho(A^{-1})}{\|A^{-1}\|}$ and $\eta =$

$\rho(A^{-1}) + \epsilon \|A^{-1}\| < 1$, then it follows from Lemma 6.2 that for any $k \in \mathbb{N}$,

$$\sum_{j=k+1}^{\infty} \phi^H(j, k) C^H C \phi(j, k) \leq N \|C\|^2 \sum_{j=k+1}^{\infty} \eta^{2(t_k - t_j)} I \leq \frac{N \|C\|^2}{1 - \eta^2} I \triangleq \beta_0 I, \quad (6.23)$$

where the last inequality is due to that $t_k^* \geq 1$ for all $k \in \mathbb{N}$. Let $\beta_1 = \min(1, \beta_0^{-1})$ and $\tilde{\beta} = \beta \beta_1$, we further obtain that

$$\Xi_k \geq \sum_{j=0}^k \phi^H(j, 0) C^H C \phi(j, 0) + \beta_0^{-1} \phi^H(k, 0) \left(\sum_{j=k+1}^{\infty} \phi^H(j, k) C^H C \phi(j, k) \right) \phi(k, 0) \geq \beta_1 \Xi,$$

where the second inequality is due to (6.23). Then, the right hand side of the inequality of (6.21) trivially follows from (6.22). Similar to (6.22), the left hand side of (6.21) can be shown by using Fatou's Lemma [4]. $\blacksquare$

6.6.1 Proof of Theorem 6.3

Proof. On one hand, assume that $\sup_{k \in \mathbb{N}} \mathbb{E}[P_k] < \infty$. By (6.2), the Markov process has a unique stationary distribution given as follows,

$$\mathbb{P}\{\gamma_{\infty} = i\} = \lim_{k \rightarrow \infty} \mathbb{P}\{\gamma_k = i\} = \frac{p^{1-i} q^i}{p + q}, \forall i \in \mathbb{S}. \quad (6.24)$$

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

$$\Pi^- = (\mathbb{P}\{\gamma_k = j | \gamma_{k+1} = i\})_{i, j \in \mathbb{S}} = \begin{bmatrix} 1 - q & q \\ p & 1 - p \end{bmatrix}. \quad (6.25)$$

Given a measurable function $f : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^{n \times n}$, we obtain that:

$$\begin{aligned} \mathbb{E}[f(\gamma_k, \dots, \gamma_0)] &= \sum_{i_j \in \mathbb{S}, 0 \leq j \leq k} f(i_k, \dots, i_0) \mathbb{P}\{\gamma_k = i_k, \dots, \gamma_0 = i_0\} \\ &= \sum_{i_j \in \mathbb{S}, 0 \leq j \leq k} f(i_k, \dots, i_0) \mathbb{P}\{\gamma_0 = i_0\} \prod_{j=0}^{k-1} \mathbb{P}\{\gamma_{j+1} = i_{j+1} | \gamma_j = i_j\} \end{aligned} \quad (6.26)$$

$$= \sum_{i_j \in \mathbb{S}, 0 \leq j \leq k} f(i_k, \dots, i_0) \mathbb{P}\{\gamma_k = i_0\} \prod_{j=0}^{k-1} \mathbb{P}\{\gamma_j = i_{j+1} | \gamma_{j+1} = i_j\} \quad (6.27)$$

$$= \mathbb{E}[f(\gamma_0, \dots, \gamma_k)] = \mathbb{E}[f(\gamma_1, \dots, \gamma_{k+1})], \quad (6.28)$$

where (6.26) follows from the Markov property of $\{\gamma_k\}_{k \geq 0}$ while (6.27) is due to (6.2), (6.25) 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 [79], there exists a positive constant β_1 such that $P_{k+1} \geq \beta_1 (\sum_{i=1}^{k+1} \gamma_{k+1-i} (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1}$. Together with (6.28), we have that

$$\begin{aligned} \mathbb{E}[P_{k+1}] &\geq \beta_1 \mathbb{E} \left(\sum_{i=1}^{k+1} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1} \right)^{-1} \\ &\geq \beta_1 \mathbb{E} \left(\sum_{i=1}^{\infty} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1} \right)^{-1}. \end{aligned} \quad (6.29)$$

Under Assumption 6.2, (6.29) is decreasing w.r.t. k . Jointly with monotone convergence theorem [4], implies that $\sup_{k \in \mathbb{N}} \mathbb{E}[P_k] \geq \beta_1 \mathbb{E}[(\sum_{i=1}^{\infty} \gamma_i (A^{-i})^H C^H C A^{-i})^{-1}] \geq \beta_1 \mathbb{E}[\Xi^{-1}]$, where the last equality follows from the definition of Ξ in (6.19). Define a stopping time μ as the time at which the first packet is received, i.e., $\mu = \inf\{k \mid \gamma_k = 1, \forall k \in \mathbb{N}\}$. Since μ is a stopping time adapted to the Markov process $\{\gamma_k\}_{k \geq 0}$, we know that $\mathcal{G}_\mu \triangleq \sigma(\gamma_0, \dots, \gamma_\mu)$ is a well defined σ -field. By iterated conditioning, it follows that

$$\begin{aligned} \mathbb{E}[\Xi^{-1}] &= \mathbb{E}[A^\mu (\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} (A^\mu)^H] \\ &= \mathbb{E}[A^\mu \mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \mathcal{G}_\mu] (A^\mu)^H]. \end{aligned}$$

By (6.2), it is clear that γ_k is a strong Markov process [77]. This implies that

$$\mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \mathcal{G}_\mu] = \mathbb{E}[(\sum_{j=0}^{\infty} \gamma_{j+\mu} (A^{-j})^H C^H C A^{-j})^{-1} | \gamma_\mu].$$

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 6.4, it follows that $\sup_{k \in \mathbb{N}} \mathbb{E}^1[M_k] < \infty$. In view of Theorem 6.2, it implies that $\rho(A)^2(1-q) < 1$. This implies that $\mathbb{E}^0[A^\mu (A^\mu)^H] = q \sum_{i=1}^{\infty} A^i (A^i)^H (1-q)^{i-1} < \infty$. 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 6.1, 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 6.1 and 6.4, we obtain that $\mathbb{E}^1[\Xi^{-1}] < \infty$. By Theorem 6.1, 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 [79], there exists a positive 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 (6.28), we have that

$$\mathbb{E}[P_{k+1}] \leq \beta_2 \mathbb{E}(\sum_{i=1}^{k+1} \gamma_i (A^{-i})^H C^H C A^{-i} + (A^{-k-1})^H A^{-k-1})^{-1} = \beta_2 A \mathbb{E}[\Theta_k^{-1}] A^H, \quad (6.30)$$

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 (6.16). Similar to (6.23), there exists 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 (6.30), 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 (6.24), we obtain that $\mathbb{E}^0[P_k] < \infty$ and $\mathbb{E}^1[P_k] < \infty$ for all $k \in \mathbb{N}$. By Lemma 6.1, the proof is completed. $\blacksquare$

6.6.2 Proof of Theorem 6.4

Proof. Define the integer valued set $\mathcal{S}_d = \{kd | \forall k \in \mathbb{N}\}$ and $\theta = \sum_{j \in \mathcal{S}_d} \mathbb{P}\{t_1^* = j | \gamma_0 = 1\}$. Let $E_k, k \geq 1$ be a sequence of events defined as follows: $E_1 = \{t_1^* \notin \mathcal{S}_d\}$, $E_k \triangleq \{t_1^* \in \mathcal{S}_d, \dots, t_{k-1}^* \in \mathcal{S}_d, t_k^* \notin \mathcal{S}_d\}$, 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 $E_i \cap E_j = \emptyset$ 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 $\mathbb{P}(F|\gamma_0 = 1) = \mathbb{P}(\bigcup_{j=1}^{\infty} E_j|\gamma_0 = 1) = \sum_{j=1}^{\infty} \mathbb{P}(E_j|\gamma_0 = 1) = 1$. 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 [4], it follows that $\mathbb{E}^1[\Xi^{-1}] = \mathbb{E}^1[\Xi^{-1}1_F] = \mathbb{E}^1[\Xi^{-1}(\lim_{k \rightarrow \infty} 1_{F_k})] = \lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[\Xi^{-1}1_{E_j}]$.

Proof of part (a).

“ $\Leftarrow$.” By (6.19), 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

$$\begin{aligned} & \sum_{i=j-1}^j \phi^H(i, 0)C^H C \phi(i, 0) \\ &= \phi^H(j-1, 0) \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \begin{bmatrix} 1 + \lambda_1^{-2t_j^*} & 1 + \lambda_1^{-t_j^*} \lambda_2^{-t_j^*} \\ 1 + \lambda_1^{-t_j^*} \lambda_2^{-t_j^*} & 1 + \lambda_2^{-2t_j^*} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \phi(j-1, 0). \end{aligned} \quad (6.31)$$

Define $\Sigma_j = \begin{bmatrix} 1 + \lambda_1^{-2t_j^*} & 1 + \lambda_1^{-t_j^*} \lambda_2^{-t_j^*} \\ 1 + \lambda_1^{-t_j^*} \lambda_2^{-t_j^*} & 1 + \lambda_2^{-2t_j^*} \end{bmatrix}$, then if $t_j^* \notin \mathcal{S}_d$, it yields that $\Sigma_j^{-1} \leq \frac{4}{\lambda_1^{-2t_j^*} + \lambda_2^{-2t_j^*} - 2\lambda_1^{-t_j^*} \lambda_2^{-t_j^*}} I \leq \frac{2|\lambda_1|^{2t_j^*}}{1 - \cos(\frac{2\pi}{d})} I$. Let $c = \max(c_1^{-2}, c_2^{-2})$, it follows from (6.31) that if $t_j^* \notin \mathcal{S}_d$, then $(\sum_{i=j-1}^j \phi^H(i, 0)C^H C \phi(i, 0))^{-1} \leq \frac{2c|\lambda_1|^{2t_j}}{1 - \cos(\frac{2\pi}{d})} I$. Combining the above, we get that $\mathbb{E}^1[\Xi^{-1}] \leq \frac{2cI}{1 - \cos(\frac{2\pi}{d})} \lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}]$. By Lemma 5.1, the following statements are in force:

$$\begin{aligned} \lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}] &= \lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[(\prod_{i=1}^{j-1} |\lambda_1|^{2t_i^*} 1_{\{t_i^* \in \mathcal{S}_d\}}) |\lambda_1|^{2t_j^*} 1_{\{t_j^* \notin \mathcal{S}_d\}}] \\ &\leq \lim_{k \rightarrow \infty} \mathbb{E}^1[|\lambda_1|^{2t_1^*}] \sum_{j=1}^k (\mathbb{E}^1[|\lambda_1|^{2t_1^*} 1_{\{t_1^* \in \mathcal{S}_d\}}])^{j-1}, \end{aligned} \quad (6.32)$$

which is finite if and only if $\mathbb{E}^1[|\lambda_1|^{2t_1^*}] < \infty$ and $\mathbb{E}^1[|\lambda_1|^{2t_1^*} 1_{\{t_1^* \in \mathcal{S}_d\}}] < 1$. 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 $\mathbb{E}^1[|\lambda_1|^{2t_1^*}] < \infty$ and $\mathbb{E}^1[|\lambda_1|^{2t_1^*} 1_{\{t_1^* \in \mathcal{S}_d\}}] = \frac{pq}{(1-q)^2} \frac{(|\lambda_1|^2(1-q))^d}{1 - (|\lambda_1|^2(1-q))^d} < 1$. Then, we conclude that $\mathbb{E}^1[\Xi^{-1}] < \infty$. By Lemma 6.4, 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 6.1, 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 (6.23), it is easy to derive that $\Xi = \Xi'_{j-1} + \phi^H(j, 0)(C^H C + \sum_{i=j+1}^{\infty} \phi^H(i, j)C^H C \phi(i, j))\phi(j, 0) \leq \Xi'_{j-1} + \phi^H(j, 0)(C^H C + \beta_0 I)\phi(j, 0)$, where β_0 is given in (6.23). Let $\beta_5^{-1} = \max(\frac{1}{1-|\lambda_1|^{-2}}, 1, \beta_0)$, it follows that if $1_{E_j} = 1$, then

$$\begin{aligned} \Xi^{-1} &\geq (\Xi'_{j-1} + \phi^H(j, 0)(C^H C + \beta_0 I)\phi(j, 0))^{-1} \\ &= \left(\sum_{i=0}^{j-1} |\lambda_1|^{-2t_i} C^H C + \phi^H(j, 0)(C^H C + \beta_0 I)\phi(j, 0) \right)^{-1} \\ &\geq \left(\frac{1}{1-|\lambda_1|^{-2}} C^H C + \phi^H(j, 0)(C^H C + \beta_0 I)\phi(j, 0) \right)^{-1} \\ &\geq \beta_5 (C^H C + \phi^H(j, 0)(C^H C + I)\phi(j, 0))^{-1}. \end{aligned} \quad (6.33)$$

By the definition of the indicator function, it is clear that $\Xi^{-1}1_{E_j} \geq \beta_5(C^H C + \phi^H(j, 0)(C^H C + I)\phi(j, 0))^{-1}1_{E_j}$. In view of Lemma 6.4, 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 $\lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[(C^H C + \phi^H(j, 0)(C^H C + I)\phi(j, 0))^{-1}1_{E_j}] < \infty$. By some manipulations, there exists a positive constant $\beta_6 > 0$ such that $\text{tr}(C^H C + \phi^H(j, 0)(C^H C + I)\phi(j, 0))^{-1}1_{E_j} \geq \beta_6 |\lambda_1|^{2t_j} 1_{E_j}$. Thus, we obtain that

$$\lim_{k \rightarrow \infty} \sum_{j=1}^k \mathbb{E}^1[|\lambda_1|^{2t_j} 1_{E_j}] = \mathbb{E}^1[|\lambda_1|^2 1_{\{t_1^* \notin \mathcal{S}_d\}}] \lim_{k \rightarrow \infty} \sum_{j=1}^k (\mathbb{E}^1[|\lambda_1|^2 1_{\{t_1^* \in \mathcal{S}_d\}}])^{j-1} < \infty. \quad (6.34)$$

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 $\text{rank}(C) = 2$ or A has two eigenvalues but with distinct magnitudes, this indicates that (C, A) is a non-degenerate pair. It is proved in Theorem 6.5.
- (ii) If $\text{rank}(C) = 1$ and A contains two identical eigenvalues, it is proved in Theorem 6.6. Note that for this case, A can not be 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 $\text{rank}(C) = 1$ and $A = \text{diag}(\lambda_1, \lambda_2)$, where $\lambda_2 = \lambda_1 \exp(2\pi\varphi i)$ 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 $\lim_{k \rightarrow \infty} \varphi_k = \varphi$, the integers r_k and d_k are irreducible and d_k goes into infinity as $k \rightarrow \infty$. 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 6.1. ■

6.6.3 Proofs of Results in Section 6.4

Lemma 6.5. [79] *Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A and $|\lambda_1| \geq \dots \geq |\lambda_n|$. If the pair (C, A) satisfies Assumptions 6.2-6.3 and 6.5, then the following inequality holds*

$$\limsup_{\Delta_1, \dots, \Delta_n \rightarrow \infty} \frac{(\sum_{j=1}^n (A^{-k_j})^H C^H C A^{-k_j})^{-1}}{\prod_{j=1}^n |\lambda_j|^{2\Delta_j}} \leq \beta_7 I, \quad (6.35)$$

where β_7 is a positive constant, $k_1 < k_2 < \dots < k_n \in \mathbb{N}$, $\Delta_1 = k_1$, $\Delta_j = k_j - k_{j-1}$ for all $j \in \{2, \dots, n\}$.

Proof of Theorem 6.5: “ $\Leftarrow$ ” By Lemma 6.5, there exists a sufficiently large $\Delta > 0$ such that for all $\Delta_j > \Delta$, it holds

$$(\sum_{j=1}^n (A^{-k_j})^H C^H C A^{-k_j})^{-1} \leq \beta_7 \prod_{j=1}^n |\lambda_j|^{2\Delta_j} I.$$

Now, select $k_j = t_{i_j}$, where $i_1 > \Delta$, $i_j - i_{j-1} > \Delta$ for all $j \in \{2, \dots, n\}$ and t_{i_j} is a stopping time defined in (5.4). Then, it is obvious that $t_{i_j} - t_{i_{j-1}} \geq i_j - i_{j-1} > \Delta$, which jointly with (6.19), implies that

$$\mathbb{E}^1[\Xi^{-1}] \leq \beta_7 \mathbb{E}^1[\prod_{j=1}^n |\lambda_j|^{2\Delta_j}] I = \beta_7 \mathbb{E}^1[\prod_{j=1}^n |\lambda_j|^{2(t_{i_j} - t_{i_{j-1}})}] I < \infty, \quad (6.36)$$

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 6.4, 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 6.1.

“ $\Rightarrow$ ” It is proved in Theorem 6.1. ■

Lemma 6.6. *For any integer k_i such that $k_{i+1} > k_i$, let $\mathcal{B} \in \mathbb{R}^{n \times n}$ be a matrix with (i, j) -th element given by $\mathcal{B}_{ij} = \binom{k_i}{j-1}$. Then, the determinant of $\mathcal{B}$ is computed as*

$$\det(\mathcal{B}) = \frac{1}{\prod_{i=0}^{n-1} i!} \prod_{1 \leq j < i \leq n} (k_i - k_j),$$

where $i!$ is the factorial of a positive integer i .

Proof. It is clear that $\det(\mathcal{B})$ is an alternative, i.e., swapping the i -th and j -th rows is the same as changing values of k_i and k_j . Moreover, $\det(\mathcal{B})$ is an $(n-1)$ -th order multivariate polynomial in $k_1, \dots, k_n$. For example, $\det(\mathcal{B})$ 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 $\det(\mathcal{B})$ 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 $\det(\mathcal{B})$, modulo a constant α_n , due to that $\prod_{1 \leq j < i \leq n} (k_i - k_j)$ and $\det(\mathcal{B})$ are both $(n-1)$ -th order, from which we get the following equality:

$$\det(\mathcal{B}) = \alpha_n \prod_{1 \leq j < i \leq n} (k_i - k_j). \quad (6.37)$$

It remains to show that $\alpha_n = 1/\prod_{i=0}^{n-1} i!$. We do so by mathematical induction. For $n = 1$, $\det \mathcal{B}(k_1) = 1$. 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 $n = m$, suppose it holds that $\alpha_m = 1/\prod_{i=0}^{m-1} i!$. Then, for $n = m+1$, let $k_1, \dots, k_m$ be fixed and k_{m+1} go to infinity. Note that $\lim_{k_{m+1} \rightarrow \infty} \frac{\binom{k_{m+1}}{i}}{k_{m+1}^i} = \frac{1}{i!}$, we have that

$$\lim_{k_{m+1} \rightarrow \infty} \frac{\det(\mathcal{B}(k_1, \dots, k_{m+1}))}{k_{m+1}^m} = \lim_{k_{m+1} \rightarrow \infty} \left(\frac{\binom{k_{m+1}}{m} \det(\mathcal{B})}{k_{m+1}^m} + \sum_{i=0}^{m-1} \frac{O(k_{m+1}^i)}{k_{m+1}^m} \right)$$

$$= \det(\mathcal{B}(k_1, \dots, k_m))/m! = \frac{1}{\prod_{i=0}^m i!} \prod_{1 \leq j < i \leq m} (k_i - k_j), \quad (6.38)$$

where $O(k_{m+1}^i)$ in the first equality means that $\lim_{k_{m+1} \rightarrow \infty} \frac{O(k_{m+1}^i)}{k_{m+1}^i} < \infty$. In light of (6.37), it yields that $\lim_{k_{m+1} \rightarrow \infty} \frac{\det(\mathcal{B}(k_1, \dots, k_{m+1}))}{k_{m+1}^m} = \alpha_{m+1} \prod_{1 \leq j < i \leq m} (k_i - k_j)$.

Combining the above, we immediately obtain that $\alpha_{m+1} = 1/\prod_{i=0}^m i!$. Hence, $\alpha_n = 1/\prod_{i=0}^{n-1} i!$ holds for all $n \geq 1$. $\blacksquare$

Lemma 6.7. *For any integer k_i such that $k_{i+1} > k_i$, let $\mathcal{B}' \in \mathbb{R}^{n \times n}$ be a matrix such that the (i, j) -th element is given by $\mathcal{B}'_{ij} = \binom{k_i}{j}$. Then, the determinant of $\mathcal{B}'$ is computed as $\det(\mathcal{B}') = (\prod_{i=1}^n \frac{k_i}{i!}) \prod_{1 \leq j < i \leq n} (k_i - k_j)$.*

Proof. Similar to Lemma 6.6, $\det(\mathcal{B}')$ is an alternative and n -th order multivariate polynomial. It is straightforward that $\det(\mathcal{B}')$ contains $\prod_{i=1}^n k_i$ as a factor. Thus, we further obtain that $\det(\mathcal{B}')$ contains $\prod_{i=1}^n k_i \prod_{1 \leq j < i \leq n} (k_i - k_j)$ as a factor, which is also the only factor containing k_i . Hence, the following is in force: $\det(\mathcal{B}') = \alpha'_n (\prod_{i=1}^n k_i \prod_{1 \leq j < i \leq n} (k_i - k_j))$. Using similar induction arguments as in Lemma 6.6, one can easily show that $\alpha'_n = 1/\prod_{i=1}^n i!$. $\blacksquare$

Lemma 6.8. *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 $i \geq 2$. Denote $\mathcal{O}(\{k_i\}_1^{n-1}) = [C^H, (A^{-k_1})^H C^H, \dots, (A^{-k_{n-1}})^H C^H]^H$ and $D_\lambda(\{k_i\}_1^{n-1}) = \prod_{v=1}^m \lambda_v^{-k(v) + \frac{n_v(n_v-1)}{2}}$, where $k(1) = k_1 + \dots + k_{n_1-1}$ and $k(v) = k_{n_1+\dots+n_{v-1}} + \dots + k_{n_1+\dots+n_v-1}$ if $v \geq 2$. Under Assumptions 6.2 and 6.6, we can asymptotically compute the determinant of $\mathcal{O}(\{k_i\}_1^{n-1})$. 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*

$$\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{\det \mathcal{O}(\{k_i\}_1^{n-1})}{D_\lambda(\{k_i\}_1^{n-1}) \psi(\{k_i\}_1^{n-1})} = 1.$$

Proof. Under Assumption 6.6, partition the observation matrix C in conformity with the block diagonal matrix A . Let $C_i = [c_{i1}, \dots, c_{in_i}]$, it is easy to verify that

$$C_i N_i^k = [\underbrace{0, \dots, 0}_k, c_{i1}, \dots, c_{i(n_i-k)}]$$

for any $k \leq n_i - 1$. We further obtain that for any $k > n_i$,

$$\begin{aligned} C_i(\lambda_i^{-1}I_i + N_i)^k &= \left[\frac{c_{i1}}{\lambda_i^k}, \frac{c_{i2}}{\lambda_i^k} + \binom{k}{1} \frac{c_{i1}}{\lambda_i^{k-1}}, \dots, \sum_{j=0}^{n_i-1} \binom{k}{j} \frac{c_{i(n_i-j)}}{\lambda_i^{k-j}} \right] \\ &\triangleq \lambda_i^{-k} \left[1, \binom{k}{1}, \dots, \binom{k}{n_i-1} \right] \tilde{C}_i, \end{aligned}$$

where $\tilde{C}_i$ is defined as

$$\tilde{C}_i = \text{diag}(1, \lambda_i, \dots, \lambda_i^{n_i-1}) \begin{bmatrix} c_{i1} & c_{i2} & \dots & c_{in_i} \\ & c_{i1} & \dots & c_{i(n_i-1)} \\ & & \ddots & \\ & & & c_{i1} \end{bmatrix}$$

and $\det(\tilde{C}_i) = \lambda_i^{n_i(n_i-1)/2} c_{i1}^{n_i}$. By the above results, it follows that

$$\begin{aligned} \det \mathcal{O}(\{k_i\}_1^{n-1}) &= \det \begin{bmatrix} C_1 & \dots & C_m \\ C_1(\lambda_1^{-1}I_1 + N_1)^{k_1} & \dots & C_m(\lambda_m^{-1}I_m + N_m)^{k_1} \\ \ddots & & \ddots \\ C_1(\lambda_1^{-1}I_1 + N_1)^{k_{n-1}} & \dots & C_m(\lambda_m^{-1}I_m + N_m)^{k_{n-1}} \end{bmatrix} \\ &= \det \begin{bmatrix} 1 & 0 & \dots & 0 \\ \binom{k_1}{0} \lambda_1^{-k_1} & \binom{k_1}{1} \lambda_1^{-k_1} & \dots & \binom{k_1}{n_1-1} \lambda_1^{-k_1} \\ & & \ddots & \\ \binom{k_{n-1}}{0} \lambda_1^{-k_{n-1}} & \binom{k_{n-1}}{1} \lambda_1^{-k_{n-1}} & \dots & \binom{k_{n-1}}{n_1-1} \lambda_1^{-k_{n-1}} \\ \dots & 1 & 0 & \dots & 0 \\ \dots & \binom{k_1}{0} \lambda_m^{-k_1} & \binom{k_1}{1} \lambda_m^{-k_1} & \dots & \binom{k_1}{n_m-1} \lambda_m^{-k_1} \\ \ddots & & & \ddots & \\ \dots & \binom{k_{n-1}}{0} \lambda_m^{-k_{n-1}} & \binom{k_{n-1}}{1} \lambda_m^{-k_{n-1}} & \dots & \binom{k_{n-1}}{n_m-1} \lambda_m^{-k_{n-1}} \end{bmatrix} \\ &\quad \times \det(\text{diag}(\tilde{C}_1, \dots, \tilde{C}_m)) \\ &\triangleq (D_1 + \dots + D_m) \det(\text{diag}(\tilde{C}_1, \dots, \tilde{C}_m)), \end{aligned} \tag{6.39}$$

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 i -th block. For example, the first block consists of the first n_1 columns and the followed n_2 columns forms the second block.

