Quantisation Issues in Feedback Control

Hernan Haimovich

A thesis submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering
and Computer Science

The University of Newcastle
Callaghan, NSW 2308
Australia

March, 2006

The UNIVERSITY of NEWCASTLE
AUSTRALIA

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.

Hernan Haimovich

Acknowledgements

First, I would like to deeply thank my supervisors, Prof. Graham Goodwin and Dr. María Seron. The process of doing this PhD would have been extremely difficult had it not been for Graham's enthusiastic encouragement and María's wisdom and in-depth ever-present help. I am also grateful to them for providing both financial support and a great working environment during my studies.

I would like to acknowledge Prof. Ernesto Kofman for providing the seed for one of the two topics of this thesis. Very special thanks go to Dr. Julio Braslavsky and Prof. Arie Feuer for the many interesting technical discussions.

For reading and providing comments on different parts of this thesis, I would like to express my gratitude to Juan Carlos Agüero, José Mare, Elena Prieto, and Daniel Quevedo.

I would also like to thank Jacqui Ramagge for encouraging and helping me to learn more advanced mathematical topics. Thanks as well to the other members of our maths and flatness study group: Udo Baumgartner, Aidan Sims, José De Doná and María Seron.

Life has been so much easier here at uni due to the always keen and helpful, superheroes-in-secretary-disguise, Dianne Piefke and Jayne Disney. Thanks very much, Dianne and Jayne. Thanks also to Frank Sobora for his help with any computer-related problem.

I am truly grateful to the many academics back in Argentina who, in spite of the many difficulties and the lack of resources, have always worked extremely hard to give their best to me (and to all their students). In particular, thanks very much Professors Sergio Junco, Juan Carlos Gómez and Federico Miyara.

Coming to Australia has been an incredible experience, especially because of the great people that I have met. Thanks Tristan and Jae, Osvaldo and Olga, Juan Carlos, James and Patty, Alejandro, Boris, José, Claudio and Claudia, Elena and Rhyall, Juan and Paz, José and Sandra, José and Patricia, Mario, Daniela, Milan and Erica, Eduardo and Andrea, and Manolo. Very special thanks to Julio, Marimar and

Ruth for sharing so much with me: playing music and dancing together, making me feel at home at their place, going on holidays together, and so much more. Thanks a lot, guys.

I am greatly indebted to my parents, Beatriz and Jorge, for always giving me their love and support. Thanks to my sister, Alejandra, for being such a great friend, and to my grandparents Fany and Abraham, and Raquel. Thanks also to my parents-in-law, Stella and Carlos, for sharing so many memorable moments during my PhD studies.

And last but not least at all, thanks to my lovely and always joyful wife Silvana Abecasis for her constant support, her beautiful smile, and for sharing her life with me. Thanks also, Sil, for coming to Australia with me, an endeavour which, I know, was not easy for you in the beginning.

To Silvana

Contents

Abstract

Systems involving quantisation arise in many areas of engineering, especially when digital implementations are involved. In this thesis we consider different aspects of quantisation in feedback control systems. We study two topics of interest: (a) quantisers that quadratically stabilise a given system and are efficient in the use of their quantisation levels and (b) the derivation of ultimate bounds for perturbed systems, especially when the perturbations arise from the use of quantisers.

In the first part of the thesis we address problem (a) above. We consider quadratic stabilisation of discrete-time multiple-input systems by means of quantised static feedback and we measure the efficiency of a quantiser via the concept of quantisation density. Intuitively, the lower the density of a quantiser is, the more separated its quantisation levels are. We thus deal with the problem of optimising density over all quantisers that quadratically stabilise a given system with respect to a given control Lyapunov function. Most of the available results on this problem treat single-input systems, and the ones that deal with the multiple-input case consider only two-input systems. In this thesis, we derive several new results for multiple-input systems and also provide an alternative approach to deal with the single-input case. Our new results for multiple-input systems include the derivation of the structure of optimal quantisers and the explicit design of multivariable quantisers with finite density that are able to quadratically stabilise systems having an arbitrary number of inputs. For single-input systems, we provide an alternative approach to the analysis and design of optimal quantisers by establishing a link between the separation of the quantisation levels of a quantiser and the size of its quantisation regions.