Let $\sigma = [\sigma_1, \dots, \sigma_m]$ be a permutation of $\{1, \dots, n-1\}$ such that $\#\sigma_1 = n_1 - 1$, $\#\sigma_2 = n_2, \dots, \#\sigma_m = n_m$, where $\#\sigma_i$ denotes the order of the permutation σ_i .

Then, it follows from the Leibnitz formula [48] for the determinant of a matrix that

$$D_1 = \sum_{\sigma} \text{sgn}(\sigma) h(k_{\sigma_j}) \left(\prod_{j=1}^m \lambda_j^{-k_{\sigma_j}} \right), \quad (6.40)$$

where the signature of permutation σ is denoted as $\text{sgn}(\sigma)$, 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 $|\lambda_1| > \dots > |\lambda_m|$, it is clear that $\lambda^{\sigma} \triangleq |\prod_{j=1}^m \lambda_j^{-k_{\sigma_j}}|$ achieves the maximum when $\lambda_1^{-k_{\sigma_1}} = \lambda_1^{-\sum_{i \in \{1, \dots, n_1-1\}} k_i}$, $\lambda_2^{-k_{\sigma_2}} = \lambda_2^{-\sum_{i \in \{n_1, \dots, n_1+n_2-1\}} k_i}$, $\dots$, $\lambda_m^{-k_{\sigma_m}} = \lambda_m^{-\sum_{i \in \{n_1+\dots+n_{m-1}, \dots, n-1\}} k_i}$. Thus, denote the set of permutations having the above property by $\mathcal{P}_{\sigma}^*$. Given any permutation σ which does not belong to $\mathcal{P}_{\sigma}^*$, we always have $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{\lambda^{\sigma}}{\lambda^{\sigma^*}} = 0$ for all $\sigma^* \in \mathcal{P}_{\sigma}^*$ and $\sigma \notin \mathcal{P}_{\sigma}^*$. Consequently, $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{D_1}{D_{1\infty}} = 1$, where $D_{1\infty} = \prod_{j=1}^m D_{1j}$ and

$$\begin{aligned} D_{11} &= \det \begin{bmatrix} \binom{k_1}{1} \lambda_1^{-k_1} & \cdots & \binom{k_1}{n_1-1} \lambda_1^{-k_1} \\ & \ddots & \\ \binom{k_{n_1-1}}{1} \lambda_1^{-k_{n_1-1}} & \cdots & \binom{k_{n_1-1}}{n_1-1} \lambda_1^{-k_{n_1-1}} \end{bmatrix} \\ &= \left(\prod_{i=1}^{n_1-1} \frac{k_i}{i!} \right) \prod_{1 \leq j < i \leq n_1-1} (k_i - k_j) \lambda_1^{-k(1)}, \end{aligned}$$

where the last equality follows from Lemma 6.7. Using Lemma 6.6, we can similarly derive that

$$D_{12} = \det \begin{bmatrix} \binom{k_{n_1}}{0} \lambda_2^{-k_{n_1}} & \cdots & \binom{k_{n_1}}{n_2-1} \lambda_2^{-k_{n_1}} \\ & \ddots & \\ \binom{k_{n_1+n_2-1}}{0} \lambda_2^{-k_{n_1+n_2-1}} & \cdots & \binom{k_{n_1+n_2-1}}{n_2-1} \lambda_2^{-k_{n_1+n_2-1}} \end{bmatrix}$$

$$= \left(\prod_{i=0}^{n_2-1} \frac{1}{i!} \prod_{0 \leq j < i \leq n_2-1} (k_{n_1+i} - k_{n_1+j}) \right) \lambda_2^{-k(2)}.$$

Applying the same arguments to the rest of D_{1j} , we obtain that

$$\begin{aligned} D_{1\infty} &= \prod_{i=1}^{n_1-1} \frac{k_i}{i!} \left(\prod_{v=2}^m \prod_{i=0}^{n_v-1} \frac{1}{v!} \prod_{0 \leq j < i \leq n_v-1} (k_{n_1+\dots+n_{v-1}+i} - k_{n_1+\dots+n_{v-1}+j}) \right) \prod_{v=1}^m \lambda_v^{-k(v)} \\ &\triangleq \psi_1(\{k_i\}_1^{n-1}) D_\lambda^1(\{k_i\}_1^{n-1}), \end{aligned} \quad (6.41)$$

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 $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{D_2}{D_{2\infty}} = 1$ 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 $D_\lambda^2(\{k_i\}_1^{n-1}) = \frac{\prod_{v=3}^m \lambda_v^{-k(v)}}{\lambda_1^{k_1+\dots+k_{n_1}} \lambda_2^{k_{n_1+1}+\dots+k_{n_1+n_2-1}}}$. Then, it follows that $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{D_{2\infty}}{D_{1\infty}} = \left(\frac{\lambda_2}{\lambda_1} \right)^{k_{n_1}} \frac{\psi_2(\{k_i\}_1^{n-1})}{\psi_1(\{k_i\}_1^{n-1})} = 0$. 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 $D_3, \dots, D_m$, e.g., $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{D_{v\infty}}{D_{1\infty}} = 0, \forall v \geq 3$. To sum up, we finally get that $\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{D_1+\dots+D_m}{D_{1\infty}} = 1$. By setting $\psi(\{k_i\}_1^{n-1}) = (\prod_{i=1}^m c_{i1}^{n_i}) \psi_1(\{k_i\}_1^{n-1})$ 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 (6.39) and (6.41) that

$$\lim_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{\det \mathcal{O}(\{k_i\}_1^{n-1})}{\psi(\{k_i\}_1^{n-1}) D_\lambda(\{k_i\}_1^{n-1})} = 1,$$

which completes the proof. ■

Proof of Theorem 6.6: “ $\Leftarrow$ ” In order to simplify notation, we remove the dependence of $\{k_i\}_1^{n-1}$ for quantities in Lemma 6.8, i.e., rewrite $\mathcal{O}(\{k_i\}_1^{n-1})$ as $\mathcal{O}$. Then, it yields that

$$(\mathcal{O}^H \mathcal{O})^{-1} \leq \text{tr}(\mathcal{O}^H \mathcal{O})^{-1} I = \sum_{i,j} \left(\frac{[\text{adj}(\mathcal{O})]_{ij}}{\det(\mathcal{O})} \right)^2 I. \quad (6.42)$$

Here $\text{adj}(\mathcal{O})$ is the adjoint matrix of $\mathcal{O}$ and $[\text{adj}(\mathcal{O})]_{ij}$ is the (i, j) -th element of $\text{adj}(\mathcal{O})$. Following a similar line of Lemma 6.8, we can show that there exist constant numbers $\beta_{i,j} = \beta_{i,j}(\lambda_1, \dots, \lambda_m)$ such that for sufficiently large Δ_j , we have that $\text{adj}(\mathcal{O})_{ij} \leq \beta_{i,j} |\psi| \prod_{v=1}^m |\lambda_v|^{-k'(v)}$, where $k'(v) = k_1 + \dots + k_{n_1-2}$ and

$k'(v) = k_{n_1+\dots+n_{v-1}-1} + \dots + k_{n_1+\dots+n_v-2}$ if $v \geq 2$. In light of (6.42), it follows that there exist constant numbers $\tilde{\beta}_{ij} = \tilde{\beta}_{ij}(\lambda_1, \dots, \lambda_m)$ such that

$$\limsup_{\Delta_1, \dots, \Delta_{n-1} \rightarrow \infty} \frac{(\mathcal{O}^H \mathcal{O})^{-1}}{\prod_{i=1}^m |\lambda_v|^{2\Delta(v)}} \leq \left(\sum_{i,j} \tilde{\beta}_{i,j} \right) I, \quad (6.43)$$

where $\Delta(1) = \Delta_1 + \dots + \Delta_{n-1}$ and $\Delta(v) = \Delta_{n_1+\dots+n_{v-1}} + \dots + \Delta_{n_1+\dots+n_v-1}$, $v \geq 2$.

The rest of the proof directly follows from that of Theorem 6.5.

“ $\Rightarrow$:” It is proved in Theorem 6.1.

6.7 Summary

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.

Chapter 7

Quantized Filtering of Linear Stochastic Systems

In Chapter 6, the effect of packet loss on the stability of Kalman filtering was examined. This chapter proposes 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 mini-max quantized filter when the innovation covariance is not exactly known. Simulation results illustrate the effectiveness and advantages of the proposed quantized filter.

This chapter is organized in the following fashion. The problem is formulated in Section 7.1. The multi-level quantized innovations Kalman filter is derived in Section 7.2, where we first derive the filter based on the general quantization scheme and then proceed to seek the optimal quantization. In Section 7.3, a max-min quantization optimization is described to address the robust quantization problem. Simulation and experiments are carried out in Section 7.3 to illustrate the performance of the

proposed filter. Some concluding remarks are drawn in Section 7.5.

7.1 Problem Formulation

Consider the following discrete-time linear stochastic system:

$$x_{k+1} = Ax_k + w_k, \quad (7.1)$$

$$y_k = Cx_k + v_k, \quad (7.2)$$

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 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. 7.1) 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 $\epsilon_k = \eta_k/\sigma_k$. Due to the symmetric property of the standard Gaussian distribution, we consider the associated symmetric quantizer (denoted by ‘ Q ’ in Fig. 7.1) with *quantization thresholds* $\{\pm z_i\}_{i=1}^N$ and $0 = z_0 < z_1 < \dots < z_N < \infty = z_{N+1}$, namely, the output b_k of the quantizer $Q(\cdot)$ is given by

$$b_k \triangleq Q(\epsilon_k) = \begin{cases} z_N, & \epsilon_k < -z_N \\ \vdots & \vdots \\ z_0, & -z_0 < \epsilon_k \leq -z_1 \\ -q(-\epsilon_k), & \epsilon_k \leq 0. \end{cases} \quad (7.3)$$

Note that when $b_k = z_0$, it will not be transmitted to the EC. That is, for a 1-bit

Figure 7.1: Network configuration

budget, the quantizer has 3 quantization levels unlike that in [97] 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. Let $\mathbf{b}_{0:k} \triangleq \{b_0, b_1, \dots, b_k\}$ and $\hat{x}_{k|k}$ represent the MMSE estimate of x_k given $\mathbf{b}_{0:k}$, i.e.,

$$\hat{x}_{k|k} \triangleq \mathbb{E}[x_k | \mathbf{b}_{0:k}] = \int_{\mathbb{R}^n} x_k p[x_k | \mathbf{b}_{0:k}] dx_k. \quad (7.4)$$

The following equalities can be easily obtained [97]:

$$\hat{x}_{k|k-1} \triangleq \mathbb{E}[x_k | \mathbf{b}_{0:k-1}] = A\hat{x}_{k-1|k-1}, \quad (7.5)$$

$$\hat{y}_{k|k-1} \triangleq \mathbb{E}[y_k | \mathbf{b}_{0:k-1}] = C\hat{x}_{k|k-1}. \quad (7.6)$$

Their filtering error covariance matrices are respectively defined by

$$P_{k|k} \triangleq \mathbb{E}[(\hat{x}_{k|k} - x_k)(\hat{x}_{k|k} - x_k)^T], \quad (7.7)$$

$$P_{k|k-1} \triangleq \mathbb{E}[(\hat{x}_{k|k-1} - x_k)(\hat{x}_{k|k-1} - x_k)^T] = AP_{k-1|k-1}A^T + Q. \quad (7.8)$$

7.2 Multi-level 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 quan-

tization 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 $z_1^N = (z_1, z_2, \dots, z_N)$.

7.2.1 Multi-level Quantized Filtering

Theorem 7.1. *Consider the systems (7.1) and (7.2), given a multi-level quantization of normalized innovations in (7.3), 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*

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \frac{f(z_1^N, b_k) P_{k|k-1} C^T}{\sqrt{C P_{k|k-1} C^T + R}}, \quad (7.9)$$

$$P_{k|k} = P_{k|k-1} - F(z_1^N) \frac{P_{k|k-1} C^T C P_{k|k-1}}{C P_{k|k-1} C^T + R}, \quad (7.10)$$

with

$$f(z_1^N, b_k) = \sum_{j=-N}^N 1_{\{z_j\}}(b_k) \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})}, \quad (7.11)$$

$$F(z_1^N) = 2 \sum_{j=0}^N \frac{[\phi(z_j) - \phi(z_{j+1})]^2}{T(z_j) - T(z_{j+1})}, \quad (7.12)$$

where $z_{-j} = -z_j$. $\phi(\cdot)$ 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

$$\hat{x}_{k|k} = \mathbb{E}[x_k | \mathbf{b}_{1:k}] = \mathbb{E}[\mathbb{E}[x_k | \mathbf{b}_{1:k-1}, y_k] | \mathbf{b}_{1:k}]. \quad (7.13)$$

And the posterior density $p[x_k | \mathbf{b}_{1:k-1}, y_k]$ is obtained as

$$p[x_k | \mathbf{b}_{1:k-1}, y_k] = \frac{p(y_k | x_k) p(x_k | \mathbf{b}_{1:k-1})}{\int_{\mathbb{R}^n} p(y_k | x_k) p(x_k | \mathbf{b}_{1:k-1}) dx_k}. \quad (7.14)$$

By the Gaussian assumption and following the technique of the Kalman filter, the

inner conditional expectation in (7.13) can be easily obtained as follows:

$$\hat{x}_{k|k}^* \triangleq \mathbb{E}[x_k | \mathbf{b}_{1:k-1}, y_k] = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}), \quad (7.15)$$

where $K_k = (C^T P_{k|k-1} C + R)^{-1} P_{k|k-1} C^T$. By (7.3) and Gaussian assumption, it follows that

$$\mathbb{E}[\epsilon_k | b_k = z_j, \mathbf{b}_{1:k-1}] = \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})}, \quad (7.16)$$

Note that $\hat{x}_{k|k-1}$ is measurable with respect to the σ -algebra generated by $\mathbf{b}_{1:k}$. Inserting (7.15) and (7.16) into (7.13) yields (7.9).

Applying the same arguments, we have that

$$P_{k|k} = \mathbb{E}[\mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T | \mathbf{b}_{1:k-1}, y_k]], \quad (7.17)$$

where the inner expectation is computed as follows:

$$\begin{aligned} & \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T | \mathbf{b}_{1:k-1}, y_k] \\ &= \mathbb{E}[(x_k - \hat{x}_{k|k}^*)(x_k - \hat{x}_{k|k}^*)^T | \mathbf{b}_{1:k-1}, y_k] + \mathbb{E}[(\hat{x}_{k|k} - \hat{x}_{k|k}^*)(\hat{x}_{k|k} - \hat{x}_{k|k}^*)^T | \mathbf{b}_{1:k-1}, y_k] \\ &= P_{k|k}^* + \mathbb{E}[(\hat{x}_{k|k} - \hat{x}_{k|k}^*)(\hat{x}_{k|k} - \hat{x}_{k|k}^*)^T | \mathbf{b}_{1:k-1}, y_k], \end{aligned} \quad (7.18)$$

where $P_{k|k}^* = P_{k|k-1} - P_{k|k-1} C^T (C P_{k|k-1} C^T + R)^{-1} C P_{k|k-1}$. Under the Gaussian assumption, we have that

$$\begin{aligned} & \mathbb{E}[(\hat{x}_{k|k} - \hat{x}_{k|k}^*)(\hat{x}_{k|k} - \hat{x}_{k|k}^*)^T] = \mathbb{E}[\mathbb{E}[(\hat{x}_{k|k} - \hat{x}_{k|k}^*)(\hat{x}_{k|k} - \hat{x}_{k|k}^*)^T | \mathbf{b}_{1:k-1}]] \\ &= (C P_{k|k-1} C^T + R) K_k \mathbb{E}[\mathbb{E}[(\epsilon_k - f(z_1^N, b_k))^2 | \mathbf{b}_{1:k-1}]] K_k^T \\ &= \frac{P_{k|k-1} C^T C P_{k|k-1}}{C P_{k|k-1} C^T + R} (1 - \mathbb{E}[\mathbb{E}[f^2(z_1^N, b_k) | \mathbf{b}_{1:k-1}]]) \\ &= \frac{P_{k|k-1} C^T C P_{k|k-1}}{C P_{k|k-1} C^T + R} (1 - F(z_1^N)). \end{aligned} \quad (7.19)$$

Together with (7.17) and (7.18), one can easily derive (7.10). ■

Remark 7.1. *Strictly speaking, the system state conditioned on the quantized innovations can not remain Gaussian due to the nonlinear operator of quantization. To*

Table 7.1: Solutions to (7.20) for optimal quantization thresholds

N=1	N=2	N=3	N=4
$z_1^* = 0.612$	$z_1^* = 0.382$ $z_2^* = 1.244$	$z_1^* = 0.280$ $z_2^* = 0.874$ $z_3^* = 1.611$	$z_1^* = 0.221$ $z_2^* = 0.681$ $z_3^* = 1.198$ $z_4^* = 1.866$
$F(z_1^*) = 0.810$	$F((z^*)_1^2) = 0.920$	$F((z^*)_1^3) = 0.956$	$F((z^*)_1^4) = 0.972$

enable the development of a simple and practically useful recursive filter, we have made the Gaussian assumption as in [97]. A similar hypothesis can also be found in [58]. 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 (7.12) 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*.

7.2.2 Optimal Quantization Thresholds

The performance of the MLQ-KF can be approximately measured by the quantity of $\text{tr}[P_{k|k}]$ or $\text{tr}[P_{k|k-1}]$. According to Theorem 7.1, the optimal thresholds $(z^*)_1^N$ of the quantizer in (7.3) 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,

$$[x, y] = \text{fmincon}(\text{fun}, x_0, A, b, Aeq, beq, lb, ub). \quad (7.20)$$

Thus, the optimal quantizer can be numerically obtained, see Table 7.1. 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. 7.2. Apparently, for single bit quantization, the quantized filter with dead zone gives a much improved performance

Figure 7.2: $F((z^*)_1^N)$ versus R bits

recovery factor (0.8098) than the SOI-KF [97] ($2/\pi \approx 0.6366$). Fig. 7.2 also indicates that the higher the bit rate, the higher the performance recovery factor. The relationship given in Fig. 7.2 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:

$$P_{k+1} = AP_k A^T + Q - \beta^* AP_k C^T (CP_k C^T + R)^{-1} CP_k A^T \quad (7.21)$$

with $\beta^* = F((z^*)_1^N)$. The corresponding algebraic Riccati equation (ARE) is

$$P = APA^T + Q - \beta^* APC^T (CPC^T + R)^{-1} CPA^T. \quad (7.22)$$

Corollary 7.1. *Consider MRDE (7.21) and assume that $(A, Q^{1/2})$ and (C, A) are respectively controllable and detectable. Then for an unstable A , the ARE (7.22) has a positive definite solution P and the MRDE (7.21) admits a unique positive definite solution P_k satisfying $P_k \rightarrow P$ 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 7.1 reveals that for systems with faster growth rate, a higher bits number is required for the quantizer to ensure the convergence of MRDE.

7.2.3 Convergence Analysis

The following result establishes the convergence of the MLQ-KF to the standard Kalman filter when $N \rightarrow \infty$.

Theorem 7.2. Let $\Delta = \sup_{j \in \mathbb{N}} \Delta_j$, where $\Delta_j = |z_j - z_{j+1}|$ and assume that the quantization thresholds in (7.3) satisfy

1. $\Delta_j \leq \Delta \rightarrow 0$;
2. $S(N) = \sum_{k=1}^{N-1} \Delta_j \rightarrow \infty$ as $N \rightarrow \infty$.

Then, it follows that

$$\frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})} \rightarrow z_{j+1}. \quad (7.23)$$

and

$$F(z_1^N) \rightarrow 1 \text{ as } N \rightarrow \infty. \quad (7.24)$$

Proof. Using some basic results in mathematical analysis, we obtain that

$$\begin{aligned} \lim_{z_j \rightarrow z_{j+1}} \frac{\phi(z_j) - \phi(z_{j+1})}{T(z_j) - T(z_{j+1})} &= \lim_{\Delta_j \rightarrow 0} \frac{\phi(z_{j+1} - \Delta_j) - \phi(z_{j+1})}{\int_{z_{j+1}-\Delta_j}^{z_{j+1}} \phi(t) dt} \\ &= \lim_{\Delta_j \rightarrow 0} \frac{\phi'(z_{j+1} - \Delta_j)}{\phi(z_{j+1} - \Delta_j)} = z_{j+1}. \\ S &\triangleq \sum_{k=1}^{\infty} \frac{[\phi(z_j) - \phi(z_{j+1})]^2}{T(z_j) - T(z_{j+1})} = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{[e^{(-\frac{z_j^2}{2})} - e^{(-\frac{z_{j+1}^2}{2})}]^2}{\int_{z_j}^{z_{j+1}} e^{(-\frac{t^2}{2})} dt} \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{[e^{(-\frac{z_j^2}{2})} - e^{(-\frac{(z_j+\Delta_j)^2}{2})}]^2}{e^{-\frac{(z_j+\theta_j\Delta_j)^2}{2}} \Delta_j} \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} e^{-\frac{z_j^2}{2}} \frac{[1 - e^{(-\frac{\Delta_j^2}{2} - z_j\Delta_j)]^2}{\Delta_j e^{-\frac{(\theta_j\Delta_j)^2}{2} - \theta_j z_j \Delta_j}}, \end{aligned}$$

where $0 \leq \theta_j \leq 1$. Using the Taylor expansion, we have that $S \triangleq S_0 + S_1$, where $S_1 = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{\infty} z_j^2 e^{-\frac{z_j^2}{2}} \Delta_j \rightarrow \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^2 e^{-t^2/2} dt = \frac{1}{2}$ as $\Delta \rightarrow 0$, and there exists $|c_{i,j,v,u}| < \infty$ such that $S_0 = \frac{1}{\sqrt{2\pi}} \sum_{i,j,u,v} \sum_{k=1}^{\infty} [c_{i,j,v,u} e^{(-\frac{z_k^2}{2})} \Delta_k^{2+i} z_k^j \theta_j^u o(\Delta_k^v)] < \Delta \sum_{i,j,u,v} c_{i,j,v,u} \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} z_k^j e^{(-\frac{z_k^2}{2})} \Delta_k \triangleq C\Delta \rightarrow 0$ as $\Delta \rightarrow 0$, where C is a finite constant since for $\Delta \rightarrow 0$, $\frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} z_k^j e^{-\frac{z_k^2}{2}} \Delta z_k \rightarrow \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^j e^{-t^2/2} dt < \infty$ and the nonnegative integers i, j, u, v can only take a finite number of elements. ■

Remark 7.2. As $z_j \rightarrow z_{j+1}$, then $\epsilon_k \rightarrow z_{j+1}$ and $f(z_1^N, b_k) \rightarrow \epsilon_k$ by (7.23). In light of (7.9), (7.10) and (7.24), the MLQ-KF converges to the standard Kalman filter

since by Theorem 7.2, we obtain that

$$\hat{x}_{k|k} \rightarrow \hat{x}_{k|k-1} + \frac{P_{k|k-1}C^T\eta_k}{CP_{k|k-1}C^T + R} \text{ and } P_{k|k} \rightarrow P_{k|k-1} - \frac{P_{k|k-1}C^TCP_{k|k-1}}{CP_{k|k-1}C^T + R}.$$

7.3 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 every time instant due to 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. $\underline{\delta}$, $\bar{\delta}$ are known lower and upper bounds of δ_k . Denote $\hat{\epsilon}_k = \eta_k/\sigma_{k,e}$, the quantization scheme (7.3) is modified as:

$$b_k \triangleq Q(\hat{\epsilon}_k) = \begin{cases} z_N, & \delta z_N < \hat{\epsilon}_k \\ \vdots & \vdots \\ z_0, & 0 < \hat{\epsilon}(n) \leq \delta z_1 \\ -q(\hat{\epsilon}_k), & \hat{\epsilon}(n) \leq 0. \end{cases} \quad (7.25)$$

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

$$\frac{1}{\max_{z_1^N} \min\{F(\underline{\delta} z_1^N), F(\bar{\delta} z_1^N)\}} = \min_{z_1^N} \max \left\{ \frac{1}{F(\underline{\delta} z_1^N)}, \frac{1}{F(\bar{\delta} z_1^N)} \right\}.$$

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 $\bar{\delta} \rightarrow \infty$, the robust quantized filter will reduce to the SOI-KF.

Figure 7.3: Comparison of the error variance of the position estimate given by 1-LQ-KF and SOI-KF based on 500 samples.

7.4 Simulation

Consider the following discrete time equations of motion for target tracking [102]:

$$x_{k+1} = \begin{bmatrix} 1 & \tau & \tau^2/2 \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} x_k + w_k, \quad (7.26)$$

$$y_k = [1 \ 0 \ 0] x_k + v_k, \quad (7.27)$$

where x_k is the state vector with each 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 7.1 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 is x_0

Figure 7.4: Comparison of the position error variance computed by (7.21) and Monte Carlo method based on Monte Carlo methods with 500 samples.

is a random vector with zero mean and covariance matrix [102]:

$$P_{0|0} = \begin{bmatrix} \sigma_R^2 & \sigma_R^2/\tau & 0 \\ \sigma_R^2/\tau & 2\sigma_R^2/\tau^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The filtering error variances of the position for 1-LQ-KF and SOI-KF are compared in Fig. 7.3, which shows that 1-LQ-KF outperforms SOI-KF. Fig. 7.4 illustrates that for the 2-bit quantization, the computed variance by (7.21) 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 distributed over $[0.8, 1.2]$. We design a 2-bit robust mini-max quantizer (2-LQ-RKF) according to (7.25) with $(\bar{z}^*)_1^2 = [0.3035 \ 1.0094]$. For comparison, a non-robust quantizer (nominal) quantizer given in (7.3) is implemented as well. Fig. 7.5 shows that 2-LQ-RKF has a better performance than the non-robust one.

7.5 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

Figure 7.5: Comparison of the position error variances by the robust and non-robust 2-bit quantized filters.

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.

Part II

**Networked Multi-agent Consensus
Control**

Chapter 8

Consensusability of Discrete-time Multi-agent Systems

In part I, all the results were established by exclusively focusing on a single plant. From now on, the system to be studied is a system of multiple subsystems, which is referred to as multi-agent system. This chapter investigates the joint effect of agent dynamics, network topology and communication data rate on consensusability of linear discrete-time multi-agent systems. Neglecting the finite communication data rate constraint and under *undirected* graphs, a necessary and sufficient condition for consensusability under a common control protocol is given, which explicitly reveals how the intrinsic entropy rate of the agent dynamics and the communication graph jointly affect consensusability. The result is established by solving a discrete-time simultaneous stabilization problem. A lower bound of the optimal convergence rate to consensus, which is shown to be tight for some special cases, is provided as well. Moreover, a necessary and sufficient condition for formationability of multi-agent systems is obtained. As a special case, the discrete-time second-order consensus is discussed where an optimal control gain is designed to achieve the fastest convergence. The effects of *undirected* graphs on consensusability/formationability and optimal convergence rate are exactly quantified by the ratio of the second smallest to the largest eigenvalues of the graph Laplacian matrix. An extension to *directed* graphs is also made. The consensus problem under a finite communication data rate

is finally investigated.

The remainder of this chapter is organized as follows. Section 8.1 introduces some basic concepts of algebraic graph theory and formulates the problem under investigation. Section 8.2 firstly presents a necessary and sufficient condition to achieve consensusability w.r.t. a common control protocol for *undirected* graphs, which is then generalized to *directed* graphs. The convergence rate to consensus is also analyzed. The result is then applied to study the formationability problem over *undirected* graphs. A necessary and sufficient condition in relation to the desired formation, agent dynamics and communication graph is provided. Section 8.3 discusses the discrete-time second-order consensus problem. To understand the quantization effects on consensusability, Section 8.4 gives a sufficient condition on the finite communication data rate for consensus on *undirected* graphs. Concluding remarks are drawn in Section 8.5.

8.1 Problem Formulation

8.1.1 Communication Graph

Denote $\mathcal{N} = \{1, \dots, N\}$ and let $\mathcal{V} = \{v_1, \dots, v_N\}$ be an index set of N agents with $i \in \mathcal{N}$ representing the i -th agent. A digraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{A}\}$ will be utilized to model the interactions among agents, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set of paired agents and $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N \times N}$ with nonnegative elements is the weighted adjacency matrix of $\mathcal{G}$. Self-edges (i, i) are not allowed, i.e., $(i, i) \notin \mathcal{E}$ for all $i \in \mathcal{N}$. An edge $(j, i) \in \mathcal{E}$ if and only if $a_{ij} > 0$, which means that agent j can send information to agent i . A sequence of edges $(i_1, i_2), (i_2, i_3), \dots, (i_{k-1}, i_k)$ with $(i_{j-1}, i_j) \in \mathcal{E}$ for all $j \in \{2, \dots, k\}$ is called a directed path from agent i_1 to agent i_k . $\mathcal{G}$ contains a spanning tree if there is a root agent that can send information to all the other agents via directed paths. $\mathcal{G}$ is called a strongly connected digraph if for any two agents $i, j \in \mathcal{V}$, there exists a directed path from agent i to agent j . If $\mathcal{A}$ is a symmetric matrix, $\mathcal{G}$ is called an *undirected* graph. A strongly connected undirected graph is simply called a connected graph. For an *undirected* graph $\mathcal{G}$, it is clear that $\mathcal{G}$ contains a spanning tree is equivalent

to that $\mathcal{G}$ is connected. A digraph is called *complete* if each pair of agents can directly connect to each other, i.e. $(i, j) \in \mathcal{E}, \forall i \neq j$. The neighborhood of the i -th agent is denoted by $\mathcal{N}_i \triangleq \{j | (j, i) \in \mathcal{E}\}$. The in-degree of agent i is represented by $\deg_i = \sum_{j=1}^N a_{ij}$. Denote $\mathcal{D} \triangleq \text{diag}(\deg_1, \dots, \deg_N)$ and the Laplacian matrix of $\mathcal{G}$ by $\mathcal{L}_{\mathcal{G}} = \mathcal{D} - \mathcal{A}$. The eigenvalues of $\mathcal{L}_{\mathcal{G}}$ are denoted by $\lambda_j \in \mathbb{C}, j \in \mathcal{N}$ and an ascending order in magnitude is written as $0 = |\lambda_1| \leq |\lambda_2| \leq \dots \leq |\lambda_N|$. Note that for an undirected graph $\mathcal{G}$, $\mathcal{L}_{\mathcal{G}}$ is a symmetric positive semi-definite matrix and for all $j \in \mathcal{N}$, $\lambda_j \geq 0$ [40].

Lemma 8.1. [34] *Let the adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ of an undirected graph $\mathcal{G}$ be a symmetric $(0, 1)$ -matrix, i.e. $a_{ij} = 1$ if $(i, j) \in \mathcal{E}$; $a_{ij} = 0$, otherwise, then $\mathcal{G}$ is complete if and only if $\mathcal{G}$ is connected and $\lambda_2 = \lambda_N$.*

Lemma 8.2. [40, 93] *Let $\mathcal{G}$ be a digraph, then all the nonzero eigenvalues of $\mathcal{L}_{\mathcal{G}}$ are in the open right half plane. Moreover, $\mathcal{G}$ has a spanning tree if and only if $\mathcal{L}_{\mathcal{G}}$ contains exactly one zero eigenvalue.*

8.1.2 Consensusability on Graphs

The dynamic of agent i in discrete time takes the following form:

$$x_i(k+1) = Ax_i(k) + Bu_i(k), k = 0, 1, \dots \quad (8.1)$$

$x_i(t) \in \mathbb{R}^n$ and $u_i(t) \in \mathbb{R}^l$ respectively represent the state and control input of agent i . $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times l}$ are the state and input matrices, respectively. Moreover, B is of full column rank, i.e., $\text{rank}(B) = l \leq n$. Without loss of generality, assume that $A = \text{diag}(A_s, A_u)$, where $A_s \in \mathbb{R}^{n_s \times n_s}$ is stable and all the eigenvalues of A_u lie on or outside the unit circle. Accordingly, let $B^T = [B_s^T, B_u^T]$.

In [71], a weighted-average protocol using relative state feedback was proposed:

$$u_i(k) = K \sum_{j=1}^N a_{ij}(x_j(k) - x_i(k)), k = 0, 1, \dots \quad (8.2)$$

where $K \in \mathbb{R}^{l \times n}$ is a fixed control gain and independent of the agent index i . Since

Figure 8.1: Encoder/decoder scheme.

$a_{ij} = 0$ for all $j \notin \mathcal{N}_i$, the control input of agent i only uses relative state information to its neighbors.

Definition 8.1. *The discrete-time multi-agent systems (8.1) with a fixed graph $\mathcal{G}$ are said to be consensusable under protocol (8.2) if for any finite $x_i(0), \forall i \in \mathcal{N}$, there exists a control gain $K \in \mathbb{R}^{l \times n}$ such that the controller of (8.2) enforces consensus, i.e.,*

$$\lim_{k \rightarrow \infty} \|x_i(k) - x_j(k)\| = 0, \forall i, j \in \mathcal{N}. \quad (8.3)$$

The above defined consensusability does not require the convergence of each agent. However, it contains the prevailing first-order average consensus [88] which corresponds to $A = B = K = 1$ and the second-order consensus [65] as special cases.

In the real world, the communication channel between each pair of agents has a limited bandwidth. Agents can only exchange information with a finite number of bits at each transmission and recover its neighbors' state using the quantized information. Thus, the consensusability problem based on the quantized information from neighbors is of theoretical and practical significance. To this purpose, denote $\hat{x}_j^i(k)$ the state estimator of agent j by agent i based on the quantized information that it has received up to time k , see Fig. 8.1 for an illustration, where encoder and decoder will be exactly described in Section 8.4. The control protocol of the following form will be investigated:

$$u_i(k) = K \sum_{j=1}^N a_{ij} (\hat{x}_j^i(k) - \hat{x}_i^i(k)). \quad (8.4)$$

The main objective of this chapter is to address the following problems:

- (a) Assume that there is no constraint on bandwidth, what is the necessary and sufficient condition on the communication graph $\mathcal{G}$ and the pair (A, B) for

consensusability of the discrete-time multi-agent systems (8.1) under protocol (8.2)?

- (b) Does there exist a coding/decoding scheme using a finite communication data rate for the controller (8.4) to enforce consensus for the multi-agent systems (8.1)? And what is the required data rate in relation to the agent dynamics and the communication graph for reaching a consensus?

The first problem reveals the joint effect of the agent dynamics and $\mathcal{G}$ on consensusability under protocol (8.2). The answer to the second problem is positive as shown later and indicates that consensusability with perfect state feedback implies consensusability with quantized state feedback provided that data rate is not less than a lower bound.

8.2 Consensusability with Perfect State Feedback

In this section, assuming that there is no limitation on communication bandwidth, we derive a necessary and sufficient condition on graphs for consensusability of the multi-agent systems (8.1) for single input system, which is essentially established by solving a discrete-time simultaneous stabilization problem. The convergence rate to consensus is evaluated and a lower bound for the optimal *asymptotic convergence rate* is given. The results are applied to investigate the formationability problem on undirected graphs. Surprisingly, all conditions on undirected graphs are completely characterized by the ratio of the second smallest to the largest eigenvalues of the corresponding graph Laplacian matrix. At last, extensions to multiple inputs and *directed* graphs (digraphs) are made as well.

8.2.1 Single Input and Undirected Graphs

We start with single input agent dynamics and *undirected* graphs which imply that the eigenvalues of the associated Laplacian matrix are real and nonnegative, e.g., for all $j \in \mathcal{N}$, $\lambda_j \geq 0$.

Assumption 8.1. *All the eigenvalues of A lie on or outside the unit circle and $\text{rank}(B) = 1$.*

Theorem 8.1. *Given a fixed undirected graph $\mathcal{G}$ and under Assumption 8.1, the discrete-time multi-agent systems (8.1) are consensusable under protocol (8.2) if and only if the following conditions hold.*

- (a) (A, B) is a controllable pair;
- (b) Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the following strict inequality:

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}, \quad (8.5)$$

where $\lambda_j^u(A)$ represents an unstable eigenvalue of A . λ_2 and λ_N are respectively the second smallest and largest eigenvalues of $\mathcal{L}_{\mathcal{G}}$.

Moreover, if the above conditions hold, select a ζ such that $\prod_j |\lambda_j^u(A)| < \zeta^{-1} \leq \frac{1+\lambda_2/\lambda_N}{1-\lambda_2/\lambda_N}$. Then, the control gain $K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B}$ solves the consensus problem, where $P > 0$ is a positive solution to the modified algebraic Riccati inequality:

$$P - A^T P A + (1 - \zeta^2) \frac{A^T P B B^T P A}{B^T P B} > 0. \quad (8.6)$$

Remark 8.1.

1. The existence of a positive solution P to (8.6) is proved in [28, 36]. In the rest of the paper, λ_2/λ_N is termed as the eigenratio of an undirected graph. By Lemmas A.1-A.2 [60], an upper bound of the eigenratio is immediately obtained, i.e.,

$$\frac{\lambda_2}{\lambda_N} \leq \frac{\min_i \deg_i}{\max_i \deg_i}.$$

2. For the average consensus problem in [88], the state of each agent is scalar and $A = B = K = 1$. The condition in item (a) is automatically satisfied while the inequality of (8.5) implies that $\lambda_2 > 0$. By Lemma 8.2, the communication

graph has to be connected which is consistent with the result in [88]. Thus, our result contains the classical average consensus as a special case.

3. *The inequality (8.5) implies that $\lambda_2 > 0$. Then, the graph is connected. In contrast with the result on continuous-time systems in [71], the case of discrete-time systems has an additional constraint given in (8.5). The eigenratio λ_2/λ_N of a Laplacian matrix is an important factor [9, 105]. A larger eigenratio corresponds to better synchronizability of the underlying communication graph. Intuitively, better network synchronizability will allow a more unstable A to achieve consensusability of the multi-agent systems and vice versa, which are confirmed by our result.*

For a continuous-time system under a sufficiently small sampling period, the unstable eigenvalues of the discretized system (8.1) can be made arbitrarily close to one and thus the inequality (8.5) will be eventually satisfied for any connected undirected graph. Thus, for the case of continuous-time agent dynamics, our result is consistent with that in [71]. In fact, for a continuous-time system, information can be transmitted arbitrarily fast so that the network synchronizability of the communication graph becomes less important for achieving consensusability.

4. *If the adjacency matrix $\mathcal{A}$ of the graph $\mathcal{G}$ is selected as a symmetric $(0,1)$ -matrix, the eigenratio $\lambda_2/\lambda_N \rightarrow 1$ means that the communication graph tends to be complete (cf. Lemma 8.1). In this case, the controller can be designed almost in a centralized fashion. Then, consensusability can be achieved for any unstable system.*
5. *The convergence rate of the average consensus over an undirected graph is determined by λ_2 [86, 88]. By the Courant-Weyl interlacing inequalities [48], adding an undirected edge to an undirected incomplete graph $\mathcal{G}$ will never decrease λ_2 , suggesting that the consensus performance will not deteriorate. However, adding an undirected edge to a graph may lead to a smaller eigenra-*

tio. For example, consider the following two graph Laplacian matrices:

$$\mathcal{L}_{\mathcal{G}_1} = \begin{bmatrix} 3 & -1 & 0 & 0 & -1 & -1 \\ -1 & 3 & -1 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & 0 & -1 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 0 & -1 & 3 \end{bmatrix}, \mathcal{L}_{\mathcal{G}_2} = \begin{bmatrix} 4 & -1 & -1 & 0 & -1 & -1 \\ -1 & 3 & -1 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & 0 & -1 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 0 & -1 & 3 \end{bmatrix}.$$

It is clear that $\mathcal{G}_2$ with an eigenratio 0.3970 is formed by adding an undirected edge to $\mathcal{G}_1$, whose eigenratio is 0.4. Thus, it is possible to lose consensusability of the multi-agent systems (8.1) under protocol (8.2) by adding an edge. It appears to be counter-intuitive since the communication graph with a “better” connectivity may result in a worse consensusable capability. Note that whether the eigenratio will increase or decrease by adding an edge is not conclusive, see [27] for more details.

6. The importance of the intrinsic entropy rate of a linear dynamical system, quantified by $\sum_j \log_2 |\lambda_j^u(A)|$, has been widely recognized in networked control systems, e.g., [36, 83, 100, 123, 124, 128], as it determines the minimum data rate for stabilization. Here the intrinsic entropy rate of the agent dynamics is firstly shown to pose a fundamental limitation on the eigenratio of an undirected graph for consensusability.

The proof of Theorem 8.1 depends on the following lemma, which gives a necessary and sufficient condition for a discrete-time simultaneous stabilization problem.

Lemma 8.3. *Given $0 < \lambda_2 \leq \dots \leq \lambda_N$ and under Assumption 8.1, a necessary and sufficient condition for the existence of a common control gain $K \in \mathbb{R}^{1 \times n}$ such that $\rho(A - \lambda_j BK) < 1$ for all $j \in \{2, \dots, N\}$ is that*

- (a) (A, B) is controllable;

(b) *The product of unstable eigenvalues of A is strictly upper bounded as follows:*

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}. \quad (8.7)$$

Proof. By the convention $\frac{2}{0} = \infty$, it is obvious that only the case with $\lambda_2/\lambda_N \neq 1$ needs to be elaborated.

Necessity: Under Assumption 8.1, it is straightforward that (A, B) is controllable. Without loss of generality (w.l.o.g.), assume that (A, B) is already in the controllable canonical form:

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots \\ \vdots & & \ddots & \ddots \\ 0 & \dots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \dots & -\alpha_{n-1} \end{bmatrix}; \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad (8.8)$$

where $\alpha_0 = \prod_j \lambda_j^u(A)$. Let $K = [-k_0, -k_1, \dots, -k_{n-1}]$ simultaneously stabilize $(A, \lambda_j B)$, it is obvious that

$$\det(zI_n - A + \lambda_j BK) = z^n + (\alpha_{n-1} - \lambda_j k_{n-1})z^{n-1} + \dots + (\alpha_0 - \lambda_j k_0). \quad (8.9)$$

Since all the eigenvalues of $A - \lambda_j BK$ are within the unit disk, it follows from (8.9) that for all $j \in \{2, \dots, N\}$,

$$|\alpha_0 - \lambda_j k_0| < 1 \Rightarrow \frac{|\alpha_0| - 1}{\lambda_j} < |k_0| < \frac{|\alpha_0| + 1}{\lambda_j}. \quad (8.10)$$

Thus, we obtain that $\bigcap_{j=2}^N \left(\frac{|\alpha_0| - 1}{\lambda_j}, \frac{|\alpha_0| + 1}{\lambda_j} \right) \neq \emptyset$, which further implies that $\frac{|\alpha_0| - 1}{\lambda_2} < \frac{|\alpha_0| + 1}{\lambda_N}$. Noting that $|\alpha_0| = \prod_j |\lambda_j^u(A)|$, the necessity follows directly.

Sufficiency: Select a ζ such that $\prod_j |\lambda_j^u(A)| < \zeta^{-1} \leq \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}$ and let $\zeta_j = 1 - \frac{2\lambda_j}{\lambda_2 + \lambda_N} \leq \zeta$ for all $j \in \{2, \dots, N\}$. Since (A, B) is controllable, there exists a positive definite solution P to the modified algebraic Riccati inequality (8.6) [36]. Let the control gain be $K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B}$, it follows that $(A - \lambda_j BK)^T P (A - \lambda_j BK) - P = A^T P A - (1 - \zeta_j^2) \frac{A^T P B B^T P A}{B^T P B} - P \leq A^T P A - (1 - \zeta^2) \frac{A^T P B B^T P A}{B^T P B} - P < 0$, which

completes the proof. ■

Proof of Theorem 8.1

Denote the average state of all agents by $\bar{x}(k) \triangleq \frac{1}{N} \sum_{i=1}^N x_i(k) = \frac{1}{N}(\mathbf{1}^T \otimes I_n)x(k)$ and the deviation of each agent from the average state by $\delta_i(k) \triangleq x_i(k) - \bar{x}(k)$. By consensusability definition, it yields that $\lim_{k \rightarrow \infty} \|\delta_i(k)\| \leq \frac{1}{N} \sum_{j=1}^N \lim_{k \rightarrow \infty} \|x_i(k) - x_j(k)\| = 0$. Conversely, $\lim_{k \rightarrow \infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{N}$ immediately implies consensusability of the multi-agent systems (8.1). Thus, consensusability is equivalent to that $\lim_{k \rightarrow \infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{N}$.

Stack x_j to get a new state vector $x(k) = [x_1^T(k), \dots, x_N^T(k)]^T$. By (8.2), the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)x(k). \quad (8.11)$$

Noting that $\mathbf{1}^T \mathcal{L}_G = \mathbf{0}^T$, the following equalities are in force:

$$\bar{x}(k+1) = \frac{1}{N}(\mathbf{1}^T \otimes A)x(k) - \frac{1}{N}(\mathbf{1}^T \mathcal{L}_G \otimes BK)x(k) = A\bar{x}(k).$$

Let $\delta(k) = [\delta_1^T(k), \dots, \delta_N^T(k)]^T$, subtracting (8.11) from (8.12) leads to that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)\delta(k). \quad (8.12)$$

Select $\phi_i \in \mathbb{R}^N$ such that $\phi_i^T \mathcal{L}_G = \lambda_i \phi_i^T$ and form $\Phi = [\frac{1}{\sqrt{N}}, \phi_2, \dots, \phi_N]$ to transform $\mathcal{L}_G$ into a diagonal form: $\text{diag}(0, \lambda_2, \dots, \lambda_N) = \Phi^T \mathcal{L}_G \Phi$. Using the property of Kronecker product gives that

$$(\Phi \otimes I_n)^T (I_N \otimes A - \mathcal{L}_G \otimes BK) (\Phi \otimes I_n) = \text{diag}(A, A - \lambda_2 BK, \dots, A - \lambda_N BK). \quad (8.13)$$

Denote $\tilde{\delta}(k) = (\Phi \otimes I_n)^T \delta(k)$ and partition $\tilde{\delta}(k) \in \mathbb{R}^{nN}$ into two parts, i.e. $\tilde{\delta}(k) = [\tilde{\delta}_1^T(k), \tilde{\delta}_2^T(k)]^T$, where $\tilde{\delta}_1(k) \in \mathbb{R}^n$ is a vector consisting of the first n elements of $\tilde{\delta}(k)$. Then, $\tilde{\delta}_1(k) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \delta_i(k) = \mathbf{0}$. In view of (8.12) and (8.13), it yields that

$$\tilde{\delta}_2(k+1) = \text{diag}(A - \lambda_2 BK, \dots, A - \lambda_N BK) \tilde{\delta}_2(k). \quad (8.14)$$

Since $\Phi \otimes I_n$ is nonsingular, it follows that $\|\delta(k)\| = \|\tilde{\delta}_2(k)\|$ for all $k \in \mathbb{N}$.

Necessity: By (8.14), it follows that $\rho(A - \lambda_i BK) < 1, \forall i \in \{2, \dots, N\}$, which in turn, implies that $\lambda_2 > 0$. Otherwise, $\lambda_2 = 0$, then $\rho(A - \lambda_2 BK) \geq 1, \forall K \in \mathbb{R}^{1 \times n}$ by Assumption 8.1. That is, $(A, \lambda_j B)$ can be simultaneously stabilized by a common control gain $K \in \mathbb{R}^{1 \times n}$. In light of Lemma 8.3, the necessity is established.

Sufficiency: Under Assumption 8.1, the inequality (8.5) implies that $\lambda_2 > 0$. By Lemma 8.3, the common gain $K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B}$, where P is a positive definite solution to the modified algebraic Riccati inequality (8.6), can simultaneously stabilize $(A, \lambda_i B), i \in \{2, \dots, N\}$, i.e., $\rho(A - \lambda_i BK) < 1$. Together with (8.14), the proof of sufficiency is completed. ■

Remark 8.2. *In the proof of sufficiency, we have constructed a specific control gain K for the multi-agent systems to reach a consensus, which is obtained by solving a modified algebraic Riccati inequality. It is interesting to note that the discrete-time consensusability problem over an undirected graph is closely related to the robust stabilization problem with a bounded uncertainty in the input gain. Here the variation of the eigenvalues of the Laplacian matrix is interpreted as parameter uncertainty in the input. Different from the classical robust stabilization problem, the uncertainty in this case takes only a finite number of specific positive values. What is particularly surprising is that the necessary and sufficient condition for the classical robust stabilization continues to hold.*

8.2.2 Performance Analysis of Consensus Protocols

Under protocol (8.2), we can evaluate the corresponding *asymptotic convergence factor* [118] defined by $r_{asym} = \sup_{\delta(0) \neq \mathbf{0}} \lim_{k \rightarrow \infty} \left(\frac{\|\delta(k)\|}{\|\delta(0)\|} \right)^{\frac{1}{k}}$. Similarly, another measure of the speed of convergence is the *per-step convergence factor* [118], which is defined as $r_{step} = \sup_{\delta(k) \neq \mathbf{0}} \frac{\|\delta(k+1)\|}{\|\delta(k)\|}$. Denote stabilizing control gains of $(A, \lambda_j B)$ by

$$\Gamma_j = \{K \in \mathbb{R}^{1 \times n} \mid \rho(A - \lambda_j BK) < 1\}. \quad (8.15)$$

Given any control gain $K \in \mathbb{R}^{1 \times n}$, define

$$J(K) = \text{diag}(A - \lambda_2 BK, \dots, A - \lambda_N BK). \quad (8.16)$$

The convergence speed is characterized in the following result:

Corollary 8.1. *Given an undirected graph $\mathcal{G}$ and under the conditions in Theorem 8.1, select a $K \in \bigcap_{j=2}^N \Gamma_j$, where Γ_j is defined in (8.15). Then, asymptotic convergence factor and per-step convergence factor for consensus are respectively evaluated by*

$$r_{\text{asym}} = \rho(J(K)) \text{ and } r_{\text{step}} = \|J(K)\|. \quad (8.17)$$

Proof. Since $\Phi \otimes I_n$ is an unitary matrix, it is obvious that $\|\delta(k)\| = \|\tilde{\delta}(k)\|, \forall k \in \mathbb{N}$. Since $\tilde{\delta}_1(k) = \mathbf{0}$, then $\|\tilde{\delta}(k)\| = \|\tilde{\delta}_2(k)\|, \forall k \in \mathbb{N}$. In view of (8.14), it follows that

$$r_{\text{asym}} = \sup_{\tilde{\delta}_2(0) \neq \mathbf{0}} \lim_{k \rightarrow \infty} \left(\frac{\|J^k(K) \tilde{\delta}_2(0)\|}{\|\tilde{\delta}_2(0)\|} \right)^{\frac{1}{k}} \leq \lim_{k \rightarrow \infty} \|J^k(K)\|^{\frac{1}{k}} = \rho(J(K)). \quad (8.18)$$

On the other hand, select $\tilde{\delta}_2(0)$ as an eigenvector of $J(K)$ corresponding to the largest eigenvalue, i.e. $J(K) \tilde{\delta}_2(0) = \lambda_M \tilde{\delta}_2(0)$ and $|\lambda_M| = \rho(J(K))$. Then, we have the following results:

$$r_{\text{asym}} \geq \lim_{k \rightarrow \infty} \left(\frac{\|J^k(K) \tilde{\delta}_2(0)\|}{\|\tilde{\delta}_2(0)\|} \right)^{\frac{1}{k}} = \lim_{k \rightarrow \infty} \left(\frac{\|\lambda_M^k \tilde{\delta}_2(0)\|}{\|\tilde{\delta}_2(0)\|} \right)^{\frac{1}{k}} = \rho(J(K)). \quad (8.19)$$

Hence, it yields that $r_{\text{asym}} = \rho(J(K))$. The second part can be shown similarly. ■

To make the convergence to consensus as fast as possible, a control gain $K \in \bigcap_{j=2}^N \Gamma_j$ should be selected to minimize the *asymptotic convergence factor* or *per-step convergence factor*. Since the spectral radius of a square matrix is not a convex function, not even Lipschitz continuous, the problem of finding an optimal K to minimize the *asymptotic convergence factor* is in general very difficult. However, a lower bound for the optimal *asymptotic convergence factor* can be derived.