In the second part of the thesis we address problem (b) above. In the presence of perturbations, asymptotic stabilisation may not be possible. However, there may exist a bounded region that contains the equilibrium point and has the property that the system trajectories converge to this bounded region. When this bounded region exists, we say that the system trajectories are ultimately bounded, and that this bounded region is an ultimate bound for the system. The size of the ultimate bound quantifies the performance of the system in steady state. Hence, it is important to derive ultimate bounds that are as tight as possible. This part of the thesis addresses the problem of ultimate bound computation in settings involving several scalar quantisers, each having different features. We consider each quantised variable in the system to be a perturbed copy of the corresponding unquantised variable. This turns the original

quantised system into a perturbed system, where the perturbation has a natural *componentwise* bound. Moreover, according to the type of quantiser employed, the perturbation bound may depend on the system state. Typical methods to estimate ultimate bounds are based on the use of Lyapunov functions and usually require a bound on the norm of the perturbation. Applying these methods in the setting considered here may disregard important information on the structure of the perturbation bound. We therefore derive ultimate bounds on the system states that explicitly take account of the componentwise structure of the perturbation bound. The ultimate bounds derived also have a componentwise form, and can be systematically computed without having to, for example, select a suitable Lyapunov function for the system. The results of this part of the thesis, though motivated by quantised systems, apply to more general perturbations, not necessarily arising from quantisation.

Notation

$\triangleq$	Definition.
$\|\cdot\|_2$	Euclidean norm of a vector and corresponding induced norm of a matrix.
$\|\cdot\|_\infty$	Infinity norm of a vector and corresponding induced norm of a matrix.
$\uplus$	Disjoint union.
$\#$	Number of elements (cardinality) of a set.
$\eta(q)$	Quantisation density of a quantiser q.
$\lambda_{\max}(\cdot)$	Maximum eigenvalue of a real symmetric matrix.
$\lambda_{\min}(\cdot)$	Minimum eigenvalue of a real symmetric matrix.
$\boldsymbol{\rho}(M)$	Spectral radius of the square matrix M, that is, maximum over the magnitude of the eigenvalues of M.
$\mathbf{0}_{n\times m}$	$n \times m$ matrix with zero entries.
$\mathbf{1}_n$	n-dimensional column vector with all components equal to 1.
$\lfloor b \rfloor$	Greatest integer not greater than b.
$\lceil b \rceil$	Least integer not less than b.
$\mathbb{C}$	Set of complex numbers.
CAQS	Coarse Almost Quadratically Stabilising.
CLF	Control Lyapunov Function.
$\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$	Diagonal matrix with main-diagonal entries $\lambda_1, \ldots, \lambda_n$.
I_n	$n \times n$ identity matrix.
$\mathrm{Im}(M)$	Range of a matrix M.
LTI	Linear Time-invariant.
$\max\{x, y\}$	Componentwise maximum of vectors x and y.
$\|M\|$	Elementwise magnitude of a matrix M with (possibly) complex entries. That is, if M has entries $M_{i,j}$, then $\|M\|$ is the matrix with entries $\|M_{i,j}\|$.

$M \prec N$	Set of componentwise inequalities $M_{i,j} < N_{i,j}$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, when $M, N \in \mathbb{R}^{m \times n}$.
$M \preceq N$	Set of componentwise inequalities $M_{i,j} \leq N_{i,j}$, for $i = 1, \ldots, m$, $j = 1, \ldots, n$, when $M, N \in \mathbb{R}^{m \times n}$.
$M \succ N$	Set of componentwise inequalities $M_{i,j} > N_{i,j}$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, when $M, N \in \mathbb{R}^{m \times n}$.
$M \succeq N$	Set of componentwise inequalities $M_{i,j} \geq N_{i,j}$, for $i = 1, \ldots, m$, $j = 1, \ldots, n$, when $M, N \in \mathbb{R}^{m \times n}$.
$n \bmod N$	Remainder of dividing n by N ($n \in \mathbb{Z}_{+,0}$, $N \in \mathbb{Z}_+$).
QS	Quadratically Stabilising.
$q, \tilde{q}, \bar{q}, \mathring{q}$	Quantisers.
$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^n_+$	Set of vectors in $\mathbb{R}^n$ with positive components.
$\mathbb{R}^n_{+,0}$	Set of vectors in $\mathbb{R}^n$ with nonnegative components.
$\mathbb{R}^{n \times m}_+$	Set of matrices in $\mathbb{R}^{n \times m}$ with positive entries.
$\mathbb{R}^{n \times m}_{+,0}$	Set of matrices in $\mathbb{R}^{n \times m}$ with nonnegative entries.
$\mathbb{R}\mathrm{e}(M)$	Elementwise real part of a matrix M. That is, if M has entries $M_{i,j}$, then $\mathbb{R}\mathrm{e}(M)$ is the matrix with entries $\mathbb{R}\mathrm{e}(M_{i,j})$.
^{T}	Transpose.
T^k	Iteration of a map $T : \mathbb{R}^n \to \mathbb{R}^n$, $T^k(x) = T(T^{k-1}(x))$, for $k = 1, 2, \ldots$, and $T^0(x) \triangleq x$.
$\mathcal{U}(q)$	Range of a quantiser q.
$V(\cdot)$	Control Lyapunov function.
$\mathbb{Z}$	Integers.
$\mathbb{Z}_+$	Positive integers.
$\mathbb{Z}_{+,0}$	Nonnegative integers.

Chapter 1

Introduction

1.1 Quantisation in Feedback Control

The term "quantisation" refers to the restriction of a variable to a discrete set of values rather than a continuous set of values. There are several reasons why quantisation needs to be considered in feedback control systems. For example, since controllers are usually implemented digitally, signals that take values in a continuous set need to be represented with finite precision to allow digital information processing in finite time (Åström and Wittenmark, 1997). In addition, sensors may produce an output that indicates only whether the value of the measured signal lies within some range. The number of different ranges that the sensor can distinguish may be severely limited in some cases. A prime example of the latter situation is provided by the exhaust gas oxygen sensor used for air-to-fuel ratio control in automotive systems (see, for example, Grizzle et al., 1991). Further motivation for considering quantisation in feedback control systems is the recent boom of interest in networked control systems (Raji, 1994; Zhang et al., 2001; Walsh and Ye, 2001). These systems are characterised by the fact that controllers and plants are interconnected over a digital communication network, making quantisation essential in order to transmit information among different parts of the system.

The analysis of the effects of quantisation in feedback control systems began as early as in the 1950s, as evidenced by the work of Kalman (1956). Kalman's work was aimed at studying the effects of nonlinearities on sampled-data systems. Although the words *quantiser* or *quantisation* do not explicitly appear in this work, the effect of including an ideal relay —one of the simplest forms of quantiser— in a sampled-data system was analysed. Also during the 1950s, the words *quantisation* and *quantised* appeared in the control systems literature (Flügge-Lotz and Taylor, 1956; Bertram, 1958). The need to analyse the effects of quantisation on control systems stemmed from the fact that controllers could be digitally implemented. Indeed, in a typical digital control scheme, quantisation arises due to the use of analog-to-digital (A/D) and digital-to-analog (D/A) converters, and because the calculations performed

by the digital processor have round-off errors.

The way in which quantisation is dealt with in feedback control systems has experienced a striking change in recent years. Traditionally, quantisation in control systems was regarded as an undesirable phenomenon. The most common approach to designing a digital controller was to disregard quantisation in a first stage, when the controller was designed. The impact of quantisation on the resulting performance was later mitigated by utilising A/D and D/A converters, and digital processors, with suitably high precision. Naturally, however, even small quantisation errors can have a negative impact on achievable performance. Different methods exist for estimating the deleterious effect of quantisation on a digital control system that is designed ignoring the presence of quantisation (Bertram, 1958; Slaughter, 1964; Yakowitz and Parker, 1973; Green and Turner, 1988; Miller et al., 1988; Farrell and Michel, 1989; Miller et al., 1989). Almost all of these methods regard a quantised variable as a perturbed copy of the unquantised variable. The effect of quantisation on the resulting performance is then analysed utilising a bound on the perturbation introduced by the quantiser. This approach is well justified when the required precision is easily achievable and the cost of the resulting implementation is reasonable.

A different approach is to regard a quantised variable as a partial observation of the unquantised variable. This means that a quantised variable provides information on a range of values that the unquantised variable may take, rather than one specific value. Relevant works that address quantisation in this manner in the context of feedback control systems are Curry (1970) and Delchamps (1990). These works have paved the way for the most recent approach to quantisation, which consists in viewing a quantiser as an *information coder*. This new approach has led to a paradigm change regarding quantisation in a feedback control system: from undesirable phenomenon to intrinsic and inescapable system component.

Indeed, in recent years, different control schemes have been proposed and analysed, which explicitly account for the fact that controller and plant(s) are connected via a communication channel (see the special issue Antsaklis and Baillieul, Guest Eds., 2004, and the references therein). The new challenges that arise from the introduction of a communication channel between controller and plant(s) are numerous. These challenges include the need to explicitly deal with variable time delays, nonuniform sampling, limited data-rate/bandwidth, data loss, and quantisation.

There has been substantial research effort directed at various aspects of the above factors. Several lines of research exist which address different groups of such issues. In particular, numerous works explicitly deal with quantisation while focusing on stabilisation in a networked control setting. Within these works, we can distinguish between the ones where the quantisation strategy is dynamic and time-varying (for example, Wong and Brockett, 1999; Brockett and Liberzon, 2000; Liberzon, 2003a,b; Liberzon and Hespanha, 2005; Nair and Evans, 2003, 2004; Li and Baillieul, 2004; Tatikonda and Mitter, 2004a,b; Tatikonda and Elia, 2004) and where it is fixed and static (for example, Elia and Mitter, 2001; Elia and Frazzoli, 2002; Kao and Venkatesh, 2002; Fu and Xie, 2003, 2005; Baillieul, 2002; Ishii

and Francis, 2002b, 2003; Ishii and Başar, 2005; Goodwin et al., 2004).