Theorem 8.2. *Given an undirected graph $\mathcal{G}$ and under the conditions in Theorem 8.1, the optimal asymptotic convergence factor is lower bounded as follows:*

$$r_{asy}^* = \inf_{K \in \bigcap_{j=2}^N \Gamma_j} \rho(J(K)) \geq |\det(A)|^{1/n} \left(\frac{1 - \lambda_2/\lambda_N}{1 + \lambda_2/\lambda_N} \right)^{1/n}. \quad (8.20)$$

Proof. It essentially follows from the necessity of Lemma 8.3 with some modifications. Without loss of generality, let (A, B) be given in the controllable canonical form. For any control gain $K \in \bigcap_{j=2}^N \Gamma_j$, it follows from the definition of $J(K)$ in (8.16) that $\rho(A - \lambda_j BK) \leq \rho(J(K)) < 1, \forall j \in \{2, \dots, N\}$. In consideration of (8.9) and letting $K = [-k_0, \dots, -k_{n-1}]$, we obtain that

$$|\det(A) - \lambda_j k_0| \leq \rho(J(K))^n \Rightarrow \frac{|\det(A)| - \rho(J(K))^n}{\lambda_j} \leq |k_0| \leq \frac{|\det(A)| + \rho(J(K))^n}{\lambda_j}.$$

By using the arguments in the necessity of Lemma 8.3, it can be shown that

$$\frac{|\det(A)| - \rho(J(K))^n}{\lambda_2} \leq \frac{|\det(A)| + \rho(J(K))^n}{\lambda_N}.$$

With some simple algebraic manipulations, we obtain that

$$\rho(J(K)) \geq |\det(A)|^{1/n} \left(\frac{1 - \lambda_2/\lambda_N}{1 + \lambda_2/\lambda_N} \right)^{1/n}, \forall K \in \bigcap_{j=2}^N \Gamma_j.$$

Taking infinitum on both sides of the above inequality completes the proof. $\blacksquare$

Remark 8.3. *The lower bound of (8.20) is attainable for some special cases:*

1. *The interaction graph is complete and the adjacency matrix $\mathcal{A}$ is a symmetric $(0,1)$ -matrix. Lemma 8.2 implies that $\lambda_2 = \lambda_N > 0$. Due to that $(A, \lambda_2 B)$ is controllable, there exists a control gain K^* such that $\rho(A - \lambda_2 B K^*) = 0$. Thus, $r_{asy} = \rho(J(K^*)) = 0$ and the consensus can be achieved in finite time.*
2. *The agent dynamics is an unstable scalar system and the communication graph is undirected, i.e., $x_i(k+1) = ax_i(k) + bu_i(k)$, where $a \geq 1$ and $b \neq 0$. Let $k^* = \frac{2a}{b(\lambda_2 + \lambda_N)}$, then $a - \lambda_2 b k^* = \frac{a(1 - \lambda_2/\lambda_N)}{1 + \lambda_2/\lambda_N}$ and $|a - \lambda_j b k^*| \leq \frac{a(1 - \lambda_2/\lambda_N)}{1 + \lambda_2/\lambda_N}, \forall j \in \{2, \dots, N\}$. Thus, $r_{asy}^* = \rho(J(k^*)) = \frac{a(1 - \lambda_2/\lambda_N)}{1 + \lambda_2/\lambda_N}$.*

3. The lower bound of (8.20) can also be approached for the discrete-time second-order consensus with an undirected graph, see Section 8.3 for the design of a control gain to reach this bound.

On the other hand, $K \in \Gamma_j$ is equivalent to that there exists a positive definite P_j such that $(A - \lambda_j BK)^T P_j (A - \lambda_j BK) - P_j < 0$. The optimal control gain K to minimize the *per-step convergence factor* is thus obtained by solving the following optimization:

$$\text{minimize } r \tag{8.21}$$

$$\begin{aligned} &\text{subject to: } P_j > 0, Q_j > 0, P_j Q_j = I_n, \forall j \in \{2, \dots, N\} \\ &\begin{bmatrix} -rI_n & (A - \lambda_j BK)^T \\ A - \lambda_j BK & -I_n \end{bmatrix} < 0, \begin{bmatrix} -P_j & (A - \lambda_j BK)^T \\ A - \lambda_j BK & -Q_j \end{bmatrix} < 0. \end{aligned} \tag{8.22}$$

Unfortunately, the constraint in the above optimization is non-convex, leaving an open question for finding an optimal control gain although the cone-complementary approach [13] can be applied to obtain a suboptimal solution.

8.2.3 Application to Formationability on Graphs

As an important application, the result on consensusability is extended to study formationability of the discrete-time multi-agent systems (8.1). Specifically, given a formation vector $H = [H_1^T, H_2^T, \dots, H_N^T]^T$, the following control protocol is adopted to study the formation problem of the discrete-time multi-agent systems (8.1):

$$u_i(k) = K \sum_{j=1}^N a_{ij} [(x_j(k) - H_j) - (x_i(k) - H_i)], \tag{8.23}$$

where $H_i - H_j$ is the desired distance vector between agent i and agent j . In the context of formation control, the protocol (8.23) has been widely adopted for continuous-time systems [33, 59, 70]. As in these works, the common knowledge of the directions of reference axes is required for all the agents.

Definition 8.2. The discrete-time multi-agent systems (8.1) with a fixed graph $\mathcal{G}$ are said to be formationable under protocol (8.23) if for any finite $x_i(0), \forall i \in \mathcal{N}$,

there exists a control gain $K \in \mathbb{R}^{1 \times n}$ in (8.23) such that $\lim_{k \rightarrow \infty} \|(x_i(k) - H_i) - (x_j(k) - H_j)\| = 0, \forall i, j \in \mathcal{N}$.

Based on Theorem 8.1, a necessary and sufficient condition on formationability of the discrete-time multi-agent systems is stated as follows.

Theorem 8.3. *Given a set of desired formation vectors $H_i, i \in \mathcal{N}$ and an undirected graph $\mathcal{G}$. Assume that Assumption 8.1 holds, the discrete-time multi-agent systems (8.1) are formationable under protocol (8.23) if and only if the following conditions hold.*

- (a) (A, B) is a controllable pair and $A(H_i - H_j) = H_i - H_j, \forall i, j \in \mathcal{N}$;
- (b) Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the following strict inequality:

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}. \quad (8.24)$$

Proof. Denote the average formation vector $\bar{H} \triangleq \frac{1}{N} \sum_{i=1}^N H_i$ and $\delta_i(k) \triangleq x_i(k) - H_i - (\bar{x}(k) - \bar{H})$. Similarly, it is easy to verify that formationability is equivalent to that $\lim_{k \rightarrow \infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{N}$. The following dynamical equation can be easily derived:

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k) + (I_N \otimes (A - I_n)) \begin{bmatrix} H_1 - \bar{H} \\ \vdots \\ H_N - \bar{H} \end{bmatrix}. \quad (8.25)$$

Thus, to reach the desired formation, we have that $(A - I_n)(H_i - \bar{H}) = 0, \forall i \in \mathcal{N}$, which implies that $A(H_j - H_i) = H_j - H_i, \forall i, j \in \mathcal{N}$. The rest follows from the proof of the necessity of Theorem 8.1.

Conversely, using the condition that $A(H_j - H_i) = H_j - H_i, \forall i, j \in \mathcal{N}$, (8.25) is reduced to the following form: $\delta(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BK)\delta(k)$. Again, the remainder of the proof follows from the sufficiency proof of Theorem 8.1. $\blacksquare$

Remark 8.4. *For the continuous-time case, the formation condition is modified as $A(H_i - H_j) = \mathbf{0}, \forall i, j \in \mathcal{N}$ [59, 71]. The physical meaning of the constraint $A(H_i - H_j) = H_i - H_j, \forall i, j \in \mathcal{N}$ will become clear for the second-order consensus problem in Section 8.3. For example, to maintain a fixed formation, the velocities of all agents should be the same.*

8.2.4 Extensions to Multiple Inputs and Digraphs

Similarly, one can easily show that the consensusability problem over *directed* graphs is still equivalent to a simultaneous stabilization problem, i.e., finding a common gain such that $\rho(A - \lambda_j BK) < 1$ for all $j \in \{2, \dots, N\}$. The main difference is that under *directed* graphs, the nonzero eigenvalues of the induced Laplacian matrix, $\lambda_j, j \in \{2, \dots, N\}$, are not real numbers in general.

We first study the solution to a discrete-time modified Riccati inequality, which plays an important role in the design of control gain of (8.2). By jointly taking into account the agent dynamics and communication graphs, a sufficient condition is provided for the multi-agent systems to reach a consensus under protocol (8.2). This condition is also necessary for single input agent dynamics with all the open-loop poles lying on or outside of the unit circle.

Assume that (A_u, B_u) is controllable, applying Wonham decomposition to the system (A_u, B_u) converts the multiple input system to l single input subsystems [20]. Specifically, there is a nonsingular real matrix Q with a compatible dimension such that $\tilde{A} = Q^{-1}A_uQ$ and $\tilde{B} = Q^{-1}B_u$ take the form:

$$\tilde{A} = \begin{bmatrix} A_1 & * & \dots & * \\ 0 & A_2 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_l \end{bmatrix}, \tilde{B} = \begin{bmatrix} b_1 & * & \dots & * \\ 0 & b_2 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_l \end{bmatrix}, \quad (8.26)$$

where ‘*’ denotes possibly nonzero parts and (A_j, b_j) with $A_j \in \mathbb{R}^{n_j \times n_j}$ and $b_j \in \mathbb{R}^{n_j}$ for all $j \in \{1, \dots, l\}$ is controllable and $\sum_{j=1}^l n_j = n_u$. With the aid of Wonham decomposition, an interesting result on the solution to a modified Riccati inequality

will be developed.

Lemma 8.4. *Assume that $M_1 \in \mathbb{R}^{n_1 \times n_1}$ and $M_2 \in \mathbb{R}^{n_2 \times n_2}$ are two positive definite matrices. Given any $M_{12} \in \mathbb{R}^{n_1 \times n_2}$ and let*

$$\beta > \frac{\bar{\sigma}(M_{12}^T M_1^{-1} M_{12})}{\underline{\sigma}(M_2)}, \quad (8.27)$$

then $M \triangleq \begin{bmatrix} M_1 & M_{12} \\ M_{12}^T & \beta M_2 \end{bmatrix}$ is a positive definite matrix.

Proof: By Shur complement, $M > 0$ if and only if $M_1 > 0$ and $\beta M_2 - M_{12}^T M_1^{-1} M_{12} > 0$, which is obvious by (8.27). ■

Lemma 8.5. *Given a positive $\delta \in \mathbb{R}$, consider the modified Riccati inequality (8.6). Assume that A is unstable and (A, B) is stabilizable, then there is a critical value $\delta_c \in (0, 1)$ such that for any positive $\delta < \delta_c$, there always exists a positive definite solution P to (8.6). Moreover, if B is invertible, $\delta_c = 1/\max_i |\lambda_i^u(A)|$. If B is of rank one, $\delta_c = 1/(\prod_i |\lambda_i^u(A)|)$. Otherwise, δ_c is lower bounded by*

$$\delta_c \geq 1/\prod_i |\lambda_i^u(A_{l^*})| \triangleq \delta'_c, \quad (8.28)$$

where l^ is defined by $l^* = \arg \max_{j \in \{1, \dots, l\}} (\prod_i |\lambda_i^u(A_j)|)$ and A_j is given in (8.26).*

Proof: In view of [36, 100], only the inequality (8.28) needs to be elaborated. Since (A, B) is stabilizable, then (A_u, B_u) is controllable. By Wonham decomposition, it follows that (A_j, b_j) is controllable. For brevity and without loss of generality, we assume that $l = 2$. By the result on the case that B is of rank one, it implies that given any positive $\delta < \delta'_c$, there exists a $P_j > 0$ such that

$$\begin{aligned} & P_j - A_j^T P_j A_j + (1 - \delta^2) A_j^T P_j b_j (b_j^T P_j b_j)^{-1} b_j^T P_j A_j \\ &= P_j - \delta^2 A_j^T P_j A_j - (1 - \delta^2) (A_j - b_j F_j)^T P_j (A_j - b_j F_j) > 0, \end{aligned} \quad (8.29)$$

where the control gain $F_j = (b_j^T P_j b_j)^{-1} b_j^T P_j A_j$ for all $j \in \{1, 2\}$. Note that given any positive β , then $\tilde{P}_j = \beta P_j > 0$ solves the modified Riccati inequality (8.29) as

well. Now, let $\tilde{P} = \text{diag}(P_1, \beta P_2)$ and $\tilde{F} = \text{diag}(F_1, F_2)$. It follows that

$$\tilde{P} - \delta^2 \tilde{A}^T \tilde{P} \tilde{A} - (1 - \delta^2)(\tilde{A} - \tilde{B}\tilde{F})^T \tilde{P}(\tilde{A} - \tilde{B}\tilde{F}) \triangleq \begin{bmatrix} \tilde{M}_1 & \tilde{M}_{12} \\ \tilde{M}_{12}^T & \beta \tilde{M}_2 \end{bmatrix} \triangleq \tilde{M}, \quad (8.30)$$

where $\tilde{M}_j = P_j - \delta^2 A_j^T P_j A_j - (1 - \delta^2)(A_j - b_j F_j)^T P_j (A_j - b_j F_j) > 0$ for all $j \in \{1, 2\}$ and $\tilde{M}_{12}$ is independent of β . By Lemma 8.4, we can find a sufficiently large β such that $\tilde{M} > 0$. Let $P_u = Q^{-T} \tilde{P} Q^{-1}$ and $F_u = \tilde{F} Q^{-1}$, it follows from (8.30) that $P_u - \delta^2 A_u^T P_u A_u - (1 - \delta^2)(A_u - B_u F_u)^T P_u (A_u - B_u F_u) > 0$. Note that A_s is stable, there exists a positive definite P_s such that $P_s - A_s^T P_s A_s > 0$. Let $F = \text{diag}(0, F_u)$ and $P = \text{diag}(P_s, \beta_1 P_u)$, we can similarly establish that there exists a sufficiently large β_1 such that $P - \delta^2 A^T P A - (1 - \delta^2)(A - BF)^T P(A - BF) > 0$. Jointly with Lemma 1 of [103], it follows that $P > A^T P A - (1 - \delta^2) A^T P B (B^T P B)^{-1} B^T P A$. Thus, given any $\delta < \delta'_c$, we have found a $P > 0$ to solve the modified Riccati inequality (8.6). This essentially implies that $\delta_c \geq \delta'_c$. ■

It is known that the above result is of great importance in the stability analysis of Kalman filtering with intermittent observations [100] and the quadratic stabilization of an uncertain linear system (cf. Theorem 2.1 of [36]). In fact, it is also useful for the design of a control gain in (8.2) to solve the consensusability problem, which is presented as follows.

Theorem 8.4. *Given a fixed directed communication graph $\mathcal{G}$, if (A, B) is stabilizable and there exists a $\omega \in \mathbb{R}$ such that*

$$\delta(\omega) \triangleq \max_{j \in \{2, \dots, N\}} |1 - \omega \lambda_j| < \delta_c, \quad (8.31)$$

where δ_c is obtained in Lemma 8.5. Then, we have the following results.

- (a) Insert $\delta = \delta(\omega)$ to (8.6), there is a positive definite solution, denoted by $P = P(\delta(\omega))$, to the modified Riccati inequality (8.6).
- (b) The control gain $K = \omega(B^T P B)^{-1} B^T P A$ solves the consensus problem of discrete-time multi-agent systems (8.1) under protocol (8.2).

Proof. Part (a) follows from Lemma 8.5. Only part (b) needs to be proved. Clearly, it suffices to verify that $\rho(A - \lambda_j BK) < 1$ for all $j \in \{2, \dots, N\}$. To this purpose, let $\delta_j = 1 - \omega\lambda_j$. Then, it is clear that there exists a $\delta > 0$ such that $|\delta_j| < \delta < \delta_c$ for all $j \in \{2, \dots, N\}$ by (8.31). By part (a), we obtain that $P - (A - \lambda_j BK)^H P (A - \lambda_j BK) = P - A^T P A + (1 - |\delta_j|^2) A^T P B (B^T P B)^{-1} B^T P A \geq P - A^T P A + (1 - \delta^2) A^T P B (B^T P B)^{-1} B^T P A > 0$. ■

The sufficient condition is also necessary if the agent dynamics satisfies Assumption 8.1.

Theorem 8.5. *Given a fixed directed graph $\mathcal{G}$ and under Assumption 8.1, then a necessary condition for consensusability of multi-agent systems (8.1) under protocol (8.2) is that (A, B) is controllable and*

$$\min_{\omega \in \mathbb{R}} \max_{j \in \{2, \dots, N\}} |1 - \omega\lambda_j| < \frac{1}{\prod_i |\lambda_i^u(A)|}. \quad (8.32)$$

Proof. It is straightforward that (A, B) is controllable. Similarly, it follows from (8.9) that

$$|\alpha_0 - \lambda_j k_0| < 1 \Rightarrow |1 - \lambda_j k'_0| < \frac{1}{|\alpha_0|}, j \in \{2, \dots, N\},$$

where $k'_0 = k_0/|\alpha_0|$. Finally, it is easy to derive that $\inf_{\omega \in \mathbb{R}} \max_{j \in \{2, \dots, N\}} |1 - \omega\lambda_j| < \frac{1}{\prod_i |\lambda_i^u(A)|}$. Note that $|x|$ is continuous with respect to (w.r.t.) $x \in \mathbb{C}$, this implies that inf in the above inequality is achievable. Thus, the proof is completed. ■

Remark 8.5. *By Theorem 8.4, we know that if $\text{rank}(B) = 1$, the critical value $\delta_c = 1/(\prod_i |\lambda_i^u(A)|)$. Under this case and Assumption 8.1, it is clear that the multi-agent systems are consensusable under protocol (8.2) if and only if (A, B) is controllable and condition (8.32) holds.*

Remark 8.6. *If the communication graph $\mathcal{G}$ is an undirected graph, all the eigenvalues of $\mathcal{L}_{\mathcal{G}}$ are nonnegative real numbers. A closed-form solution to the min-max optimization (8.32) is obtained as follows:*

$$\min_{\omega \in \mathbb{R}} \max_{j \in \{2, \dots, N\}} |1 - \omega\lambda_j| = \frac{\lambda_N - \lambda_2}{\lambda_N + \lambda_2}. \quad (8.33)$$

Then, condition (8.32) is converted to that $\prod_i |\lambda_i^u(A)| < \frac{1+\lambda_2/\lambda_N}{1-\lambda_2/\lambda_N}$, which recovers the result in Theorem 8.1.

Even under a *directed* graph, the condition of (8.32) can be readily checked by the following lemma.

Lemma 8.6. *Let $\lambda_j = r_j \exp(\theta_j \mathbf{i})$ with $\mathbf{i}^2 = -1$ and $\Delta_m = 1/(\prod_i |\lambda_i^u(A)|)$, then the inequality (8.32) holds if and only if the intersection*

$$\bigcap_{j=2}^N \left(\frac{\cos \theta_j - \sqrt{\Delta_m^2 - \sin^2 \theta_j}}{r_j}, \frac{\cos \theta_j + \sqrt{\Delta_m^2 - \sin^2 \theta_j}}{r_j} \right) \quad (8.34)$$

is not empty, which is equivalent to that

$$\frac{1 - \Delta_m^2}{\min_{j \in \{2, \dots, N\}} r_j f(\theta_j)} < \min_{j \in \{2, \dots, N\}} \frac{f(\theta_j)}{r_j}. \quad (8.35)$$

Here $f(\theta) = \cos \theta + \sqrt{\Delta_m^2 - \sin^2 \theta}$ is a decreasing function w.r.t. $\theta \in (0, \arcsin(\Delta_m))$, where $\arcsin(x)$ is the inverse sine of x .

Proof: One can easily verify that the inequality (8.32) holds if and only if the set $\bigcap_{j=2}^N \{\omega \in \mathbb{R} \mid |1 - \omega \lambda_j| < \Delta_m\} \neq \emptyset$, which is equivalent to (8.34). The equivalence between (8.34) and (8.35) is trivial as well. ■

Remark 8.7. *By Lemma 3.3 of [93], all the nonzero eigenvalues of $\mathcal{L}_G$ lie in the open right half plane, which implies that $-\pi/2 \leq \theta_j \leq \pi/2$. Then, it follows from (8.32) that*

$$|\theta_j| < \arcsin\left(\frac{1}{\prod_i |\lambda_i^u(A)|}\right), \forall j \in \{2, \dots, N\}. \quad (8.36)$$

This means that the more unstable of the open loop matrix, a stronger condition is required on the directed graph to achieve consensusability under protocol (8.2). Since $|1 - \omega \lambda_j| \geq |1 - |\omega| \cdot |\lambda_j||$ for all $\omega \in \mathbb{R}$ and $j \in \{2, \dots, N\}$, it follows that $\min_{\omega \in \mathbb{R}} \max_{j \in \{2, \dots, N\}} |1 - |\omega| \cdot |\lambda_j|| \leq \min_{\omega \in \mathbb{R}} \max_{j \in \{2, \dots, N\}} |1 - \omega \lambda_j|$. Thus, we can easily derive from (8.32) that

$$\prod_i |\lambda_i^u(A)| < \frac{|\lambda_N| + |\lambda_2|}{|\lambda_N| - |\lambda_2|}. \quad (8.37)$$

In view of Theorem 8.1, we know that under an undirected graph, the necessary condition (8.37) on graphs is also sufficient for reaching a consensus. Unfortunately, this condition is no longer strong enough for achieving consensusability of the multi-agent systems under protocol (8.2) if the communication graph is directed.

8.3 Special Case: Second-order Consensus

Recently, more attention has been paid to consensus related problems for agents with double-integrator dynamics [65, 92]. In the section, an optimal control gain is designed to achieve the optimal *asymptotic convergence factor*, which is expressed by the *eigenratio* of an *undirected* graph.

8.3.1 Second-order Consensus

Consider the sampled double-integrator systems with a sampling period $h > 0$ for each agent as follows [92]:

$$x_i(k+1) = x_i(k) + hv_i(k) + \frac{1}{2}h^2u_i(k), \quad (8.38)$$

$$v_i(k+1) = v_i(k) + hu_i(k), \forall i \in \mathcal{N}, \quad (8.39)$$

where $x_i(k) \in \mathbb{R}$ and $v_i(k) \in \mathbb{R}$ respectively correspond to the position and velocity of agent i at time kh . $u_i(k) \in \mathbb{R}$ is the control input. Denote the configuration variable of agent i at time kh by $\xi_i(k) = [x_i(k), v_i(k)]^T$, the agent dynamics is written in a vector form as follows:

$$\xi_i(k+1) = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \xi_i(k) + \begin{bmatrix} \frac{1}{2}h^2 \\ h \end{bmatrix} u_i(k), \forall i \in \mathcal{N}. \quad (8.40)$$

The following control protocol is adopted

$$u_i(k) = K \sum_{j=1}^N a_{ij}(\xi_j(k) - \xi_i(k)), \quad K \in \mathbb{R}^{1 \times 2}. \quad (8.41)$$

In view of Lemma 8.2 and Theorem 8.1, a necessary and sufficient condition

for reaching second-order consensus is given below.

Corollary 8.2. *Given an undirected graph $\mathcal{G}$, the second-order multi-agent systems (8.40) are consensusable under protocol (8.41) if and only if $\mathcal{G}$ is connected.*

In this occasion, a control gain K that solves the second-order consensus can be designed without resorting to the algebraic Riccati inequality (8.6).