If the quantisation strategy is dynamic and time-varying, then quantisers having a finite number of levels may be employed to achieve asymptotic stabilisation. On the other hand, if the quantisation strategy is fixed and static, employing a quantiser with a finite number of levels can only yield local practical stability. Asymptotic stability can be achieved by utilising quantisers with a countably infinite number of levels, having increasingly higher precision towards the origin (such as logarithmic quantisers). Quantisers with a finite number of levels have greater practical significance than quantisers with an infinite number of levels. However, the latter quantisers have been very useful for proving many important results in networked control.

Throughout this thesis, we will regard a quantiser as a fixed and static component of the system and will analyse different aspects of quantisation in feedback control systems. Specifically, we will deal with the following two topics: (a) quadratic stabilisation by means of quantised static feedback and (b) the derivation of ultimate bounds for perturbed systems, especially when the perturbations arise due to the use of quantisers.

1.1.1 Quadratic Stabilisation via Quantised Static Feedback

In Part I of the thesis, we deal with quadratic stabilisation of discrete-time linear systems by means of static feedback employing quantisers. The approach that we follow is related to the work of Elia and Mitter (2001); Elia and Frazzoli (2002); Elia (2002); Kao and Venkatesh (2002); Fu and Xie (2003, 2005). Elia and Mitter (2001) introduce a measure of density of quantisation. Intuitively, the density of a quantiser is lower than that of another quantiser if the values of the former are more separated than those of the latter. In this sense, a quantiser can be regarded as being more efficient in the use of its quantisation levels if its density is lower. In this context, an important question that is posed and answered in Elia and Mitter (2001) is: for a linear single-input system, what is the most efficient quantiser over all quadratically stabilising quantisers?

The interesting results of Elia and Mitter (2001) apply only to single-input systems. Generalising these results to multiple-input systems is recognised as an extremely difficult task. Indeed, for multiple-input systems, the quantisation density problem introduced in Elia and Mitter (2001) still remains largely open. Elia and Frazzoli (2002) and Elia (2002) provide lower bounds on the infimum quantisation density for two-input systems. Kao and Venkatesh (2002) analyse different quantisation schemes and their densities for linear multiple-input systems. However, explicit design of a multivariable quantiser with finite (though not necessarily infimum) quantisation density is performed only when quadratic stabilisation is possible through the use of a two-dimensional subspace of the input space.

The works of Fu and Xie (2003, 2005) employ a completely different approach to deal with the optimisation of quantisation density. These authors model a logarithmic quantiser as a nonlinearity

bounded by a sector. The system is then regarded as an uncertain system with sector-bound uncertainty and the problem is posed as a robust control problem. Their main finding is that, for single-input systems, this approach is not conservative. That is, the results of Elia and Mitter (2001) can be recovered by means of this approach. To deal with multiple-input systems, Fu and Xie utilise independent scalar quantisers for each input signal. Independently quantising the different input signals, however, leads to designs with infinite quantisation density.

Our contribution to this problem will be to give several new results related to quadratic stabilisation of single- and multiple-input systems and to quantisation density. These results are presented in Part I of the thesis and include the derivation of the structure of optimal quantisers and the explicit design of multivariable quantisers with finite density that are able to quadratically stabilise systems having an arbitrary number of inputs.

1.1.2 Componentwise Ultimate Bounds for Perturbed Systems

In Part II of the thesis, we deal with the derivation of componentwise ultimate bounds for continuous-time, discrete-time and sampled-data perturbed systems, especially when the perturbations arise due to the use of quantisers.

Quantisation in digital control systems arises due to the use of A/D, D/A and digital processors with finite precision. Since a digital control system is usually designed ignoring the presence of quantisation, it is necessary to estimate the effect that quantisation has on the resulting practical implementation. This effect can be quantified by means of bounds on the difference between the desired and the actual system behaviour. In particular, it is of interest to obtain ultimate bounds on the system variables (Yakowitz and Parker, 1973; Green and Turner, 1988; Miller et al., 1988; Farrell and Michel, 1989; Miller et al., 1989).

More recently, motivated by networked control systems, several control schemes have been considered that involve static memoryless quantisers (see for example, Elia and Mitter, 2001; Ishii and Francis, 2002b, 2003; Ishii et al., 2004; Ishii and Başar, 2005; Fu and Hara, 2005). Most of these works deal with the design of quantised control strategies to achieve different objectives, and utilise all the information provided by a quantised variable. However, some aspects of the resulting schemes can also be analysed by regarding a quantised variable as a perturbed copy of the corresponding unquantised variable. In particular, ultimate bounds on the system variables may be obtained in this manner when asymptotic stability is not possible.

Regarding a quantised variable as a perturbed copy of the unquantised variable turns a quantised control system into a perturbed system. The most general and powerful tool to analyse ultimate bounds in perturbed systems is the use of Lyapunov functions (see, for example, Khalil, 2002). This approach has the inherent difficulty of finding a suitable Lyapunov function. For linear systems, however, quadratic Lyapunov functions can be easily computed. Kofman (2005) proposes a different method to estimate

ultimate bounds for linear continuous-time perturbed systems with constant perturbation bounds. The method is based on the analysis of the system in modal coordinates and gives componentwise ultimate bounds on the system state. An example is given where the suggested method yields ultimate bounds that are substantially tighter than those derived by means of quadratic Lyapunov functions. The method of Kofman can be regarded as the continuous-time counterpart to the earlier method of Yakowitz and Parker (1973).