Theorem 8.6. *Given an undirected graph $\mathcal{G}$, let $\Omega = \{[\alpha, \beta] | \beta < \frac{2}{\lambda_N h}, 0 < \alpha < \frac{2\beta}{h}\}$. Under the conditions in Corollary 8.2, a control gain K in (8.41) solves the second-order consensus problem if and only if $K \in \Omega$.*

Proof. By (8.14) and the definition of $J(K)$ in (8.16), a control gain $K \triangleq [\alpha, \beta]$ solves the second-order consensus problem under protocol (8.41) if and only if $\rho(J(K)) < 1$. It is easy to compute that

$$\begin{aligned} \det(zI_2 - (A - \lambda BK)) &= z^2 + \left(\frac{1}{2}\alpha\lambda h^2 + \beta\lambda h - 2\right)z + \frac{1}{2}\alpha\lambda h^2 - \beta\lambda h + 1 \\ &\triangleq (z - z_1)(z - z_2). \end{aligned} \quad (8.42)$$

Applying a bilinear transformation on the above, i.e., $z = \frac{s+1}{s-1}$, we obtain the new polynomial expressed by: $\alpha\lambda h^2 s^2 + (2\beta\lambda h - \alpha\lambda h^2)s + 4 - 2\beta\lambda h \triangleq (\alpha\lambda h^2)(s - s_1)(s - s_2)$. From the property of the bilinear transformation, it follows that $|z_1| < 1$ and $|z_2| < 1$ if and only if $s_1 < 0$ and $s_2 < 0$, which are equivalent to that $\beta < \frac{2}{\lambda h}$ and $0 < \alpha < \frac{2\beta}{h}$ by the Routh's stability criterion. The rest of the proof is trivial. ■

The following result characterizes the optimal *asymptotic convergence factor* among all possible control gains that achieve a consensus under protocol (8.41).

Theorem 8.7. *Given an undirected graph $\mathcal{G}$ and under the conditions in Corollary 8.2, the optimal asymptotic convergence factor for consensus of the second-order multi-agent systems (8.40) under protocol (8.41) is $r_{asy}^* = \left(\frac{1 - \lambda_2/\lambda_N}{1 + \lambda_2/\lambda_N}\right)^{1/2}$ and $K^* = \left[\frac{1 - (r_{asy}^*)^2}{h^2\lambda_N}, \frac{3 + (r_{asy}^*)^2}{2h\lambda_N}\right]$ leads to the optimal asymptotic convergence factor.*

Proof. By Theorem 8.2, the optimal *asymptotic convergence factor* is lower bounded by

$$r_{asy}^* \geq \left(\frac{1 - \lambda_2/\lambda_N}{1 + \lambda_2/\lambda_N}\right)^{1/2}. \quad (8.43)$$

Next, a control gain $K = [\alpha, \beta]$ is to be constructed to show that the lower bound is tight. For notational simplicity, let $\alpha_0 = \frac{1}{2}\alpha h^2 + \beta h$, $\beta_0 = \frac{1}{2}\alpha h^2 - \beta h$ and $\sigma = \left(\frac{1-\lambda_2/\lambda_N}{1+\lambda_2/\lambda_N}\right)^{1/2}$. Then, we obtain that $\det(zI_2 - (A - \lambda BK)) = z^2 + (\alpha_0\lambda - 2)z + \beta_0\lambda + 1 \triangleq (z - z_+(\lambda))(z - z_-(\lambda))$. It is clear that $\forall \lambda \geq \frac{4(\alpha_0+\beta_0)}{\alpha_0^2}$, $z_+(\lambda)$ and $z_-(\lambda)$ are real numbers and can be expressed by

$$z_+(\lambda) = 1 - \frac{\alpha_0 + \beta_0}{\frac{\alpha_0}{2} + \sqrt{\frac{\alpha_0^2}{4} - \frac{\alpha_0 + \beta_0}{\lambda}}}, z_-(\lambda) = 1 - \frac{\alpha_0 + \beta_0}{\frac{\alpha_0}{2} - \sqrt{\frac{\alpha_0^2}{4} - \frac{\alpha_0 + \beta_0}{\lambda}}}. \quad (8.44)$$

Setting $z_+(\lambda_N) = \sigma$ and $z_-(\lambda_N) = -\sigma$, we get a solution of $\alpha_0^* = \frac{2}{\lambda_N}$ and $\beta_0^* = -\frac{\sigma^2+1}{\lambda_N}$. Note that $\frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} = (1-\sigma^2)\lambda_N \in [\lambda_2, \lambda_N)$, there exists a $k \in \{2, \dots, N-1\}$ such that $\frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} \leq \lambda_{k+1}$ and $\frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} \geq \lambda_k$. Thus, $z_+(\lambda_j)$ and $z_-(\lambda_j)$, $j \in \{k+1, \dots, N\}$ are real numbers. For any control gain K solving the second-order consensus, Theorem 8.6 assures that $\alpha_0 + \beta_0 = \alpha h^2 > 0$. Then, $z_+(\lambda)$ and $z_-(\lambda)$ are respectively increasing and decreasing functions w.r.t. $\lambda > \frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2}$. It implies that $z_+(\lambda_N) \geq \dots \geq z_+(\lambda_{k+1}) \geq z_-(\lambda_{k+1}) \geq \dots \geq z_-(\lambda_N)$ and

$$\max_{j \in \{k+1, \dots, N\}} \{|z_+(\lambda_j)|, |z_-(\lambda_j)|\} = \sigma. \quad (8.45)$$

On the other hand, for any $\lambda \leq \frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2}$, $z_+(\lambda)$ and $z_-(\lambda)$ are a pair of conjugate complex numbers and $|z_+(\lambda)|^2 = |z_-(\lambda)|^2 = 1 + \frac{1}{2}\lambda^2(\alpha_0^*)^2 - \lambda(\beta_0^* + 2\alpha_0^*) = 2\left(\frac{\lambda}{\lambda_N}\right)^2 + (\sigma^2 - 3)\left(\frac{\lambda}{\lambda_N}\right) + 1$. In particular, for all $j \in \{2, \dots, k\}$, $z_+(\lambda_j)$ and $z_-(\lambda_j)$ are complex numbers. Moreover, $|z_+((1-\sigma^2)\lambda_N)|^2 - \sigma^2 = \sigma^2(\sigma^2 - 1) \leq 0$ and $|z_+(\lambda_2)|^2 - \sigma^2 = 2\left(\frac{1-\sigma^2}{1+\sigma^2}\right)^2 + (\sigma^2 - 3)\left(\frac{1-\sigma^2}{1+\sigma^2}\right) + 1 - \sigma^2 = -\frac{2\sigma^2(\sigma^2-1)^2}{(1+\sigma^2)^2} \leq 0$. Together with the fact that $\lambda_k \leq \frac{4(\alpha_0^*+\beta_0^*)}{(\alpha_0^*)^2} = (1-\sigma^2)\lambda_N$, it follows that $\max_{j \in \{2, \dots, k\}} |z_+(\lambda_j)|^2 \leq \max\{|z_+(\lambda_2)|^2, |z_+((1-\sigma^2)\lambda_N)|^2\} \leq \sigma^2$. Combing the above, we conclude that $\max_{j \in \{2, \dots, N\}} \{|z_+(\lambda_j)|, |z_-(\lambda_j)|\} = \sigma$. Thus, the lower bound (8.43) is attainable and $K^* = [\frac{1-\sigma^2}{h^2\lambda_N}, \frac{3+\sigma^2}{2h\lambda_N}]$ leads to the optimal *asymptotic convergence factor* $r_{asy}^* = \sigma$. ■

8.3.2 Illustrative Example

We use a formation example with four agents moving on a plane to illustrate the results on the second-order consensus. The position of agent i at time kh is rep-

resented by the pair $[x_i(k), y_i(k)]^T$. Two configuration variables ($x_i(k)$ and $y_i(k)$) are decoupled and have a dynamic equation as in (8.40). An *undirected* communication graph with $(0, 1)$ -weights to model interactions among the four agents is shown in Fig. 8.2. The desired formation is specified as the vertices of a reg-

Figure 8.2: Communication graph

ular parallelogram. In particular, $F_1 = [0 \ 0]^T$, $F_2 = [5 \ -5]^T$, $F_3 = [10 \ 0]^T$ and $F_4 = [5 \ 5]^T$. All the agents are lined up in a column at the beginning, e.g., $[x_i(0), y_i(0)]^T = [0, 20 - 5i]^T$, $i = 1, \dots, 4$. Let the sampling interval $h = 0.25s$ and initial velocities be $\dot{x}(0) = [2 \ 0.5 \ 0.5 \ 2]^T$, $\dot{y}(0) = [0.5 \ -1 \ 0.5 \ 3]^T$, the control gain to achieve the optimal *asymptotic convergence factor* for each configuration variable is solved as $K^* = [1.6 \ 1.8]$ by Theorem 8.7. In view of Theorem 8.6, the set of feasible control gains solving the second-order consensus is computed as $\Omega = \{[\alpha, \beta] | \beta \in (0, 2), \alpha \in (0, 8\beta)\}$. Thus, we pick a feasible control gain $K_1 = [4, 1]$ and an infeasible control gain $K_2 = [3, 2]$ to compare their performance. It can be easily seen that the consensus by the optimal control gain (see Fig. 8.3) converges faster than the one by K_1 (see Fig. 8.4) and it fails to reach a consensus by using the control gain K_2 , see Fig. 8.5.

8.4 Consensusability with Encoded State Feedback

In this section, the exact state information of each agent's neighbor is not available due to the limited bandwidth of communication channels. Thus, at each time slot, agents can only transmit information with a finite number bits to their neighbors,

Figure 8.3: $K = K^*$: the formationability is achieved with the optimal *asymptotic convergence factor* $r_{asym}^* = 0.7746$.

Figure 8.4: $K = K_1$: the formationability is achieved with an *asymptotic convergence factor* $r_{asym} = 0.9354$.

which will be used to recover the state of the transmitter. Because of the information loss induced by quantization, consensusability of the multi-agents systems may be lost. The more number of bits can be sent for each transmission, the less information loss will be induced. The fundamental question is how many bits are needed to be communicated between each pair of agents to preserve consensusability? This is an interesting and difficult problem for multi-agent systems. The primary object of interest in this section is to provide a partial answer to the above question. Under *undirected* graphs, a dynamic encoding-decoding scheme will be constructed to estimate the state of the transmitter. Sufficient conditions on the data rate in relation to the agent dynamics and communication graphs are provided to achieve consensusability of the multi-agent systems. It is worth pointing out that to our knowledge, the quantized higher-order consensus problems have not been discussed in the literature.

Figure 8.5: $K = K_2$: it is impossible to achieve formationability.

8.4.1 Encoder/decoder Description

A uniform quantizer will be used to quantize the state variable of each agent. Precisely, define a $(2L + 1)$ -level uniform quantizer as follows

$$q_L(x) = \begin{cases} 0, & \text{if } -1/2 < x < 1/2; \\ i, & \text{if } \frac{2i-1}{2} \leq x < \frac{2i+1}{2}, i \in \{1, \dots, L\}, \\ L, & \text{if } x \geq L + \frac{1}{2}, \\ -q_L(-x), & \text{if } x \leq -1/2. \end{cases} \quad (8.46)$$

If $|x| \leq L + \frac{1}{2}$, the quantization error of the above $(2L + 1)$ -level uniform quantizer is computed as $|x - q_L(x)| \leq 1/2$. Motivated by [60] and neglecting the transmission delay between each pair of neighboring agents, the state estimator $\hat{x}_i(k)$ of $x_i(k)$, which can be simultaneously updated by agent i and its neighbor $j \in \mathcal{N}_i$, is proposed as follows:

$$\begin{cases} \hat{x}_i(0) = \mathbf{0}, \\ \hat{x}_i(k+1) = l(k)s_i(k) + A\hat{x}_i(k). \end{cases} \quad (8.47)$$

Here the quantizer output $s_i(k) \triangleq Q(\frac{x_i(k+1) - A\hat{x}_i(k)}{l(k)})$ is encoded into binary data and broadcasted to agent i 's neighbors via finite data rate communication channels. Since the quantized message $s_i(k)$ is accessible by agent i and its neighbors, (8.47) can be simultaneously processed by agent i and its neighbors. The product quantizer $Q(\cdot)$ is defined by $Q(\cdot) = [q_L(\cdot), \dots, q_L(\cdot)]^T$, whose dimension is compatible with its

input. The scaling function to make the quantizer unsaturated is designed by

$$l(k) = l_0 \gamma^k, \quad (8.48)$$

where $l_0 > 0$ and $\gamma \in (0, 1)$ are to be determined.

8.4.2 Consensusability Analysis

The main purpose of this subsection is to show that consensusability with perfect state feedback implies consensusability with encoded state feedback subject to a finite communication data rate. We need the following assumption:

Assumption 8.2. $\|x_i(0)\|_\infty \leq C_x, \forall i \in \mathcal{N}$ and C_x is known by all agents.

With the aid of the estimator in (8.47), the following controller is proposed:

$$u_i(k) = K \sum_{j=1}^N a_{ij} (\hat{x}_j(k) - \hat{x}_i(k)). \quad (8.49)$$

The main idea of using the above controller is to make the average state of the closed-loop systems evolve as an open-loop system. Specifically, by applying the control of (8.49) to the multi-agent systems, it follows that $\bar{x}(k+1) = A\bar{x}(k)$.

Theorem 8.8. *Given an undirected graph $\mathcal{G}$, suppose that Assumption 8.2 and conditions in Theorem 8.1 hold.*

- (a) *Select a $K \in \mathbb{R}^{1 \times n}$ such that $\rho(J(K)) < 1$, where $J(K)$ is defined in (8.16).*
- (b) *Choose any $\gamma \in (\rho(J(K)), 1)$ and $\epsilon \in (0, \frac{\gamma - \rho(J(K))}{\|J(K)\|})$. Let $\eta = \rho(\frac{J(K)}{\gamma}) + \epsilon \|\frac{J(K)}{\gamma}\| < 1$, $d_{\max} = \max_i \deg_i$ and*

$$M(K, \gamma) \triangleq \frac{\|A\|_\infty + 2d_{\max}\|BK\|_\infty}{2\gamma} + \frac{n^{3/2}\sqrt{N}M\lambda_N^2\|BK\|_\infty^2}{2\gamma^2(1-\eta)}.$$

- (c) *Select the quantization level parameter $L > M(K, \gamma) - \frac{1}{2}$ and the scaling parameter*

$$l_0 > \max \left\{ \frac{\|A\|_\infty C_x}{L + 1/2}, \frac{4\gamma^2 C_x (1 - \eta)}{\sqrt{n} \lambda_N \|BK\|_\infty} \right\}.$$

Then, under the coding/decoding scheme described in (8.47)-(8.48) and the controller (8.49), the consensus of the discrete-time multi-agent systems (8.1) is achieved. Moreover, the asymptotic convergence factor is upper bounded by γ , i.e., $r_{\text{asym}} \leq \gamma$.

Remark 8.8. The existence of a control gain in item (a) is based on Theorem 8.1. In Theorem 8.8, a smaller γ may lead to a better asymptotic convergence factor which also requires a higher number of quantization levels since for a fixed control gain K , $M(K, \gamma)$ increases as γ decreases. For the extreme case, $\gamma \searrow \rho(J(K))$, where $\searrow$ means approaching from the right hand side, then $M(K, \gamma)$ goes to infinity since $\epsilon \searrow 0$ and $1 - \eta \searrow 1 - \frac{\rho(J(K))}{\gamma} \searrow 0$. Thus, to achieve an asymptotic convergence factor, $r_{\text{asym}} = \rho(J(K))$, an infinite number of quantization levels is needed. Although it is unclear about the conservativeness of the lower bound of $M(K, \gamma)$ at this moment, it provides some insights on the relationship among the convergence rate, system dynamic and data rate.

Proof of Theorem 8.8:

Denote $\hat{x}(k) = [\hat{x}_1^T(k), \dots, \hat{x}_N^T(k)]^T$ and $s(k) = [s_1^T(k), \dots, s_N^T(k)]^T$, it follows from (8.47) that $\hat{x}(k+1) = l(k)s(k) + (I_N \otimes A)\hat{x}(k)$. Let the estimation error be $E_i(k) = x_i(k) - \hat{x}_i(k)$. Inserting the controller (8.49) into the discrete-time system (8.1) leads to that

$$x_i(k+1) = Ax_i(k) + BK \sum_{j=1}^N a_{ij}(x_j(k) - x_i(k)) - BK \sum_{j=1}^N a_{ij}(E_j(k) - E_i(k)). \quad (8.50)$$

Collect $E(k) = [E_1^T(k), \dots, E_N^T(k)]^T$, the whole system is compactly written by

$$x(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)x(k) + (\mathcal{L}_G \otimes BK)E(k). \quad (8.51)$$

Noting that $\bar{x}(k+1) = A\bar{x}(k)$ and $\mathcal{L}_G \mathbf{1} = \mathbf{0}$, it yields that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)\delta(k) + (\mathcal{L}_G \otimes BK)E(k). \quad (8.52)$$

Also, it is easy to derive that $x_i(k+1) - A\hat{x}_i(k) = AE_i(k) + BK \sum_{j=1}^N a_{ij}(\delta_j(k) -$

$\delta_i(k)) - BK \sum_{j=1}^N a_{ij}(E_j(k) - E_i(k))$. We rewrite it in a compact form as follows:

$$x(k+1) - (I_N \otimes A)\hat{x}(k) = (I_N \otimes A + \mathcal{L}_G \otimes BK)E(k) - (\mathcal{L}_G \otimes BK)\delta(k). \quad (8.53)$$

The dynamical equation to the state estimation error is expressed by

$$\begin{aligned} E(k+1) &= x(k+1) - \hat{x}(k+1) = x(k+1) - (I_N \otimes A)\hat{x}(k) - l(k)s(k) \\ &= (I_N \otimes A + \mathcal{L}_G \otimes BK)E(k) - (\mathcal{L}_G \otimes BK)\delta(k) - l(k) \\ &\quad \times Q \left(\frac{(I_N \otimes A + \mathcal{L}_G \otimes BK)E(k) - (\mathcal{L}_G \otimes BK)\delta(k)}{l(k)} \right). \end{aligned} \quad (8.54)$$

Let $W(k) = \frac{1}{l(k)}\delta(k)$, $Z(k) = \frac{1}{l(k)}E(k)$, then (8.52) and (8.54) jointly imply that

$$\begin{aligned} \gamma Z(k+1) &= (I_N \otimes A + \mathcal{L}_G \otimes BK)Z(k) - (\mathcal{L}_G \otimes BK)W(k) \\ &\quad - Q((I_N \otimes A + \mathcal{L}_G \otimes BK)Z(k) - (\mathcal{L}_G \otimes BK)W(k)), \end{aligned} \quad (8.55)$$

$$\gamma W(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)W(k) + (\mathcal{L}_G \otimes BK)Z(k). \quad (8.56)$$

We shall prove that the quantizer will never be saturated. First, due to that $\hat{x}(0) = \mathbf{0}$ and $\mathcal{L}_G \mathbf{1} = \mathbf{0}$, then $\|(I_N \otimes A + \mathcal{L}_G \otimes BK)Z(0) - (\mathcal{L}_G \otimes BK)W(0)\|_\infty = \frac{1}{l_0}\|(I_N \otimes A + \mathcal{L}_G \otimes BK)x(0) - (\mathcal{L}_G \otimes BK)(x(0) - \mathbf{1} \otimes \bar{x}(0))\|_\infty = \frac{1}{l_0}\|(I_N \otimes A)x(0)\|_\infty < L + 1/2$, which implies that the quantizer is unsaturated at time $k = 0$. Assume that the quantizer is unsaturated before time $k \geq 0$. In view of (8.55), it follows that $\sup_{1 \leq i \leq k+1} \|Z(i)\|_\infty \leq \frac{1}{2\gamma}$. Let $\widetilde{W}(k) = (\Phi \otimes I_n)^T W(k)$, $k \in \mathbb{N}$ and partition $\widetilde{W}(k)$ as $\widetilde{W}(k) = [\widetilde{W}_1^T(k), \widetilde{W}_2^T(k)]^T$, where $\widetilde{W}_1(k)$ is the collection of the first n elements of $\widetilde{W}(k)$. Then, one can easily verify that $\widetilde{W}_1(k) = \mathbf{0}$. Define $\widetilde{Z}(k) = (\Phi^T \mathcal{L}_G \otimes BK)Z(k) = [\widetilde{Z}_1^T(k), \widetilde{Z}_2^T(k)]^T$ in conformity with the partition of $\widetilde{W}(k)$, it follows from (8.56) that

$$\begin{aligned} \widetilde{W}_2(k+1) &= \frac{\text{diag}(A - \lambda_2 BK, \dots, A - \lambda_N BK)}{\gamma} \widetilde{W}_2(k) + \frac{1}{\gamma} \widetilde{Z}_2(k) \\ &= \left(\frac{J(K)}{\gamma} \right)^{k+1} \widetilde{W}_2(0) + \gamma^{-1} \sum_{i=0}^k \left(\frac{J(K)}{\gamma} \right)^{k-i} \widetilde{Z}_2(i). \end{aligned} \quad (8.57)$$

Since $\epsilon < \frac{\gamma - \rho(J(K))}{\|J(K)\|}$ and $\eta = \rho(\frac{J(K)}{\gamma}) + \epsilon \|\frac{J(K)}{\gamma}\| < 1$, it follows from Lemma 6.2 that

$$\left\| \left(\frac{J(K)}{\gamma} \right)^k \right\| \leq M\eta^k, \forall k \in \mathbb{N}, \quad (8.58)$$

where $M = \sqrt{n}(1 + \frac{2}{\epsilon})^{n-1}$. By (8.57), it is easy to get that $\|\widetilde{W}_2(k+1)\| \leq M\eta^{k+1}\|\widetilde{W}_2(0)\| + \gamma^{-1}M \sum_{i=0}^k \eta^{k-i}\|\widetilde{Z}_2(i)\|$. Now, we evaluate the two terms on the right hand side of the above inequality. It is obvious that $\|\widetilde{W}_2(0)\| = \|\widetilde{W}(0)\| = \|W(0)\| \leq \frac{\sqrt{Nn}}{l_0}\|\delta(0)\|_\infty \leq \frac{2C_x\sqrt{Nn}}{l_0}$. Similarly, we obtain that $\gamma^{-1} \sum_{i=0}^k \eta^{k-i}\|\widetilde{Z}_2(i)\| \leq \frac{1}{\gamma} \sum_{i=0}^k \eta^{k-i}\|P^T \mathcal{L}_G \otimes BK\| \|Z(i)\| = \frac{\lambda_N \|BK\|}{\gamma} \sum_{i=0}^k \eta^{k-i}\|Z(i)\| \leq \frac{n\lambda_N \sqrt{N}\|BK\|_\infty}{2\gamma^2(1-\eta)}(1 - \eta^{k+1})$. Since $\eta \in (0, 1)$, we obtain that $\|\widetilde{W}_2(k+1)\| < \max\{\frac{2MC_x\sqrt{Nn}}{l_0}, \frac{n\lambda_N M\sqrt{N}\|BK\|_\infty}{2\gamma^2(1-\eta)}\}$.

Thus, the quantizer at time $k+1$ is unsaturated since

$$\begin{aligned} & \|(I_N \otimes A + \mathcal{L}_G \otimes BK)Z(k+1) - (\mathcal{L}_G \otimes BK)W(k+1)\|_\infty \\ & \leq (\|I_N \otimes A\|_\infty + \|\mathcal{L}_G \otimes BK\|_\infty)\|Z(k+1)\|_\infty + \|\mathcal{L}_G \otimes BK\| \|W(k+1)\| \\ & \leq \frac{\|A\|_\infty + 2d_{\max}\|BK\|_\infty}{2\gamma} + \sqrt{n}\lambda_N\|BK\|_\infty\|W(k+1)\| \leq M(K, \gamma) < L + \frac{1}{2}. \end{aligned}$$

By induction, the $(2L+1)$ -level uniform quantizer will never be saturated. Together with the fact that $\|W(0)\|_\infty = \frac{\|\delta(0)\|_\infty}{l_0} \leq \frac{2C_x}{l_0}$, it yields that

$$\sup_{k \in \mathbb{N}} \|W(k)\|_\infty \leq \max \left\{ \frac{2C_x}{l_0}, \frac{n\lambda_N M\sqrt{N}\|BK\|_\infty}{2\gamma^2(1-\eta)} \right\} < \infty. \quad (8.59)$$

By the relationship between $W(k)$ and $\delta(k)$, it finally implies that $\lim_{k \rightarrow \infty} \|\delta(k)\|_\infty = 0$. Together with the fact that $\lim_{k \rightarrow \infty} a^{1/k} = 1, \forall a > 0$, (8.59) results in that $\frac{r_{asym}}{\gamma} = \sup_{\delta(0) \neq \mathbf{0}} \lim_{k \rightarrow \infty} \left(\frac{\|\gamma^{-k}\delta(k)\|}{\|\delta(0)\|} \right)^{\frac{1}{k}} = \sup_{\delta(0) \neq \mathbf{0}} \lim_{k \rightarrow \infty} \left(\frac{l_0\|W(k)\|}{\|\delta(0)\|} \right)^{\frac{1}{k}} \leq 1$. $\blacksquare$

8.4.3 Illustrative Example

The feasibility of the control protocol (8.4) is demonstrated by multi-agent systems with a scalar agent dynamics in this subsection. We consider four agents with an *undirected* graph shown in Fig. 8.2. Let the system matrices in (8.1) be $A = 1.1$ and $B = 1$. We choose a control gain $K = 0.21$ and $\gamma = 0.995$. Assume the initial states

of all agents are given by $x_i(0) = i$. By Theorem 8.8, the quantizer parameters are selected as follows: $L = 8$ and $l_0 = 3$. In this case, the number of quantization level is $2L + 1 = 17$, which can be represented by 4 bits since zero level in (8.46) is not necessary to be transmitted. Note that the lower bound on the number of quantizer parameter L in Theorem 8.8 may not be the minimal number for consensusability. From Fig. 8.6, it is clear to see that the deviation of each agent from the average state asymptotically converges to zero, i.e., the multi-agent systems eventually reach a consensus.