In Part II of the thesis, we will bring together the above earlier ideas and more recent work. In particular, motivated by the results of Yakowitz and Parker (1973) and Kofman (2005), we develop systematic methods to obtain componentwise ultimate bounds in continuous-time, discrete-time and sampled-data perturbed systems, especially when the perturbations arise due to the use of quantisers. We allow for different types of quantisers: uniform, logarithmic and semitruncated logarithmic. Our developments require several extensions of the methods of Yakowitz and Parker and Kofman. We also show how our methods can be applied to a class of nonlinear systems. The main features of our methods are their systematic nature and their flexibility in dealing with highly structured perturbation schemes. This latter feature allows us to deal with systems involving many different quantisers in the same setting.

1.2 Thesis Overview

The contents of the thesis are presented in 6 core chapters, which have been organised into two parts: the first part deals with stabilisation by means of quantisers, and the second part addresses the derivation of ultimate bounds in the presence of quantisation. A final chapter presents a summary and conclusions.

Part I (Chapters 2 to 5) addresses quantisation density in the context of quadratic stabilisation of discrete-time systems by means of static feedback utilising quantisers. The works most related to this part of the thesis are Elia and Mitter (2001); Elia and Frazzoli (2002); Elia (2002); Kao and Venkatesh (2002); Fu and Xie (2003, 2005). A more detailed description of the various chapters follows.

In Chapter 2, we first briefly review quadratic stabilisation of linear discrete-time systems and then focus on quantisation density in the context of multiple-input systems. We generalise the definition of quantisation density of Elia and Mitter (2001) to multiple-input systems and derive several new results regarding quantisation density. We also pose the problem of optimising quantisation density over all quantisers that quadratically stabilise a given multiple-input system and derive an important result that reveals the structure of a quantiser that optimises density. The different results of this chapter are employed in the remaining chapters of Part I of the thesis.

In Chapter 3, we focus on the characterisation of quantisers that quadratically stabilise a given multiple-input system. As a first step toward this characterisation, we consider quantisers having a form that can be interpreted as the simplest possible in some appropriate sense. We derive necessary and sufficient conditions for these quantisers to quadratically stabilise a system, and we do this by

means of explicit geometric considerations. We thus develop a novel geometric approach to quadratic stabilisation of multiple-input systems by means of quantisers. The geometric approach derived in this chapter will provide the framework for results derived in subsequent chapters. We also employ this geometric approach to design quantisers with finite density that can stabilise multiple-input systems having an arbitrary number of inputs.

In Chapter 4, we deal with single-input systems. For these systems, we enhance the geometric approach of Chapter 3 to explore quantiser coarseness from a state-space standpoint, as opposed to the standard input-space-based concept of quantisation density. We introduce a novel type of quantisers, namely CAQS (Coarse-Almost-Quadratically-Stabilising) quantisers, and analyse the relationships between CAQS quantisers and quantisers that minimise quantisation density in the standard sense. We also show how to directly utilise CAQS quantisers to design static output feedback strategies that employ quantisers of infimum density. We conclude this chapter by showing how to recover a well-known result on infimum quantisation density by means of our approach.

In Chapter 5, we solve a special case of infimum quantisation density problem for multiple-input systems. Specifically, we derive the infimum density over all quantisers that quadratically stabilise the system and have levels in a one-dimensional subspace of the input space. We also show that our result conflicts with a previously published result, and we provide a counterexample to the latter result.

Part II (Chapters 6 and 7) addresses the derivation of ultimate bounds in systems involving quantisation. Throughout this part, a quantised variable will be regarded as a perturbed copy of the corresponding unquantised variable.

In Chapter 6, we derive componentwise ultimate bound expressions for discrete-time and sampled-data perturbed systems, especially when the perturbations arise due to quantisation. A very important feature of our results is that they can directly accommodate feedback schemes where quantisers of different characteristics and/or types affect different signals in the same system. We demonstrate the applicability and potential of the method by means of an example taken from recent literature on the topic of control over communication networks.

In Chapter 7, we extend the results of Chapter 6 to deal with perturbed systems where the perturbation bounds have more general forms. In this case, we focus on continuous- and discrete-time perturbed systems. Since the perturbations are allowed to be bounded by state-dependent functions, the method can then be applied to nonlinear systems by regarding them as perturbed linear systems.

In Chapter 8, we summarize and give suggestions for future work.