Figure 8.6: A consensus is asymptotically reached.

8.5 Summary

We have addressed a fundamental problem dealing with consensusability of discrete-time multi-agents systems w.r.t. a common control protocol. For a group of identical general linear systems with all the poles on or outside the unit circle, necessary and sufficient conditions on agent dynamics and communication graphs were obtained for consensusability, respectively under *undirected* and *directed* graphs. It was shown that the intrinsic entropy rate of the agent dynamics poses a fundamental limitation on the communication graph to reach a consensus. We also provided a simple necessary and sufficient condition for the existence of a common control gain to simultaneously stabilize a number of systems of a same state matrix but with their input matrices scaled by different positive constants. To evaluate the consensus performance, *asymptotic convergence factor* and *per-step convergence factor* were

introduced. We proved that the optimal *asymptotic convergence factor* among all possible control gains has a lower bound, which is again described by the intrinsic entropy rate and the *eigenratio* of an *undirected* graph and shown to be tight for some special cases. The results were generalized to study the formationability problem on *undirected* graphs and a necessary and sufficient condition for formationability was given as well. As an interesting special case, we discussed the discrete-time second-order consensus problem and constructed a control gain to achieve the optimal *asymptotic convergence factor*. A formation example demonstrated our main results on the second-order consensus. Motivated by the limited bandwidth constraint of a real communication channel, a dynamic encoding/decoding scheme and the corresponding control protocol were proposed to show that consensusability with perfect state feedback implies consensusability with encoded state feedback provided that the data rate is greater than an explicitly determined lower bound.

Chapter 9

Coordination of Multi-agent Systems via Output Feedback

This chapter studies the joint effect of agent dynamics and network topology on consensusability of linear discrete-time multi-agent systems via relative output feedback. We restrict ourselves to the single input agent dynamics and *undirected* graphs.¹ An observer-based distributed control protocol is proposed. A necessary and sufficient condition for consensusability under this control protocol is given, which explicitly reveals how the intrinsic entropy rate of the agent dynamics and the *eigenratio* of the *undirected* graph affect consensusability. As a special case, multi-agent systems with discrete-time double-integrator dynamic are discussed where a simple control protocol directly using two-step relative position feedback is provided to reach a consensus. Finally, the result is extended to solve the formation and formation-based tracking problems. The theoretical results are illustrated by simulations.

This chapter is organized as follows. Problem formulation is described in Section 9.1. Consensusability analysis is performed in Section 9.2, where some necessary and sufficient conditions are given for consensusability under two control protocols. In Section 9.3, consensusability of the discrete-time double-integrator systems is discussed by designing a simple control protocol. As important applications, the results are extended to investigate the formation and formation-based tracking problems in

¹The extension to multiple inputs and directed graphs can be made as in the previous chapter.

Section 9.4. Simulations are included in Section 9.5. Conclusion remarks are drawn in Section 9.6.

9.1 Problem Formulation

The dynamic of agent i takes the following form:

$$\begin{cases} x_i(k+1) &= Ax_i(k) + Bu_i(k), \\ y_i(k) &= Cx_i(k), \forall k \in \mathbb{N}, i \in \mathcal{V}, \end{cases} \quad (9.1)$$

where $x_i(k) \in \mathbb{R}^n$, $u_i(k) \in \mathbb{R}$ and $y_i(k) \in \mathbb{R}^m$ represent the state, control input and output of agent i , respectively. $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^n$ and $C \in \mathbb{R}^{m \times n}$ are respectively the state, input and observation matrices.

Consider the situation that each agent does not know its exact output but can measure the output relative to those of its neighboring agents. For instance, in vehicle coordination, the vision-based sensor on a vehicle can not directly locate the position of the vehicle in a global coordinate system but can measure the relative position to its neighbors. While in networked clock synchronization, we are more concerned with the time difference between each pair of clocks. In addition, the communication link of $\mathcal{E}$ is assumed to be perfect in the sense that we ignore effects due to quantization, packet dropout, transmission errors and delays.

By adapting to the available information for each agent, we say a control protocol *admissible* if each agent generates its control input signal by relying on relative outputs. Generally, *admissible* control protocols can be categorized depending on whether they are *static* or *dynamic*. A *dynamic* protocol uses memory and can be potentially more powerful. In the sequel, two *admissible* control protocols will be proposed. Precisely, we first adopt a *static* control protocol:

$$u_i(k) = F \sum_{j \in \mathcal{N}_i} a_{ij}(y_j(k) - y_i(k)) \triangleq F\zeta_i(k), F \in \mathbb{R}^{1 \times m}. \quad (9.2)$$

Definition 9.1. Given a graph $\mathcal{G}$, the discrete-time multi-agent systems (9.1) are said to be consensusable under the static protocol (9.2) if for any finite $x_i(0), \forall i \in \mathcal{V}$,

the protocol can asymptotically drive all agents close to each other, i.e.,

$$\lim_{k \rightarrow \infty} \|x_i(k) - x_j(k)\| = 0, \forall i, j \in \mathcal{V}. \quad (9.3)$$

The second *admissible* control protocol is an observer-based *dynamic* protocol that depends on an internal controller state. In the previous chapter, the following consensus protocol using the relative state feedback is exploited:

$$u_i(k) = K \sum_{j \in \mathcal{N}_i} a_{ij}(x_j(k) - x_i(k)) \triangleq K\xi_i(k), K \in \mathbb{R}^{1 \times n}. \quad (9.4)$$

Since $\xi_i(k)$ is no longer available in the current framework, a very natural thing is to design an observer to estimate $\xi_i(k)$ for the control design. In consideration of the agent dynamics, an observer-based control protocol for agent i will be studied.

$$\begin{cases} \hat{\xi}_i(k+1) &= A\hat{\xi}_i(k) + B \sum_{j \in \mathcal{N}_i} a_{ij}(u_j(k) - u_i(k)) + L(\zeta_i(k) - C\hat{\xi}_i(k)), \\ u_i(k) &= K\hat{\xi}_i(k), L \in \mathbb{R}^{n \times m}, K \in \mathbb{R}^{1 \times n}. \end{cases} \quad (9.5)$$

At time k , agent i computes the aggregate relative measurements to those of its neighbors, denoted by $\zeta_i(k)$. Together with control inputs from its neighbors, $u_j(k), j \in \mathcal{N}_i$, which will be received before time $k+1$, the agent updates its internal controller state to obtain $\hat{\xi}_i(k+1)$ and produces the control input $u_i(k+1)$. It is clear that the *dynamic* control protocol (9.5) is *admissible*. Compared to the *static* protocol (9.2), this *dynamic* protocol requires each agent to broadcast its control input to its neighboring agents.

Observe the special case that the initial estimate is perfect, i.e., $\xi_i(0) = \hat{\xi}_i(0)$, it can be easily shown that $\xi_i(k) = \hat{\xi}_i(k), \forall k \in \mathbb{N}$. When the consensus is reached, the internal controller state $\hat{\xi}_i(k)$ of this case becomes zero. By taking this into consideration, it is reasonable to impose an additional condition on the definition of consensus that all the internal states $\hat{\xi}_i(k), \forall i \in \mathcal{V}$ should asymptotically converge to zero.

Definition 9.2. Given a graph $\mathcal{G}$, the discrete-time multi-agent systems (9.1) are said to be consensusable under the dynamic protocol (9.5) if for any finite $x_i(0), \forall i \in$

$\mathcal{V}$, the protocol can asymptotically drive the states of all agents close to each other and all the controller internal states to zero, i.e.,

$$\lim_{k \rightarrow \infty} \|x_i(k) - x_j(k)\| = 0 \ \& \ \lim_{k \rightarrow \infty} \|\hat{\xi}_i(k)\| = 0, \forall i, j \in \mathcal{V}. \quad (9.6)$$

9.2 Consensusability Analysis

In this section, a necessary and sufficient condition for consensusability under the *static* control protocol (9.2) is established at first. Noting that the verification of this condition is nontrivial, we proceed to seek a necessary and sufficient condition for consensusability under the *dynamic* control protocol (9.5). The roles of *undirected* graphs and agent dynamics on consensusability are explicitly quantified for agent dynamics satisfying that all the open-loop poles lie on or outside the unit circle.

9.2.1 Consensus under Static Protocol

Theorem 9.1. *Given an undirected graph $\mathcal{G}$, the discrete-time multi-agent systems (9.1) are consensusable under the static control protocol (9.2) if and only if there exists a common gain $F \in \mathbb{R}^{1 \times m}$ such that $\rho(A - \lambda_j BFC) < 1, \forall j \in \{2, \dots, N\}$.*

Proof. Denote the average state of all agents by $\bar{x}(k) \triangleq \frac{1}{N} \sum_{i=1}^N x_i(k) = \frac{1}{N} (\mathbf{1}^T \otimes I_n) x(k)$, where $x(k) \triangleq [x_1^T(k), \dots, x_N^T(k)]^T$, and the displacement vector by $\delta_i(k) \triangleq x_i(k) - \bar{x}(k)$. It is clear that $\lim_{k \rightarrow \infty} \|\delta_i(k)\| = 0, \forall i \in \mathcal{V}$ is inequivalent to consensusability of the multi-agent systems (9.1). Inserting the control protocol (9.2) into the agent dynamics, the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_{\mathcal{G}} \otimes BFC)x(k). \quad (9.7)$$

Considering that $\mathbf{1}^T \mathcal{L}_{\mathcal{G}} = \mathbf{0}^T$, we obtain

$$\begin{aligned} \bar{x}(k+1) &= \frac{1}{N} (\mathbf{1}^T \otimes A)x(k) - \frac{1}{N} (\mathbf{1}^T \mathcal{L}_{\mathcal{G}} \otimes BFC)x(k) \\ &= A\bar{x}(k). \end{aligned} \quad (9.8)$$

Let $\delta(k) = [\delta_1^T(k), \dots, \delta_N^T(k)]^T$, subtracting (9.7) from (9.8) leads to that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BFC)\delta(k). \quad (9.9)$$

Select $\phi_i \in \mathbb{R}^N$ such that $\phi_i^T \mathcal{L}_G = \lambda_i \phi_i^T$ and let $\Phi = [\frac{1}{\sqrt{N}}, \phi_2, \dots, \phi_N]$ to transform $\mathcal{L}_G$ into a diagonal form: $\text{diag}(0, \lambda_2, \dots, \lambda_N) = \Phi^T \mathcal{L}_G \Phi$. Further, using the property of Kronecker product gives that

$$\begin{aligned} (\Phi \otimes I_n)^T (I_N \otimes A - \mathcal{L}_G \otimes BFC) (\Phi \otimes I_n) &= I_N \otimes A - \Phi^T \mathcal{L}_G \Phi \otimes BFC \\ &= \text{diag}(A, A - \lambda_2 BFC, \dots, A - \lambda_N BFC). \end{aligned} \quad (9.10)$$

Denote $\tilde{\delta}(k) = (\Phi \otimes I_n)^T \delta(k)$ and partition $\tilde{\delta}(k) \in \mathbb{R}^{nN}$ into two parts, i.e., $\tilde{\delta}(k) = [\tilde{\delta}_1^T(k), \tilde{\delta}_2^T(k)]^T$, where $\tilde{\delta}_1(k) \in \mathbb{R}^n$ is a vector consisting of the first n elements of $\tilde{\delta}(k)$. Then, $\tilde{\delta}_1(k) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \delta_i(k) = \mathbf{0}$. In view of (9.9) and (9.10), it yields that

$$\tilde{\delta}_2(k+1) = \text{diag}(A - \lambda_2 BFC, \dots, A - \lambda_N BFC) \tilde{\delta}_2(k). \quad (9.11)$$

The rest of the proof is straightforward. ■

This result reveals how the eigenvalues of the Laplacian matrix affect consensusability. However, the verification of the condition in Theorem 9.1 is nontrivial although some conservative sufficient conditions can be given in terms of linear matrix inequalities [13]. An explicit characterization of the effect of these eigenvalues on consensusability is given in the following subsection.

9.2.2 Consensus under Dynamic Protocol

Theorem 9.2. *Given an undirected graph $\mathcal{G}$, the discrete-time multi-agent systems (9.1) satisfying that all the eigenvalues of A lie on or outside the unit circle are consensusable under the dynamic control protocol (9.5) if and only if the following conditions hold.*

- (a) (A, B, C) are controllable and observable;

- (b) *Each agent cannot change too fast. Precisely, the product of the unstable eigenvalues of A is upper bounded by the following strict inequality:*

$$\prod_j |\lambda_j^u(A)| < \frac{1 + \lambda_2/\lambda_N}{1 - \lambda_2/\lambda_N}. \quad (9.12)$$

Moreover, if the above conditions hold, select a ζ such that $\prod_j |\lambda_j^u(A)| < \zeta^{-1} \leq \frac{1+\lambda_2/\lambda_N}{1-\lambda_2/\lambda_N}$. Then, the control gain $K = \frac{2}{\lambda_2 + \lambda_N} \frac{B^T P A}{B^T P B}$ solves the consensus problem, where $P > 0$ is a positive solution to the following modified algebraic Riccati inequality:

$$P - A^T P A + (1 - \zeta^2) \frac{A^T P B B^T P A}{B^T P B} > 0. \quad (9.13)$$

The observer gain L is chosen to make $\rho(A - LC) < 1$.

Proof. Define $\tilde{\xi}_i(k)$ as the estimation error of $\xi_i(k)$, i.e., $\tilde{\xi}_i(k) = \hat{\xi}_i(k) - \xi_i(k)$. Inserting the control protocol (9.5) into each agent dynamics, the dynamical equation of $x(k)$ can be written as

$$x(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)x(k) + (I_N \otimes BK)\tilde{\xi}(k). \quad (9.14)$$

Similar to the proof of Theorem 9.1, it is easy to show that

$$\delta(k+1) = (I_N \otimes A - \mathcal{L}_G \otimes BK)\delta(k) + (I_N \otimes BK)\tilde{\xi}(k). \quad (9.15)$$

Let $E(k) = (\Phi \otimes I_n)^T (I_N \otimes BK)\tilde{\xi}(k)$ and partition it into two parts $E(k) = [E_1^T(k), E_2^T(k)]^T$ where $E_1(k) \in \mathbb{R}^n$ is a vector consisting of the first n elements of $E(k)$. Following a similar arguments of the proof of Theorem 9.1, we have that

$$\begin{cases} \tilde{\delta}_1(k) &= 0, \forall k \in \mathbb{N}. \\ \tilde{\delta}_2(k+1) &= \text{diag}(A - \lambda_2 BK, \dots, A - \lambda_N BK)\tilde{\delta}_2(k) + E_2(k). \end{cases} \quad (9.16)$$

Necessity: By (9.1), it follows that

$$\xi_i(k+1) = A\xi_i(k) + BK \sum_{j \in \mathcal{N}_i} a_{ij}(\hat{\xi}_j(k) - \hat{\xi}_i(k)). \quad (9.17)$$

Together with (9.5), the error dynamic of $\tilde{\xi}_i(k)$ is written by $\tilde{\xi}_i(k+1) = (A - LC)\tilde{\xi}_i(k)$. Assume that the multi-agent systems (9.1) reach a consensus under the *dynamic* protocol (9.5), it yields that

$$\begin{aligned} \lim_{k \rightarrow \infty} \|\tilde{\xi}_i(k)\| &= \lim_{k \rightarrow \infty} \|\hat{\xi}_i(k) - \xi_i(k)\| \leq \lim_{k \rightarrow \infty} \|\hat{\xi}_i(k)\| + \lim_{k \rightarrow \infty} \|\xi_i(k)\| \\ &\leq \|K\| \sum_{j \in \mathcal{N}_i} a_{ij} \lim_{k \rightarrow \infty} \|x_j(k) - x_i(k)\| = 0, \forall i \in \mathcal{V}. \end{aligned} \quad (9.18)$$

Thus, we get that $\rho(A - LC) < 1$. This implies that (C, A) is detectable.

Now, we consider a special case that the initial estimate of $\xi_i(0), \forall i \in \mathcal{V}$ is perfect. By the error dynamic of $\tilde{\xi}_i(k)$, it is easy to see that $\tilde{\xi}_i(k) = 0, \forall i \in \mathcal{V}$, which further implies that $E_2(k) = 0, \forall k \in \mathbb{N}$. In light of (9.16), it immediately follows that $\rho(A - \lambda_j BK) < 1, \forall j \in \{2, \dots, N\}$. The rest of the proof of the necessity follows from Lemma 8.3.

Sufficiency: Since (A, B) is stabilizable, there exists a positive definite solution P to the algebraic Riccati inequality (9.13). In view of Lemma 8.3, the proposed control gain K can simultaneously stabilize $(A, \lambda_j B), \forall j \in \{2, \dots, N\}$, i.e.,

$$\varrho \triangleq \max_{j \in \{2, \dots, N\}} \rho(A - \lambda_j BK) < 1. \quad (9.19)$$

In addition, the observer gain L will make the estimation error asymptotically converge to zero, i.e., $\lim_{k \rightarrow \infty} \tilde{\xi}_i(k) = 0$, which further follows that $\lim_{k \rightarrow \infty} \|E_2(k)\| = 0$. Denote $J(K) = \text{diag}(A - \lambda_2 BK, \dots, A - \lambda_N BK)$, it follows from (9.16) that

$$\tilde{\delta}_2(k+1) = J(K)^{k+1} \tilde{\delta}_2(0) + \sum_{i=0}^k J(K)^{k-i} E_2(i). \quad (9.20)$$

Select a positive ϵ such that $\epsilon < \frac{1-\varrho}{\|J(K)\|}$ and $\eta = \varrho + \epsilon \|J(K)\| < 1$. By Lemma 6.2, it follows that $\|J(K)^k\| \leq M \eta^k$, where $M = \sqrt{n}(1 + 2/\epsilon)^{n-1}$. Thus, we obtain that $\|\tilde{\delta}_2(k+1)\| \leq M \left(\eta^{k+1} + \sum_{i=0}^k \eta^{k-i} \|E_2(i)\| \right)$. Consider an auxiliary system as follows: $z_{k+1} = \eta z_k + \|E_2(k)\|, z_0 = 1$. In view of Lemma 4.1, we have that

$$\lim_{k \rightarrow \infty} z_k = \frac{\lim_{k \rightarrow \infty} \|E_2(k)\|}{1 - \eta} = 0. \quad (9.21)$$

By iteration, it is clear that $z_{k+1} = \eta^{k+1} + \sum_{i=0}^k \eta^{k-i} \|E_2(i)\|$. Hence, we have proved that $\lim_{k \rightarrow \infty} \|\tilde{\delta}_2(k)\| = 0$. Together with that $\tilde{\delta}_1(k) = 0, \forall k \in \mathbb{N}$, it follows that $\lim_{k \rightarrow \infty} \|\delta(k)\| = 0$. Thus, we get that $\lim_{k \rightarrow \infty} \|x_i(k) - x_j(k)\| = 0, \forall i, j \in \mathcal{V}$, which further implies that $\lim_{k \rightarrow \infty} \|\xi_i(k)\| = 0, \forall i \in \mathcal{V}$. Moreover, the following statement is straightforward: $\lim_{k \rightarrow \infty} \|\hat{\xi}_i(k)\| \leq \lim_{k \rightarrow \infty} \|\xi_i(k)\| + \lim_{k \rightarrow \infty} \|\tilde{\xi}_i(k)\| = 0$. By Definition 9.2, the proof is completed. $\blacksquare$

9.3 Special Case: Double-Integrator Systems

Consider discrete-time double-integrator systems for each agent as follows:

$$\begin{cases} x_i(k+1) = x_i(k) + hv_i(k), \\ v_i(k+1) = v_i(k) + hu_i(k), \end{cases} \quad \forall i \in \mathcal{V}, k \in \mathbb{N}, \quad (9.22)$$

where h is the sampling interval, $x_i(k) \in \mathbb{R}$ and $v_i(k) \in \mathbb{R}$ respectively correspond to the position and velocity of agent i at time kh . $u_i(k) \in \mathbb{R}$ is the control input.

Under this setting, A, B, C are respectively written as

$$A = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ h \end{bmatrix}, C = [1 \ 0].$$

9.3.1 One-step Relative Position Feedback

Consider the situation that each agent does not know its position in a global coordinate system but can measure the position relative to those of its neighboring agents. One may attempt to reach a consensus by adopting a control protocol the form:

$$u_i(k) = \gamma \sum_{j=1}^N a_{ij} (x_j(k) - x_i(k)), \quad \gamma \in \mathbb{R}. \quad (9.23)$$

Intuitively, the above control protocol only uses relative position information, it may not be able to drive the multi-agent system to reach a consensus.

Theorem 9.3. *The second-order multi-agent systems (9.22) can not reach a consensus under the control protocol (9.23) for any graph.*

Proof. In view of (9.16), it can be similarly established that

$$\tilde{\delta}_j(k+1) = (A - \lambda_j \gamma BC) \tilde{\delta}_j(k), \forall j \in \{2, \dots, N\}. \quad (9.24)$$

It is straightforward that

$$\det(zI_2 - (A - \lambda_j \gamma BC)) = z^2 - 2z + 1 + \lambda_j h^2 \gamma. \quad (9.25)$$

Together with (9.24), we can not guarantee that for any finite initial state $\xi_0(k)$, $\lim_{k \rightarrow \infty} \|\tilde{\delta}(k)\| \neq 0$ since (9.25) contains at least one unstable root. ■

9.3.2 Two-step Relative Position Feedback

Due to the distinct feature of the double integrator system, the relative velocity can be accessed by using the relative position information with one-step delay. For example, by (9.22), it follows that $v_j(k-1) = \frac{x_j(k) - x_j(k-1)}{h}$. Thus, we study the following control protocol. Let $x_j(k) = 0, \forall k < 0$,

$$\begin{aligned} u_i(k) &= \sum_{j=1}^N a_{ij} [\gamma_0(x_j(k) - x_i(k)) + \gamma_1(v_j(k-1) - v_i(k-1))] \\ &\triangleq \sum_{j=1}^N a_{ij} [\alpha(x_j(k) - x_i(k)) + \beta(x_j(k-1) - x_i(k-1))]. \end{aligned} \quad (9.26)$$

The above protocol requires each agent to store the relative position feedback at the previous step. Under such a simple protocol (9.26), a connected graph is also necessary and sufficient for reaching a consensus.

Theorem 9.4. *Given an undirected graph $\mathcal{G}$, the second-order multi-agent systems (9.22) are consensusable under the control protocol (9.26) if and only if the communication graph is connected. Moreover, if this condition holds, (α, β) in the protocol (9.26) can be selected from the set*

$$\Omega_c \triangleq \left\{ (\alpha, \beta) \mid \max\left\{-\frac{1}{h^2}, -\frac{1}{\lambda_N h^2}\right\} < \beta < 0, \alpha = -\frac{\lambda_N h^2 \beta^2 + 3\beta}{2} \right\}. \quad (9.27)$$

Proof. Similar to the proof of Theorem 9.1, we obtain that $\forall j \in \{2, \dots, N\}$,

$$\tilde{\delta}_j(k+1) = (A - \alpha\lambda_j BC)\tilde{\delta}_j(k) - \beta\lambda_j BC\tilde{\delta}_j(k-1). \quad (9.28)$$

Let $\Delta_j(k) = [\tilde{\delta}_j^T(k-1), \tilde{\delta}_j^T(k)]^T$, where $\tilde{\delta}_j(k) = 0, \forall k < 0$. In view of (9.28), the dynamical equation of $\Delta_j(k)$ is expressed by

$$\begin{aligned} \Delta_j(k+1) &= \begin{bmatrix} 0 & I_2 \\ -\beta\lambda_j BC & A - \alpha\lambda_j BC \end{bmatrix} \Delta_j(k) \\ &\triangleq M_j(\alpha, \beta)\Delta_j(k), \forall j \in \{2, \dots, N\}. \end{aligned} \quad (9.29)$$

Thus, a necessary and sufficient condition for the multi-agent systems (9.22) to reach a consensus is that $\rho(M_j(\alpha, \beta)) < 1, \forall j \in \{2, \dots, N\}$.

Necessity: If the graph is not connected, it immediately follows that $\lambda_2 = 0$, which implies that $\rho(M_2(\alpha, \beta)) = 1, \forall \alpha, \beta \in \mathbb{R}$. In view of (9.29), we cannot guarantee that $\lim_{k \rightarrow \infty} \|\Delta_j(k)\| = 0, \forall \Delta_j(0) \in \mathbb{R}^4$. This contradicts with Definition 9.1.

Sufficiency: We show that for any connected graph and any $(\alpha, \beta) \in \Omega_c$, it holds that $\rho(M_j(\alpha, \beta)) < 1, \forall j \in \{2, \dots, N\}$. It is easy to compute that

$$\det(zI_4 - M_j(\alpha, \beta)) = z(z^3 - 2z^2 + (1 + \lambda_j h^2 \alpha)z + \lambda_j h^2 \beta). \quad (9.30)$$

Let the polynomial be $f(z) = z^3 - 2z^2 + (1 + x)z + y$. By using the Jury's stability test [3], it can be shown that all roots of $f(z)$ are inside the unit circle *if and only if* $(x, y) \in \Omega$, where Ω is the *stability region* shown in Fig 9.1 and computed as follows: $\Omega \triangleq \{(x, y) | -y < x < -y^2 - 2y\}$. Finally, it can be verified that for any $(\alpha, \beta) \in \Omega_c$, $(\lambda_j \alpha h^2, \lambda_j \beta h^2) \in \Omega, \forall j \in \{2, \dots, N\}$.

Together with (9.30), the proof is completed. ■

9.4 Applications

In this section, the result of Section 9.3 is applied to solve the formation and formation-based tracking problems.

Figure 9.1: Stability region: Ω .

9.4.1 Vehicle Formation on Graphs

As an important application, we study the vehicle formation problem via relative position feedback. The vehicle dynamical equation is described by the discrete-time double-integrator system (9.22). Given an arbitrary formation vector $f = [f_1, \dots, f_N]^T \in \mathbb{R}^N$, where f_i represents the desired separation of agent i from the centroid of all agents, the objective is to design a simple control protocol such that the vehicles reach the desired formation. Motivated by (9.26), the formationability of the following controller will be investigated:

$$u_i(k) = \sum_{j=1}^N a_{ij} [\alpha(x_j(k) - x_i(k) - f_j + f_i) + \beta(x_j(k-1) - x_i(k-1) - f_j + f_i)]. \quad (9.31)$$

Definition 9.3. *Given an undirected graph $\mathcal{G}$, the second-order multi-agent systems (9.22) are said to be formationable under the protocol (9.31) if for any finite initial position $x_i(0)$ and velocity $v_i(0)$, $i \in \mathcal{V}$, there exists a pair of $(\alpha, \beta) \in \mathbb{R}^2$ such that*

$$\lim_{k \rightarrow \infty} \|(x_i(k) - f_i) - (x_j(k) - f_j)\| = 0, \forall i, j \in \mathcal{V}. \quad (9.32)$$

Theorem 9.5. *Given an undirected graph $\mathcal{G}$, the second-order multi-agent systems (9.22) are formationable under the control protocol (9.31) if and only if the communication graph is connected. Moreover, if this condition holds, (α, β) in the protocol (9.31) can be chosen from the set Ω_c of (9.27).*

Proof. Let $\bar{f} = \frac{1}{N} \sum_{j=1}^N f_j$ be the average of the formation vector. Denote the displacement vector by $\delta_i(k) = [x_i(k) - f_i - \bar{x}(k) + \bar{f}, v_i(k) - \bar{v}(k)]^T$ and $\delta(k) = [\delta_1(k)^T, \dots, \delta_N(k)^T]^T$. Inserting the controller of (9.31) into (9.22) leads to that

$$\delta(k+1) = (I_N \otimes A - \alpha \mathcal{L}_{\mathcal{G}} \otimes BC) \delta(k) - (\beta \mathcal{L}_{\mathcal{G}} \otimes BC) \delta(k-1). \quad (9.33)$$

The rest of the proof follows from that of Theorem 9.4. ■

9.4.2 Formation-based Tracking Problem

The dynamic of the uncooperative leader is described by the following constant velocity model:

$$\begin{cases} x_0(k+1) &= x_0(k) + hv_0(k), \\ v_0(k+1) &= v_0(k) + hu_0(k), k \in \mathcal{V}, \end{cases} \quad (9.34)$$

where the control input $u_0(k)$ is an independent and identically distributed (i.i.d.) random process with zero mean and $\mathbb{E}[|u_0(k)|^2] = \sigma^2$.

The connection weights vector between followers and the leader is denoted by $b = [b_1, \dots, b_N]^T$, where b_i is positive if and only if the leader is a neighbor of agent i , otherwise $b_i = 0$. The leader's neighboring agent can measure its position relative to the leader. The goal is to design a simple distributed controller such that the center of all vehicles (except the leader), denoted by $\bar{x}(k) = \frac{1}{N} \sum_{i=1}^N x_i(k)$, asymptotically tracks the leader while keep a given formation vector f . To this purpose, assume that the average of the formation vector $\bar{f}$ is accessible to those vehicles that are connected to the leader. We propose the following control protocol:

$$\begin{aligned} u_i(k) = & \sum_{j=1}^N a_{ij} [\alpha(x_j(k) - x_i(k) - f_j + f_i) + \beta(x_j(k-1) - x_i(k-1) - f_j + f_i)] \\ & - b_i [\alpha(x_i(k) - x_0(k) - f_i + \bar{f}) + \beta(x_i(k-1) - x_0(k-1) - f_i + \bar{f})]. \end{aligned} \quad (9.35)$$

The summation of the first square bracket is to make the vehicles to maintain the given formation vector f while the rest is used to drive the center of the vehicles to

asymptotically track the leader.

Denote $\mathbb{R}_{\geq 0}$ the set of non-negative real numbers. A function $g : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ is said to be of class $\mathcal{K}_\infty$ if it is continuous, strictly increasing and unbounded and it crosses the origin.

Definition 9.4. *Given an undirected graph $\mathcal{G}$, the formation-based tracking problem associated with the second-order multi-agent systems (9.22) and the leader (9.34) is said to be solvable under the protocol (9.35) if for any finite initial position $x_i(0)$ and velocity $v_i(0), i \in \mathcal{V} \cup \{0\}$, there exist a pair of $(\alpha, \beta) \in \mathbb{R}^2$ and $g \in \mathcal{K}_\infty$ such that*

$$\begin{cases} \limsup_{k \rightarrow \infty} \mathbb{E}[\|(x_i(k) - f_i) - (x_j(k) - f_j)\|^2] \leq g(\sigma^2), \\ \limsup_{k \rightarrow \infty} \mathbb{E}[\|\bar{x}(k) - x_0(k)\|^2] \leq g(\sigma^2), \forall i, j \in \mathcal{V}, \end{cases} \quad (9.36)$$

where the mathematical expectation is taken w.r.t. the process $\{u_0(k)\}_{k \in \mathbb{N}}$.

Denote the index of the leader by 0 and $\mathcal{V}' = \mathcal{V} \cup \{0\}$. Similarly, denote the new adjacency matrix $\mathcal{A}' = \begin{bmatrix} 0 & b \\ b^T & \mathcal{A} \end{bmatrix}$ and the corresponding edge set $\mathcal{E}'$. Let $\mathcal{L}' = \mathcal{L}_\mathcal{G} + \text{diag}(b_1, \dots, b_N)$ and write the ascending order of the eigenvalues of $\mathcal{L}'$ by $\lambda'_1 \leq \lambda'_2 \leq \dots \leq \lambda'_N$. It is clear that the new graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}', \mathcal{A}'\}$ is generated by adding an *undirected* edge from the agent i to the leader if $b_i \neq 0$. Note that in fact only the follower can take relative measurements to its neighbors. The following result solves the formation-based leader-follower consensus problem.

Theorem 9.6. *Given an undirected graph $\mathcal{G}$, the control protocol (9.35) solves the formation-based tracking problem associated with the second-order multi-agent systems (9.22) and the leader (9.34) if and only if*

- (a) *The communication graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}', \mathcal{A}'\}$ is connected;*
- (b) *At least one agent connects to the leader, i.e., $\sum_{i=1}^N b_i \neq 0$.*

Moreover, if the above conditions hold, $(\alpha, \beta) \in \Omega'_c$ solves the formation-based tracking problem, where Ω'_c is given by

$$\Omega'_c \triangleq \left\{ (\alpha, \beta) \mid \max\left\{-\frac{1}{h^2}, -\frac{1}{\lambda'_N h^2}\right\} < \beta < 0, \alpha = -\frac{\lambda'_N h^2 \beta^2 + 3\beta}{2} \right\}. \quad (9.37)$$

$g \in \mathcal{K}_\infty$ can be chosen as a linear function, i.e., $g(\sigma^2) = c\sigma^2$ where the positive number c is a constant depending on h^2 , (α, β) and $\lambda'_j, j \in \mathcal{V}$.

Proof. Define the “tracking error” of agent i by $\delta_i(k) = [x_i(k) - f_i - x_0(k) + \bar{f}, v_i(k) - v_0(k)]^T$. Together with that if $g_1, g_2 \in \mathcal{K}_\infty$, $g_1 + g_2 \in \mathcal{K}_\infty$, it is easy to verify that the solvability of the formation-based leader-follower consensus problem is equivalent to that there exists a $g_i^\delta \in \mathcal{K}_\infty$ such that $\lim_{k \rightarrow \infty} \mathbb{E}[\|\delta_i(k)\|^2] \leq g_i^\delta(\sigma^2), \forall i \in \mathcal{V}$. Collect $\delta_i(k)$ to get a new vector: $\delta(k) = [\delta_1^T(k), \dots, \delta_N^T(k)]^T$. Inserting the control protocol (9.35) into (9.22) leads to that

$$\delta(k+1) = (I_N \otimes A - \alpha \mathcal{L}' \otimes BC) \delta(k) - (\beta \mathcal{L}' \otimes BC) \delta(k-1) - (\mathbf{1} \otimes \begin{bmatrix} 0 \\ h \end{bmatrix}) u_0(k). \quad (9.38)$$

Select $\psi_i \in \mathbb{R}^N$ such that $\psi_i^T \mathcal{L}' = \lambda'_i \psi_i^T, \forall i \in \mathcal{V}$. Form the unitary matrix $\Psi = [\psi_1, \psi_2, \dots, \psi_N]$ to transform $\mathcal{L}'$ into a diagonal form.

$$\text{diag}(\lambda'_1, \lambda'_2, \dots, \lambda'_N) = \Psi^T \mathcal{L}' \Psi. \quad (9.39)$$

Define $\tilde{\delta}(k) = (\Psi \otimes I_2)^T \delta(k)$ and partition it in conformity with $\delta(k)$. The same partition pattern is applied to $\Psi \mathbf{1} \otimes \begin{bmatrix} 0 \\ h \end{bmatrix} \triangleq [q_1^T, \dots, q_N^T]^T$. It follows that

$$\tilde{\delta}_j(k+1) = (A - \alpha \lambda'_j BC) \tilde{\delta}_j(k) - \beta \lambda'_j BC \tilde{\delta}_j(k-1) - q_j u_0(k). \quad (9.40)$$

Let $M'_j(\alpha, \beta) \triangleq \begin{bmatrix} 0 & I_2 \\ -\beta \lambda'_j BC & A - \alpha \lambda'_j BC \end{bmatrix}$ and $\Delta_j(k) = [\tilde{\delta}_j^T(k), \tilde{\delta}_j^T(k-1)]^T$, we obtain that $\Delta_j(k+1) = M'_j(\alpha, \beta) \Delta_j(k) + [\mathbf{0}^T, q_j^T]^T u_0(k)$. Denote $P_j(k) \triangleq \mathbb{E}[\Delta_j(k) \Delta_j(k)^T]$ and the zero matrix $0_2 \in \mathbb{R}^{2 \times 2}$, it is easy to derive that

$$P_j(k+1) = M'_j(\alpha, \beta) P_j(k) M'_j(\alpha, \beta)^T + \sigma^2 \text{diag}(0_2, q_j q_j^T). \quad (9.41)$$

Necessity: By Definition 9.4, it is clear that there exists a $g' \in \mathcal{K}_\infty$ such that $\limsup_{k \rightarrow \infty} \mathbb{E}[\|P_j(k)\|^2] \leq g'(\sigma^2)$. Jointly with (9.41), this implies that $\rho(M'_j(\alpha, \beta)) < 1, \forall j \in \mathcal{V}$. Similar to the proof of Theorem 9.4, it is obvious that a necessary

condition should be $\lambda'_j > 0, \forall j \in \mathcal{V}$. Thus, we obtain that $\sum_{i=1}^N b_i \neq 0$ since otherwise, $\mathcal{L}' = \mathcal{L}_G$, which contains at least one zero eigenvalue, e.g. $\lambda'_1 = 0$. In addition, the Laplacian matrix corresponding to the new graph $\mathcal{G}'$ can be written as

$$\mathcal{L}_{\mathcal{G}'} = \begin{bmatrix} \sum_{i=1}^N b_i & -b \\ -b^T & \mathcal{L}' \end{bmatrix}.$$

Due to that $\lambda'_1 > 0$, this means that $\mathcal{L}'$ is nonsingular. Then, the following holds:

$$\begin{bmatrix} 1 & b(\mathcal{L}')^{-1} \\ \mathbf{0} & I_N \end{bmatrix} \begin{bmatrix} \sum_{i=1}^N b_i & -b \\ -b^T & \mathcal{L}' \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ (\mathcal{L}')^{-1}b^T & I_N \end{bmatrix} = \begin{bmatrix} b\mathbf{1} - b(\mathcal{L}')^{-1}b^T & \mathbf{0}^T \\ \mathbf{0} & \mathcal{L}' \end{bmatrix}.$$

Since $\mathcal{L}'\mathbf{1} = b^T$, it follows that $b(\mathbf{1} - (\mathcal{L}')^{-1}b^T) = 0$. This implies that $\mathcal{L}_{\mathcal{G}'}$ contains one simple zero eigenvalue. Thus, the graph $\mathcal{G}'$ is connected [88].

Sufficiency: In view of the sufficiency part of Theorem 9.2, we obtain that $\varrho \triangleq \max_{j \in \mathcal{V}} \rho(M'_j(\alpha, \beta)) < 1, \forall (\alpha, \beta) \in \Omega'_c$. It follows from (9.41) that

$$\begin{aligned} \lim_{k \rightarrow \infty} P_j(k) &= \sigma^2 \sum_{k=0}^{\infty} M'_j(\alpha, \beta)^k \text{diag}(0_2, q_j q_j^T) (M'_j(\alpha, \beta)^k)^T \\ &\leq 2h^2 \sigma^2 \sum_{k=0}^{\infty} (M'_j(\alpha, \beta) M'_j(\alpha, \beta)^T)^k, \end{aligned}$$

where the first inequality is due to that $\|\text{diag}(0_2, q_j q_j^T)\| \leq 2h^2$. By Lemma 6.2 and $\varrho < 1$, it is easy to establish that there exists a finite positive number $\varsigma = \varsigma(\varrho)$ such that $\max_{j \in \mathcal{V}} \|\sum_{k=0}^{\infty} (M'_j(\alpha, \beta) M'_j(\alpha, \beta)^T)^k\| \leq \varsigma < \infty$. Because of the unitary matrix Ψ , it is trivial that $\limsup_{k \rightarrow \infty} \mathbb{E}[\|\delta_j(k)\|^2] = \limsup_{k \rightarrow \infty} \mathbb{E}[\|\tilde{\delta}_j(k)\|^2] = \frac{1}{2} \limsup_{k \rightarrow \infty} \mathbb{E}[\|\Delta_j(k)\|^2] = \frac{1}{2} \lim_{k \rightarrow \infty} \text{tr}(P_j(k)) \leq 4h^2 \varsigma \sigma^2$. Thus, the function $g \in \mathcal{K}_{\infty}$ can be selected as $g(\sigma^2) = 8h^2 \varsigma \sigma^2$. Note that ϱ depends on (α, β) and $\lambda'_j, j \in \mathcal{V}$, the proof is completed. $\blacksquare$

9.5 Illustrative Examples

Example 9.1: The first example aims at demonstrating the effectiveness of the

dynamic protocol (9.5). Consider the system (9.1) with

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -2 & 0 & 1 \\ 0.5 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 \\ -0.5 \\ 0.4 \end{bmatrix}, C^T = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

There are four agents with an undirected communication graph $\mathcal{G}$ shown in Fig. 8.2. The adjacency matrix is selected as $(0, 1)$ -weighted symmetric matrix. The eigenvalues of the corresponding Laplacian matrix are $\lambda_2 = 1, \lambda_3 = 3, \lambda_4 = 4$. It is clear that (A, B, C) are stabilizable and detectable. The product of unstable eigenvalue of A is $1.4161 < \frac{1+\lambda_2/\lambda_4}{1-\lambda_2/\lambda_4} = \frac{5}{3}$. Thus, the conditions in Theorem 9.2 are satisfied. It is easy to verify that

$$P = \begin{bmatrix} 8.5661 & -4.4156 & 0.9909 \\ -4.4156 & 5.3489 & -0.7049 \\ 0.9909 & -0.7049 & 1.1388 \end{bmatrix}$$

is a positive definite solution to (9.13) and $K = [0.3844, -0.5275, 0.1434]$. The stabilizing observer gain is selected as $L^T = [-2, -2, -0.5]$. Each element of the initial states of agents are randomly generated from the standard normal distribution. Let the initial controller state be $\hat{\xi}_i(0) = \mathbf{0}, i \in \{1, 2, 3, 4\}$. Denote the j -th element of $\delta_i(k) = x_i(k) - \bar{x}(k)$ by $\delta_i^j(k)$. By using the consensus protocol (9.5), it is shown in Fig. 9.2 that the multi-agent systems asymptotically reach a consensus.

Example 9.2: The second example is to validate the results in Section 9.3 and 8.2.3. Consider a team of four vehicles with an undirected graph $\mathcal{G}$ shown in Fig. 8.2. The adjacency matrix is selected as $(0, 1)$ -weighted symmetric matrix. The configuration variables are initialized as $x(0) = [15, 30, 0, -15]^T$ and $\dot{x}(0) = [4, 4, 2, 3]^T$. Let the sampling interval be $h = 0.25$ s and the control gain of (9.26) be $(\alpha, \beta) = (2.5, -2) \in \Omega_c$. Fig. 9.3 and Fig. 9.4 show that a consensus is reached for all vehicles.

Under the same communication topology as Fig. 8.2, we verify the formation result of the second-order consensus in Section 8.2.3 by using a formation

Figure 9.2: Consensus is asymptotically reached.

example with four agents moving on a plane. The position of agent i at time kh is represented by the pair $[x_i(k), y_i(k)]^T$. Two configuration variables ($x_i(k)$ and $y_i(k)$) are decoupled and have a same dynamic equation as in (9.22). The desired formation is specified as the vertices of a regular parallelogram. To be specific, the formation vectors in the x -axis and y -axis are $f_x = [5, 0, -5, 0]^T$ and $f_y = [0, -10, 0, 10]^T$. All the agents are lined up in a column at the start, e.g., $[x_i(0), y_i(0)]^T = [0, 20 - 5i]^T, i \in \{1, \dots, 4\}$. The velocities in the x -axis and y -axis are respectively initialized as $\dot{x}(0) = [4, 2, 3, 2]^T$ and $\dot{y}(0) = [2, 2, 4, 2]^T$. By adopting the same control gain $(\alpha, \beta) = (2.5, -2) \in \Omega_c$ in (9.31), Fig. 9.5 indicates that a formation is asymptotically produced.

At last, an uncooperative leader is added to the vehicles as shown in Fig. 9.6. The purpose is to demonstrate the proposed formation-based tracking protocol

Figure 9.3: Trajectory of velocity.

Figure 9.4: Trajectory of position.

Figure 9.5: A formation is asymptotically produced.

(9.35). The control gain (α, β) is randomly picked from the feasible set Ω'_c , which is obtained as $(\alpha, \beta) = (3.4217, -3.1106) \in \Omega'_c$ in this example. The initial position and velocity of the leader are set to $[x_0(0), y_0(0)]^T = [0, 15]^T$ and $[\dot{x}_0(0), \dot{y}_0(0)]^T = [3, 2]^T$.

Figure 9.6: Communication graph

Figure 9.7: A formation-based leader-follower consensus problem is solved.

The desired formation of the vehicles is the same as that in the formation example. The difference is that the center of the vehicles is supposed to asymptotically track the leader, which is confirmed by Fig. 9.7 where the control input $u_0(k)$ is uniformly distributed within the interval $[-1, 1]$. For this realization, the tracking errors w.r.t. time are bounded, see Fig. 9.8. By Monte Carlo method with 500 samples, the tracking error variances when $u_0(k)$ is uniformly distributed within $[-1, 1]$ and $[-5, 5]$ are respectively shown in Figures 9.9 and 9.10.

9.6 Summary

Motivated by the situation that agents can only take the relative output feedback for the distributed control design, we have studied the joint effect of the agent dynamics and the communication graph on consensusability under some *admissible* control protocols. The distinct feature of our results lies in the precise quantification of

Figure 9.8: Tracking errors are bounded.

Figure 9.9: $u_0(k)$ is uniformly distributed within $[-1, 1]$: tracking error variances by Monte Carlo method with 500 samples. Figure 9.10: $u_0(k)$ is uniformly distributed within $[-5, 5]$: tracking error variances by Monte Carlo method with 500 samples.

their effects on consensusability. To be exact, the intrinsic entropy rate of the agent dynamics poses a fundamental limitation on the *eigenratio* of the *undirected* graph Laplacian matrix. Some simple control protocols directly using two-step relative position output feedback have been proposed to reach a consensus for the discrete-time double integrator multi-agent systems, which are utilized to solve the formation and formation-based tracking problems. Our theoretical results have been verified by simulations.

Chapter 10

Conclusion and Recommendation for Future Research

10.1 Conclusion

In the past few years, considerable effort has been devoted to the subject of investigating the interactions between communication and control/estimation due to its theoretical and practical significance. This thesis is an addition to this subject. We have addressed the effect of limited data rate and stochastic packet loss on the stabilizability and performance of an NCS from the perspective of unifying control, estimation and communication in the following aspects.

- The logarithmic quantization was shown to be optimal in the sense of approaching the minimum data rate for stabilizing a discrete-time LTI system.
- Data rate theorem for stabilization over lossy channels has been enriched, which is of equal importance in the NCS 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 for single input systems under i.i.d. packet loss model, and for scalar systems under Markovian packet loss model.
- A more insightful characterization of Markovian packet loss on the stability

of Kalman filtering with intermittent observations has been made. The result significantly advances the existing literature, and gives necessary and sufficient conditions for the stability of Kalman filter for second-order and certain classes of high-order systems.

- Motivated by the limited communication bandwidth, a simple recursive quantized innovations Kalman filter has been derived. The quantized filter has the same complexity as the standard Kalman filter and exhibits a comparable filtering performance, even under a moderate number of bit rate.
- The role of the communication graph, system dynamic and limited data rate on consensusability has been revealed. It turns out that the topological entropy rate of agent dynamic poses a fundamental limitation on the graph for consensusability.

We anticipate that the techniques and results developed in the thesis will be helpful in the design and synthesis of an NCS.

10.2 Recommendation for Future Research

Research will continue on the following aspects.

- (1) Performance control via quantized feedback:

Most of the existing results on quantized feedback control are for stabilization only rather than performance control. The transient response is typically very poor due to the lack of good control design algorithms. The study of performance control via finite-level logarithmic quantization is left for our future research.

- (2) Improvement of data rate theorem for stabilization over lossy channels:

The celebrated data rate theorem provides a clear guideline for the minimal requirement on data rate for stabilization of a LTI system. Due to the possible packet loss of the underlying communication network, the development of data rate theorem for stabilization over lossy channels becomes crucial. The

main challenge lies in the fact that the optimal bit assignment is considerably complicated by the random packet losses. Currently, the minimal additional bits required to counter the effect of random packet loss is unavailable for a general LTI system. Thus, the improvement of the existing results on data rate theorem over lossy channels becomes very significant. One critical step is how to find an optimal bit assignment for each state variable by taking into account the random packet losses.

(3) A rigorous error analysis on the quantized innovations Kalman filter:

In Chapter 7, a simple quantized filter is derived under a Gaussian assumption on the predicted density, which strictly speaking, does not hold. This essentially indicates that the modified Riccati recursion can NOT exactly obtain the true error covariance. Although the low computation complexity of this filter is quite attractive, it still lacks a rigorous error analysis. For example, it is still unknown whether the true estimation error is bounded or not in theory. In the future work, we will engage in the error analysis of the proposed quantized filter.

(4) A deeper understanding of packet loss on Kalman filtering:

The existing results on the study of Kalman filtering with intermittent observations mostly focus on the stability analysis. Although some important progresses have been made towards this topic, a complete characterization of the effect of packet loss on stability for general LTI systems is still unknown. Till now, the performance analysis of the state estimation covariance matrices has been rarely touched. A natural question can be raised: to achieve a given filtering performance, what is the minimal requirement on the communication network for transmitting data packet? Apart from the theoretic merit on its own, the answer to this problem also plays an important role in application. It deserves further investigation in the future.

(5) Minimal data rate for consensusability:

In Chapter 8, we have shown that consensusability with perfect state feedback implies consensusability with quantized state feedback, provided that data rate

of each link in the communication graph is not less than an explicitly determined lower bound. Although this result is meaningful, its conservativeness is still unclear. What is of particular interest is to derive the minimal data rate among all possible causal coding for each link to preserve consensusability.

(6) Consensusability over time-varying graphs:

The assumption that the communication graph is fixed over time is somehow restrictive. Changes in the environment, such as the random presence of large metal objects between agents will inevitably affect the propagation properties of the channels. Thus, it is more interesting to consider the scenario that the communication channel is time-varying and unreliable. The investigation of consensusability over time-varying graphs may have far-reaching consequences on the understanding and engineering of networked multi-agent systems. With fixed graphs, consensusability of multi-agent systems under a common control protocol is converted into a simultaneous stabilization problem. However, this key property does not hold in the case of time-varying graphs. Perhaps a completely new method needs to be developed.