1.3 Thesis Contributions

The main contributions of the thesis are believed to be:

Chapter 2. We derive an expression for the quantisation density of multivariable quantisers having radially logarithmically spaced levels (Theorem 2.12). We establish the invariance of the density of a quantiser under a linear one-to-one transformation (Lemma 2.15). We also derive results (Theorem 2.17 and Theorem 2.19) that provide insight into the structure of quantisers that optimise density for a multiple-input system.

Chapter 3. We give a geometric interpretation to the fact that a quantiser quadratically stabilises the system. We derive necessary and sufficient conditions for a quantised feedback having levels in a minimum-dimensional subspace to quadratically stabilise a given multiple-input system (Theorem 3.14 and Theorem 3.17). We also explicitly design quantisers having finite quantisation density that are able to quadratically stabilise a system having an arbitrary number of inputs (Theorem 3.22).

Chapter 4. For single-input systems, we explore quantiser coarseness from a state-space standpoint. We introduce and characterise CAQS quantisers (Theorem 4.16) and analyse the connections between this state-space approach and the standard input-space-based quantisation density (Theorems 4.19, 4.21, 4.22 and 4.23). We also show how to directly utilise CAQS quantisers to design static output feedback strategies that employ infimum density quantisers (Theorem 4.20).

Chapter 5. We derive a new result on infimum quantisation density for multiple-input systems, optimising over the class of quantisers that have levels in a one-dimensional subspace of the input space (Theorem 5.3). We also show that our result partially replaces an incorrect intermediate result in Elia and Frazzoli (2002).

Chapter 6. We derive ultimate bound expressions for perturbed discrete-time systems (Theorem 6.5) and sampled-data systems (Theorem 6.8 and Lemma 6.9). These expressions are believed to be novel. A key feature that distinguishes these results from other ultimate-bound derivation methods is the particular componentwise form of the perturbation bound [see (6.36)]. This form for the perturbation bound is particularly well-suited to the analysis of schemes where different combinations of uniform, logarithmic and semitruncated logarithmic quantisers are simultaneously employed on the same system.

Chapter 7. We extend the results of Chapter 6 to perturbed systems with more general componentwise perturbation bounds. This extension allows us to derive ultimate bounds for a class of nonlinear systems by regarding them as perturbed linear systems, with perturbation bounds that may depend on the system state. We provide systematic methods for the derivation of ultimate bounds (Theorem 7.4 for continuous-time systems and Theorem 7.8 for discrete-time systems), jointly with a region of attraction to the ultimate bound (Algorithm 1 and Theorem 7.5 for continuous-time systems, and Theorem 7.9 for discrete-time systems).

1.4 Associated Publications and Related Work

Most of the results presented in this thesis have been published by the author in journal and conference papers. The following list details the relevant publications:

Journal Papers

- H. Haimovich, M. M. Seron and G. C. Goodwin. Geometric Characterization of Multivariable Quadratically Stabilizing Quantizers. *International Journal of Control*, 79(8):845-857, 2006.
- E. Kofman, H. Haimovich and M. M. Seron. A Systematic Method to Obtain Ultimate Bounds for Perturbed Systems. *International Journal of Control*, 2006. In press.

Conference Papers

- H. Haimovich and M. M. Seron. On infimum quantization density for multiple-input systems. In *Proc. 44th IEEE Conf. on Decision and Control, Seville, Spain*, pp. 7692-7697, 2005.
- H. Haimovich. Stabilizing static output feedback via coarsest quantizers. In *16th IFAC World Congress, Prague, Czech Republic*, 2005.

Other related works published by the author during his Ph.D. studies are:

Book Chapters

- T. Perez and H. Haimovich. Output Feedback Optimal Control with Constraints. In G. C. Goodwin, M. M. Seron and J. A. De Doná, *Constrained Control and Estimation: An Optimisation Approach*, Springer-Verlag, London, 2005, Chapter 12.
- J. Welsh, H. Haimovich and D. Quevedo. Control over Communication Networks. In G. C. Goodwin, M. M. Seron and J. A. De Doná, *Constrained Control and Estimation: An Optimisation Approach*, Springer-Verlag, London, 2005, Chapter 16.

Journal Papers

- G. C. Goodwin, H. Haimovich, D. E. Quevedo and J. S. Welsh. A moving horizon approach to networked control system design. *IEEE Trans. on Automatic Control*, 49(9):1427-1445, 2004.
- T. Perez, H. Haimovich and G. C. Goodwin. On optimal control of constrained linear systems with imperfect state information and stochastic disturbances. *International Journal of Robust and Nonlinear Control*, 14:379-393, 2004.

Conference Papers

- H. Haimovich, G. C. Goodwin and J. S. Welsh. Set-valued observers for Constrained State Estimation of Discrete-time Systems with Quantized Measurements. In *Proc. 5th Asian Control Conference*, Melbourne, Australia, pp. 1947-1955, 2004.