(7) Distributed LQR design for networked multi-agent systems:

To the best of our knowledge, the performance control of discrete-time multi-agent systems has not been adequately addressed in the literature, albeit the importance of distributed control is widely recognized. Most recent works devoting to the continuous-time systems include [7, 12, 25, 44, 57, 74]. The design procedure will impose a stringent requirement on simplifying the computational complexity as well as on finding a controller with a distributed architecture. Motivating examples of systems are found in formation flying, vehicle platoons, production units in a power plant and others. Pointed out in [42, 129], the consensusability problem for discrete-time multi-agent systems is considerably harder than its counterpart for continuous-time multi-agent systems. This strongly suggests that the generalization of the existing results to the discrete-time case is far from trivial. In the future, we will consider the distributed LQR design for discrete-time multi-agent systems.

Author's Publications

Book Chapter

- [1] K. You and L. Xie, “Quantized Kalman Filtering of Linear Stochastic Systems”, Editor: Joaquin M. Gomez, Book Chapter in Book Title: “Kalman Filtering”, Nova Science Publishers, 2010

Journal Publications

- [1] K. Hengster-Movric, K. You, F. Lewis and L. Xie, “Synchronization of Discrete-Time Multi-agent Systems on Graphs Using Riccati Design,” submitted to *Automatica* as a *regular paper*, April 2011.
- [2] D. Marelli, K. You, and M. Fu, “Identification of ARMA models using intermittent and quantized output observations,” submitted to *Automatica* as a *regular paper*, April 2011.
- [3] K. You, M. Fu, and L. Xie, “Mean Square Stability for Kalman Filtering with Markovian Packet Losses,” *Automatica*, accepted as a *regular paper*, June 2011.
- [4] K. You and L. Xie, “Network Topology and Communication Data Rate for Consensusability of Discrete-time Multi-agent Systems,” *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2262-2275, 2011(*regular paper*).
- [5] K. You and L. Xie, “Linear Quadratic Gaussian Control with Quantised Innovations Kalman filter over a Symmetric Channel,” *IET Control Theory and applications*, vol. 5, no. 3, pp. 437-446, 2011.
- [6] K. You and L. Xie, “Minimum Data Rate for Mean Square Stabilizability of Linear Systems with Markovian Packet Losses,” *IEEE Transactions on Automatic Control*, vol. 56, no. 4, pp. 772-785, 2011 (*regular paper*).
- [7] K. You and L. Xie, “Minimum Data Rate for Mean Square Stabilization of Discrete LTI Systems over Lossy Channels,” *IEEE Transactions on Automatic Control*, vol. 55, no. 10, pp. 2373–2378, 2010.
- [8] K. You, W. Su, M. Fu, and L. Xie, “Attainability of the Minimum Data Rate for Stabilization of Linear Systems via Logarithmic Quantization,” *Automatica*, vol. 47, no. 1, pp. 170-176, 2011.

- [9] K. You and L. Xie, "Coordination of Discrete-time Multi-agent Systems via Relative Output Feedback," *International Journal of Robust and Nonlinear Control*, first published October 2010.
- [10] K. You, L. Xie, S. Sun, and W. Xiao, "Quantized Filtering of Linear Stochastic Systems," *Transactions of the Institute of Measurement and Control*, first published April 2010.

Conference Publications

- [1] K. You, Z. Li and L. Xie, "Consensus for General Multi-agent Systems over Random Graphs," to appear in the 9th IEEE International Conference on Control & Automation, Santiago, Chile, December, 2011.
- [2] K. You and L. Xie, "Consensusability of Discrete-time Multi-agent Systems over Directed Graphs," Proc. 27th Chinese Control Conference, Yantai, China, 2011.
- [3] D. Marelli, K. You, and M. Fu, "Identification of ARMA models using intermittent and quantized output observations," Proc. 36th International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 2011.
- [4] K. You, M. Fu, and L. Xie, "Necessary and Sufficient Conditions for Stability of Kalman Filtering with Markovian Packet Losses," Proc. 18th World Congress of the International Federation of Automatic Control, Milan, Italy, 2010.
- [5] K. You and L. Xie, "Consensusability of Discrete-time Multi-agent Systems via Relative Output Feedback", Proc. 11th International Conference on Control, Automation, Robotics and Vision, Singapore, 2010.
- [6] K. You and L. Xie, "Necessary and Sufficient Conditions for Consensusability of Discrete-time Multi-agent Systems," Proc. 2nd IFAC Workshop on Distributed Estimation and Control in Networked Systems, Annecy, France, September 2010.
- [7] K. You and L. Xie, "Mean Square Stabilizability of Linear Systems with Limited Feedback Data Rates and Markovian Packet Losses," Proc. 29th Chinese Control Conference, Beijing, China, July 2010 (**Guan Zhao-zhi best paper award**).
- [8] T. Wongpiromsarn, K. You, and L. Xie, "A Consensus Approach to the Assignment Problem: Application to Mobile Sensor Dispatch," Proc. 8th IEEE International Conference on Control & Automation, Xiamen, China, June 2010.
- [9] K. You and Lihua Xie, "Minimum Data Rate for Mean Square Stabilization of Single Input Discrete Systems over Lossy Channels," Proc. 7th IEEE International Conference on Control & Automation, New Zealand, December 2009.

- [10] K. You, W. Su, M. Fu, and L. Xie, "Optimality of the Logarithmic quantizer for stabilization of Linear Systems: achieving the minimum data rate," Proc. 48th IEEE Conference on Decision and Control & 28th Chinese Control Conference, Shanghai, China, December 2009.
- [11] K. You, Y. Zhao, and L. Xie, "Recursive Quantized State Estimation of Discrete-Time Linear Stochastic Systems," Proc. 7th Asian Control Conference, Hong Kong, China, August 2009.
- [12] K. You, L. Xie, S. Sun, and W. Xiao, "Multiple-Level Quantized Innovation Kalman Filter", Proc. 17th International Federation of Automatic Control, Seoul, Korea, July 2008.
- [13] K. You and L. Xie, "LQG control with quantized innovation Kalman Filter," Proc. 27th Chinese Control Conference, Kunming, China, July 2008.

Bibliography

- [1] P. Antsaklis and J. Baillieul, “Guest editorial special issue on networked control systems,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1421–1423, 2004.
- [2] ———, “Special issue on technology of networked control systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 5–8, 2007.
- [3] P. Antsaklis and A. Michel, *Linear Systems*. Birkhauser, 2006.
- [4] R. Ash and C. Doléans-Dade, *Probability and Measure Theory*. Academic Press, 2000.
- [5] J. Baillieul, “Feedback coding for information-based control: operating near the data-rate limit,” *Proc. 41st IEEE Conference on Decision and Control*, 2002.
- [6] ———, “Feedback designs in information-based control,” *Lecture Notes in Control and Information Sciences*, pp. 35–58, 2002.
- [7] B. Bamieh and P. Voulgaris, “A convex characterization of distributed control problems in spatially invariant systems with communication constraints,” *Systems & Control Letters*, vol. 54, no. 6, pp. 575–583, 2005.
- [8] L. Bao, M. Skoglund, and K. Johansson, “Iterative encoder-controller design for feedback control over noisy channels,” *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 265–278, 2011.
- [9] M. Barahona and L. M. Pecora, “Synchronization in small-world systems,” *Physical review letters*, vol. 89, no. 5, p. 054101, 2002.
- [10] V. Borkar and S. Mitter, “LQG Control with Communication Constraints,” in *Communications, Computation, Control and Signal Processing: A Tribute to Thomas Kailath*. Norwell, MA:Kluwer, 1997.
- [11] V. Borkar, S. Mitter, and S. Tatikonda, “Optimal sequential vector quantization of markov sources,” *SIAM Journal on Control and Optimization*, vol. 40, no. 1, pp. 135–148, 2001.

- [12] F. Borrelli and T. Keviczky, "Distributed LQR design for identical dynamically decoupled systems," *IEEE Transactions on Automatic Control*, vol. 53, no. 8, pp. 1901–1912, 2008.
- [13] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Society for Industrial Mathematics, 1994.
- [14] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1279–1289, 2000.
- [15] L. Bushnell, "Networks and control [guest editorial]," *IEEE Control Systems Magazine*, vol. 21, no. 1, pp. 22–23, 2001.
- [16] R. Carli, F. Bullo, and S. Zampieri, "Quantized average consensus via dynamic coding/decoding schemes," *International Journal of Robust and Nonlinear Control*, vol. 20, no. 2, pp. 156–175, 2010.
- [17] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Optimal synchronization for networks of noisy double integrators," *IEEE Transactions on Automatic Control*, vol. 56, no. 5, pp. 1146 – 1152, 2011.
- [18] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, "Communication constraints in the average consensus problem," *Automatica*, vol. 44, no. 3, pp. 671–684, 2008.
- [19] F. Ceragioli and C. De Persis, "Discontinuous stabilization of nonlinear systems: Quantized and switching controls," *Systems & Control Letters*, vol. 56, no. 7, pp. 461–473, 2007.
- [20] C. Chen, *Linear System: Theory and Design*. Saunders College Publishing Philadelphia, PA, USA, 1984.
- [21] J. Cortés and F. Bullo, "Coordination and geometric optimization via distributed dynamical systems," *SIAM Journal on Control and Optimization*, vol. 44, no. 5, pp. 1543–1574, 2006.
- [22] J. Cortés, S. Martinez, T. Karatas, F. Bullo, *et al.*, "Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243–255, 2004.
- [23] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
- [24] R. Curry, *Estimation and Control with Quantized Measurements*. MIT Press, 1970.

- [25] R. D'Andrea and G. Dullerud, "Distributed control design for spatially interconnected systems," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1478–1495, 2003.
- [26] D. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Transactions on Automatic Control*, vol. 35, no. 8, pp. 916–924, 1990.
- [27] Z. Duan, G. Chen, and L. Huang, "Complex network synchronizability: Analysis and control," *Physical Review E*, vol. 76, no. 5, p. 56103, 2007.
- [28] N. Elia, "Remote stabilization over fading channels," *Systems & Control Letters*, vol. 54, no. 3, pp. 237–249, 2005.
- [29] N. Elia and S. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.
- [30] F. Fagnani and S. Zampieri, "Stability analysis and synthesis for scalar linear systems with a quantized feedback," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1569–1584, 2003.
- [31] —, "Quantized stabilization of linear systems: complexity versus performance," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1534–1548, 2004.
- [32] J. Fang and H. Li, "Distributed adaptive quantization for wireless sensor networks: From delta modulation to maximum likelihood," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5246–5257, 2008.
- [33] J. Fax and R. Murray, "Information flow and cooperative control of vehicle formations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1465–1476, 2004.
- [34] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.
- [35] M. Fu, "Linear quadratic Gaussian control with quantized feedback," *American Control Conference*, pp. 2172–2177, 2009.
- [36] M. Fu and L. Xie, "The sector bound approach to quantized feedback control," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1698–1711, 2005.
- [37] —, "Finite-level quantized feedback control for linear systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 5, pp. 1165–1170, 2009.
- [38] —, "Quantized feedback control for linear uncertain systems," *International Journal on Robust and Nonlinear Control*, vol. 20, no. 8, pp. 843–857, 2010.
- [39] H. Gao and T. Chen, "A new approach to quantized feedback control systems," *Automatica*, vol. 44, no. 2, pp. 534–542, 2007.

- [40] C. Godsil and G. Royle, *Algebraic Graph Theory*. Springer New York, 2001.
- [41] G. Gu, “On the existence of linear optimal control with output feedback,” *SIAM Journal on Control and Optimization*, vol. 28, pp. 711–719, 1990.
- [42] G. Gu, L. Marinovici, and F. Lewis, “Consensusability and consensus control of discrete-time multi-agent systems,” *submitted to IEEE Transactions on Automatic Control*, 2011.
- [43] V. Gupta, B. Hassibi, and R. Murray, “Optimal LQG control across packet-dropping links,” *Systems & Control Letters*, vol. 56, no. 6, pp. 439–446, 2007.
- [44] —, “A sub-optimal algorithm to synthesize control laws for a network of dynamic agents,” *International Journal of Control*, vol. 78, no. 16, pp. 1302–1313, 2005.
- [45] V. Gupta, N. Martins, and J. Baras, “Optimal output feedback control using two remote sensors over erasure channels,” *IEEE Transactions on Automatic Control*, vol. 54, no. 7, pp. 1463–1476, July 2009.
- [46] T. Hayakawa, H. Ishii, and K. Tsumura, “Adaptive quantized control for linear uncertain discrete-time systems,” *Automatica*, vol. 45, no. 3, pp. 692–700, 2009.
- [47] J. Hespanha, P. Naghshtabrizi, and Y. Xu, “A survey of recent results in networked control systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007.
- [48] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
- [49] M. Huang and S. Dey, “Dynamic quantizer design for hidden markov state estimation via multiple sensors with fusion center feedback,” *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2887–2896, 2006.
- [50] —, “Dynamic quantization for multisensor estimation over bandlimited fading channels,” *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4696–4702, 2007.
- [51] —, “Stability of Kalman filtering with Markovian packet losses,” *Automatica*, vol. 43, no. 4, pp. 598–607, 2007.
- [52] O. Imer, S. Yüksel, and T. Basar, “Optimal control of LTI systems over unreliable communication links,” *Automatica*, vol. 42, no. 9, pp. 1429–1439, 2006.
- [53] A. Jadbabaie, J. Lin, and A. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.
- [54] R. Kalman, “Nonlinear aspects of sampled-data control systems,” in *Proc. of the Symposium on Nonlinear Circuit Analysis*, vol. 6, 1956, pp. 273–313.

- [55] S. Kar, B. Sinopoli, and J. Moura, “Kalman filtering with intermittent observations: Weak convergence to a stationary distribution,” *accepted by IEEE Transactions on Automatic Control*, 2010.
- [56] R. Karlsson and F. Gustafsson, “Particle filtering for quantized sensor information,” *Proceedings of the 13th European Signal Processing Conference, Antalya, Turkey*, September 2005.
- [57] T. Keviczky, F. Borrelli, and G. Balas, “Decentralized receding horizon control for large scale dynamically decoupled systems,” *Automatica*, vol. 42, no. 12, pp. 2105–2115, 2006.
- [58] J. Kotecha and P. Djuric, “Gaussian particle filtering,” *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2592–2601, 2003.
- [59] G. Lafferriere, A. Williams, J. Caughman, and J. Veerman, “Decentralized control of vehicle formations,” *Systems & Control Letters*, vol. 54, no. 9, pp. 899–910, 2005.
- [60] T. Li, M. Fu, L. Xie, and J. Zhang, “Distributed Consensus with Limited Communication Data Rate,” *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 279–292, 2011.
- [61] T. Li and J. Zhang, “Mean square average-consensus under measurement noises and fixed topologies: Necessary and sufficient conditions,” *Automatica*, vol. 45, no. 8, pp. 1929–1936, 2009.
- [62] —, “Sampled-data based average consensus with measurement noises: Convergence analysis and uncertainty principle,” *Science in China Series F: Information Sciences*, vol. 52, no. 11, pp. 2089–2103, 2009.
- [63] —, “Consensus conditions of multi-agent systems with time-varying topologies and stochastic communication noises,” *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2043–2057, 2010.
- [64] Z. Li, Z. Duan, G. Chen, and L. Huang, “Consensus of multiagent systems and synchronization of complex networks: a unified viewpoint,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 1, pp. 213–224, 2010.
- [65] P. Lin and Y. Jia, “Consensus of second-order discrete-time multi-agent systems with nonuniform time-delays and dynamically changing topologies,” *Automatica*, vol. 45, no. 9, pp. 2154–2158, 2009.
- [66] J. Liu and N. Elia, “Quantized feedback stabilization of non-linear affine systems,” *International Journal of Control*, vol. 77, no. 3, pp. 239–249, 2004.

- [67] Z. Luo, "An isotropic universal decentralized estimation scheme for a bandwith constrained ad hoc sensor network," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 735–744, 2005.
- [68] —, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2210–2219, 2005.
- [69] N. Lynch, *Distributed Algorithms*. Morgan Kaufmann, 1996.
- [70] C. Ma, "System analysis and control synthesis of linear-multi-agent systems," P.h.d. dissertation, Institute of Systems Science, Chinese Academy of Sciences, China, May 2009.
- [71] C. Ma and J. Zhang, "Necessary and sufficient conditions for consensusability of linear multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 5, pp. 1263–1268, 2010.
- [72] D. Marelli, K. You, and M. Fu, "Identification of ARMA models using intermittent and quantized output observations," *submitted to Automatica*, April 2011.
- [73] N. Martins, M. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 438–447, 2006.
- [74] P. Massioni and M. Verhaegen, "Distributed control for identical dynamically coupled systems: A decomposition approach," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 124–135, 2009.
- [75] A. Matveev and A. Savkin, "Comments on "Control over noisy channels" and relevant negative results," *IEEE Transactions on Automatic Control*, vol. 50, no. 12, pp. 2105–2110, 2005.
- [76] —, "An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels," *SIAM Journal on Control and Optimization*, vol. 46, pp. 1323–1367, 2007.
- [77] S. Meyn, R. Tweedie, and J. Hiley, *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1996.
- [78] P. Minero, M. Franceschetti, S. Dey, and G. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 243–255, 2009.
- [79] Y. Mo and B. Sinopoli, "Towards finding the critical value for Kalman filtering with intermittent observations," <http://arxiv.org/abs/1005.2442>, 2010.

- [80] L. Moreau, “Stability of multiagent systems with time-dependent communication links,” *IEEE Transactions on Automatic Control*, vol. 50, no. 2, pp. 169–182, 2005.
- [81] E. Msechu, S. Roumeliotis, A. Ribeiro, and G. Giannakis, “Decentralized quantized Kalman filtering with scalable communication cost,” *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3727–3741, 2008.
- [82] G. Nair and R. Evans, “Exponential stabilisability of finite-dimensional linear systems with limited data rates,” *Automatica*, vol. 39, no. 4, pp. 585–593, 2003.
- [83] G. Nair, F. Fagnani, S. Zampieri, and R. Evans, “Feedback control under data rate constraints: an overview,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 108–137, 2007.
- [84] G. Nair and R. Evans, “Stabilizability of stochastic linear systems with finite feedback data rates,” *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.
- [85] A. Okubo, “Dynamical aspects of animal grouping: swarms, schools, flocks, and herds,” *Advances in Biophysics*, vol. 22, p. 1, 1986.
- [86] R. Olfati-Saber, J. Fax, and R. Murray, “Consensus and cooperation in networked multi-agent systems,” *Proceedings of the IEEE*, vol. 95, no. 1, 2007.
- [87] R. Olfati-Saber and R. Murray, “Flocking with obstacle avoidance: Cooperation with limited communication in mobile networks,” *Proc. 42nd IEEE Conference on Decision and Control*, 2003.
- [88] —, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.
- [89] G. Pagès and H. Pham, “Optimal quantization methods for nonlinear filtering with discrete-time observations,” *Bernoulli*, vol. 11, no. 5, pp. 893–932, 2005.
- [90] H. Papadopoulos, G. Wornell, and A. Oppenheim, “Sequential signal encoding from noisy measurements using quantizers with dynamic bias control,” *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 978–1002, 2002.
- [91] K. Plarre and F. Bullo, “On Kalman filtering for detectable systems with intermittent observations,” *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 386–390, 2009.
- [92] W. Ren and E. Atkins, “Distributed multi-vehicle coordinated control via local information exchange,” *International Journal of Robust and Nonlinear Control*, vol. 17, no. 10-11, pp. 1002–1033, 2007.

- [93] W. Ren and R. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 655–661, 2005.
- [94] W. Ren, K. Moore, and Y. Chen, "High-order consensus algorithms in cooperative vehicle systems," *Proc. IEEE International Conference on Networking, Sensing and Control (ICNSC)*, pp. 457–462, 2006.
- [95] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-part I: Gaussian case," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1131–1143, 2006.
- [96] —, "Bandwidth-constrained distributed estimation for wireless sensor networks-part II: unknown pdf," *IEEE Transactions on Signal Processing*, vol. 54, no. 7, pp. 2784–2796, 2006.
- [97] A. Ribeiro, G. Giannakis, and S. Roumeliotis, "SOI-KF: Distributed Kalman filtering with low-cost communications using the sign of innovations," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4782–4795, 2006.
- [98] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for control over a noisy communication link: Part I: Scalar Systems," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3369–3395, 2006.
- [99] L. Scardovi and R. Sepulchre, "Synchronization in networks of identical linear systems," *Automatica*, vol. 45, no. 11, pp. 2557–2562, 2009.
- [100] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. Sastry, "Foundations of control and estimation over lossy networks," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, 2007.
- [101] L. Shi, M. Epstein, and R. Murray, "Kalman filtering over a packet-dropping network: A probabilistic perspective," *IEEE Transactions on Automatic Control*, vol. 55, no. 3, pp. 594–604, 2010.
- [102] R. Singer, "Estimating optimal filter tracking performance for manned maneuvering targets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 6, no. 4, pp. 473–483, 1970.
- [103] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, and S. Sastry, "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.
- [104] V. Solo, "One step ahead adaptive controller with slowly time-varying parameters," Dept. ECE, John Hopkins University, Baltimore, Tech. Rep., 1991.
- [105] F. Sorrentino, M. Di Bernardo, and F. Garofalo, "Synchronizability and synchronization dynamics of weighed and unweighed scale free networks with degree mixing," *International Journal of Bifurcation and Chaos*, vol. 17, no. 7, pp. 2419–2434, 2007.

- [106] S. Tatikonda and S. Mitter, “Control over noisy channels,” *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1196–1201, 2004.
- [107] —, “Control under communication constraints,” *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, 2004.
- [108] S. Tatikonda, A. Sahai, and S. Mitter, “Stochastic linear control over a communication channel,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1549–1561, 2004.
- [109] K. Tsumura, H. Ishii, and H. Hoshina, “Tradeoffs between quantization and packet loss in networked control of linear systems,” *Automatica*, vol. 45, no. 12, pp. 2963–2970, 2009.
- [110] S. Tuna, “Synchronizing linear systems via partial-state coupling,” *Automatica*, vol. 44, no. 8, pp. 2179–2184, 2008.
- [111] L. Wang, G. Yin, J. Zhang, and Y. Zhao, *System identification with quantized observations*. Birkhäuser Boston, 2010.
- [112] B. Widrow, I. Kollar, and M. Liu, “Statistical theory of quantization,” *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353–361, 1996.
- [113] P. Wieland, J. Kim, H. Scheu, and F. Allgäwer, “On consensus in multi-agent systems with linear high-order agents,” *Proc. 17th IFAC World Congress, Seoul, Korea*, pp. 1541–1546, 2008.
- [114] D. Williamson, “Finite wordlength design of digital Kalman filters for state estimation,” *IEEE Transactions on Automatic Control*, vol. 30, no. 10, pp. 930–939, 1985.
- [115] W. Wong and R. Brockett, “Systems with finite communication bandwidth constraints-Part I: State estimation problems,” *IEEE Transactions on Automatic Control*, vol. 42, no. 9, pp. 1294–1299, 1997.
- [116] —, “Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback,” *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 1049–1053, 1999.
- [117] J. Xiao, A. Ribeiro, Z. Luo, and G. Giannakis, “Distributed compression-estimation using wireless sensor networks,” *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 27–41, 2006.
- [118] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.
- [119] L. Xie and L. Xie, “Stability of a random Riccati equation with Markovian binary switching,” *IEEE Transactions on Automatic Control*, vol. 53, no. 7, pp. 1759–1764, 2008.

- [120] —, “Stability analysis of networked sampled-data linear systems with markovian packet losses,” *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1368–1374, 2009.
- [121] P. Yang, R. Freeman, and K. Lynch, “Multi-agent coordination by decentralized estimation and control,” *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2480–2496, 2008.
- [122] K. You, M. Fu, and L. Xie, “Mean square stability for Kalman filtering with markovian packet losses,” *Automatica(accepted)*, June 2011.
- [123] K. You, W. Su, M. Fu, and L. Xie, “Attainability of the minimum data rate for stabilization of linear systems via logarithmic quantization,” *Automatica*, vol. 47, no. 1, pp. 170–176, 2011.
- [124] K. You and L. Xie, “Minimum data rate for mean square stabilization of discrete LTI systems over lossy channels,” *IEEE Transactions on Automatic Control*, vol. 55, no. 10, pp. 2373–2378, 2010.
- [125] —, “Consensusability of discrete-time multi-agent systems over directed graphs,” *Proc. 30th Chinese Control Conference, Yantai, China*, July 2011.
- [126] —, “Coordination of discrete-time multi-agent systems via relative output feedback,” *International Journal of Robust and Nonlinear Control*, vol. 21, no. 13, pp. 1587–1605, 2011.
- [127] —, “Linear quadratic gaussian control with quantised innovations kalman filter over a symmetric channel,” *IET Control Theory & Applications*, vol. 5, no. 3, pp. 437–446, 2011.
- [128] —, “Minimum data rate for mean square stabilizability of linear systems with Markovian packet losses,” *IEEE Transactions on Automatic Control*, vol. 56, no. 4, pp. 772–785, 2011.
- [129] —, “Network topology and communication data rate for consensusability of discrete-time multi-agent systems,” *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2262–2275, 2011.
- [130] K. You, L. Xie, S. Sun, and W. Xiao, “Multiple-level quantized innovation Kalman filter,” *Proc. 17th IFAC World Congress*, pp. 1420–1425, 2008.
- [131] S. Yuksel and T. Basar, “Minimum rate coding for LTI systems over noiseless channels,” *IEEE Transactions on Automatic Control*, vol. 51, no. 12, pp. 1878–1887, 2006.
- [132] W. Zhang, M. Branicky, and S. Phillips, “Stability of networked control systems,” *IEEE Control Systems Magazine*, vol. 21, no. 1, pp. 84–99, 2001.