- H. Haimovich, G. C. Goodwin and D. E. Quevedo. Moving horizon Monte Carlo state estimation for linear systems with output quantization. In *Proc. 42nd IEEE Conf. on Decision and Control, Maui, HI, USA*, pp. 4859-4864, 2003.

- H. Haimovich, T. Perez and G. C. Goodwin. On optimality and certainty equivalence in output feedback control of constrained uncertain linear systems. In *Proc. European Control Conference*, Cambridge, UK, 2003.

- H. Haimovich, M. M. Seron, G. C. Goodwin and J. C. Agüero. A neural approximation to the explicit solution of constrained linear MPC. In *Proc. European Control Conference*, Cambridge, UK, 2003.

Confidential Reports for Industry

- J.S. Welsh, G.C. Goodwin, H. Haimovich, H. Tidefelt, A. Rosen and R.H. Middleton, *'Adaptive Powertrain Control: Report 4 - Final Report'*, Report for General Motors, USA, January, 2004.

- A. Rosen, G.C. Goodwin, J.S. Welsh, R.H. Middleton and H. Haimovich, *'Adaptive Powertrain Control: Interim Report No. 5'*, Report for General Motors, USA, December, 2003.

- H. Tidefelt, G.C. Goodwin, J.S. Welsh, R.H. Middleton and H. Haimovich, *'Adaptive Powertrain Control: Interim Report No. 4'*, Report for General Motors, USA, October, 2003.

- J.S. Welsh, H. Haimovich, G.C. Goodwin, R.H. Middleton and J.A. De Dona, *'Adaptive Powertrain Control: Interim Report No. 3'*, Report for General Motors, USA, July, 2003.

- J.S. Welsh, H. Haimovich, G.C. Goodwin, R.H. Middleton and J.A. De Dona, *'Adaptive Powertrain Control: Interim Report No. 2'*, Report for General Motors, USA, June, 2003.

- J.S. Welsh, H. Haimovich, G.C. Goodwin, R.H. Middleton and J.A. De Dona, *'Adaptive Powertrain Control: Interim Report No. 1'*, Report for General Motors, USA, May, 2003.

- J.S. Welsh, H. Haimovich, G.C. Goodwin, R.H. Middleton and J.A. De Dona, *'Adaptive Powertrain Control: Report 3 - System Identification for Automotive Powertrain Control'*, Report for General Motors, USA, March, 2003.

- J.S. Welsh, H. Haimovich, G.C. Goodwin, R.H. Middleton, J.C. Agüero, J.A. De Dona and M.M. Seron, *'Adaptive Powertrain Control Report No.2a'*, Report for General Motors, USA, November, 2002.

Part I

Geometric Approach to Quadratic Stabilisation and Quantisation Density

Chapter 2

Quadratic Stabilisation and Quantisation Density

2.1 Overview

The concept of quantisation density was first introduced by Elia and Mitter (2001) into the context of quadratic stabilisation of linear systems. Intuitively, the density of a quantiser is lower than that of another quantiser if the values of the former are more separated than those of the latter. In this sense, a quantiser can be regarded as being more efficient in the use of its quantisation levels if its density is lower.

Elia and Mitter (2001) consider and solve the problem of finding a least dense quantiser over all quantisers that quadratically stabilise a given linear single-input system. These authors begin by considering discrete-time systems and approach the problem by dividing it into two parts. First, a least dense quantiser over all quantisers that quadratically stabilise *with respect to a given Control Lyapunov Function (CLF)* (see Definition 2.3) is explicitly found. The density of this quantiser is explicitly derived and depends, among other quantities, on the matrix defining the given quadratic CLF. The second part of the approach consists in optimising this density over all matrices that define quadratic CLFs for the system. The end result of this two-part approach is the derivation of an optimum quantisation density, corresponding to a least dense quantiser over all quadratically stabilising quantisers, and also to show how such a least dense quantiser may be constructed. Elia and Mitter then develop a comprehensive treatment of stabilisation of single-input linear systems with least dense quantisers, providing results which deal with continuous-time systems, state estimation with quantisers and practical stabilisation with finite quantisers arising from the truncation of a least dense quantiser.

For multiple-input systems the problem of optimising quantisation density over all quantisers that

quadratically stabilise a system still remains largely open. Elia and Frazzoli (2002) give a straightforward generalisation of the quantisation density definition to two-input systems and provide lower bounds on the infimum quantisation density for two-input systems. Tighter lower bounds on the infimum quantisation density for two-input systems are given in Elia (2002). Kao and Venkatesh (2002) analyse different quantisation schemes and their densities for linear multiple-input systems. However, explicit design of a multivariable quantiser with finite (though not necessarily infimum) quantisation density is performed only when quadratic stabilisation is possible through the use of a two-dimensional subspace of the input space.

Throughout Part I of this thesis, we will derive results contributing to the problem of optimising quantisation density over all quantisers that quadratically stabilise a discrete-time system. We will be mainly concerned with multiple-input systems though we will also provide a different approach to deal with single-input systems. We will focus on the optimisation of quantisation density for a given CLF, which corresponds to the first part of Elia and Mitter's two-part approach.

In this chapter, we begin with a brief review of quadratic stabilisation of linear time-invariant discrete-time systems. After this brief review, we focus on quantisation density in the context of multiple-input systems. We will provide a straightforward generalisation of the definition of quantisation density to multiple-input systems and will then derive several new results in this context.

2.2 Quadratic Stabilisation of Discrete-time Systems

We consider a discrete-time linear time-invariant system, defined by

$$x(k+1) = Ax(k) + Bu(k), \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $u(k) \in \mathbb{R}^m$ is the current control, and $x(k) \in \mathbb{R}^n$ is the current state. We assume that the matrix A has at least one eigenvalue outside or on the unit circle, B has full column rank and the pair (A, B) is stabilisable.

The main ingredient in the analysis of quadratic stabilisation of system (2.1) is a positive definite quadratic function $V : \mathbb{R}^n \to \mathbb{R}_{+,0}$, of the form

$$V(x) = x^T P x, \quad \text{where } P = P^T > 0. \tag{2.2}$$

Although system (2.1) is linear, applying a nonlinear feedback $u = q(x)$ yields a nonlinear closed-loop system. The resulting closed-loop system is said to be quadratically stable if and only if it admits a quadratic Lyapunov function. We thus employ the following definitions.

Definition 2.1 (Quadratic Stability) *A time-invariant discrete-time system of the form $x(k+1) = f(x(k))$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $f(0) = 0$, is said to be quadratically stable with respect to*

V if V is a quadratic positive definite function of the form (2.2) *and $V(f(x)) - V(x) < 0$, for all nonzero $x \in \mathbb{R}^n$. It is said to be quadratically stable if there exists V such that $x(k+1) = f(x(k))$ is quadratically stable with respect to V.*

Definition 2.2 (Quadratically Stabilising Feedback) *A static feedback $u = q(x)$ is said to quadratically stabilise system (2.1) with respect to V if the closed-loop system $x(k+1) = Ax(k) + Bq(x(k))$ is quadratically stable with respect to V. It is said to quadratically stabilise system* (2.1) *if the closed-loop system $x(k+1) = Ax(k) + Bq(x(k))$ is quadratically stable.*

Given a function V of the form (2.2), we will analyse feedback laws that quadratically stabilise system (2.1) with respect to V. However, not every function V of the form (2.2) will allow such a feedback law to exist. We therefore employ the following definition.

Definition 2.3 (CLF) *A positive definite quadratic function of the form $V(x) = x^T Px$, with $P = P^T > 0$, is said to be a control Lyapunov function (CLF) for system (2.1) if a static feedback $u = q(x)$ exists such that the closed-loop system $x(k+1) = Ax(k) + Bq(x(k))$ is quadratically stable with respect to V.*

The concept of control Lyapunov function is restricted neither to quadratic functions nor to open-loop linear systems, and is thus much more general than what we will utilise in this thesis. The reader is referred to Sontag (1998) for further information.

In the sequel, we will employ the increment of a function V of the form (2.2) along the trajectories of system (2.1), defined as

$$\Delta V(x,u) \triangleq V(Ax+Bu) - V(x) = x^T Lx + 2x^T Mu + u^T B^T PBu, \tag{2.3}$$

where

$$L \triangleq A^T PA - P, \quad M \triangleq A^T PB. \tag{2.4}$$

Since, by assumption, $P = P^T > 0$ and B has full column rank, then $B^T PB > 0$ and hence $B^T PB$ is invertible. We then define the matrices

$$Q \triangleq M(B^T PB)^{-1} M^T - L \quad \text{and} \quad K_{GD} \triangleq -(B^T PB)^{-1} M^T. \tag{2.5}$$

The following result shows how to determine whether a given function V of the form (2.2) is a CLF for system (2.1).

Lemma 2.4 *A function $V : \mathbb{R}^n \to \mathbb{R}_{+,0}$ of the form* (2.2) *is a CLF for system (2.1) if and only if $Q > 0$, where Q was defined in* (2.5) *with L and M as in* (2.4)*.*

Proof. Necessity. Consider $\Delta V(x,u)$ in (2.3). Since $P > 0$ and B has full column rank, then $B^T PB > 0$. Then, given $x \in \mathbb{R}^n$, there exists a unique $u \in \mathbb{R}^m$ that minimises the increment $\Delta V(x,u)$ over

all $u \in \mathbb{R}^m$. This minimiser can be straightforwardly calculated by finding the partial derivative of $\Delta V(x,u)$ with respect to u and equating to zero. Thus, we have

$$\frac{\partial \Delta V(x,u)}{\partial u} = 2x^T M + 2u^T B^T P B. \tag{2.6}$$

Equating (2.6) to zero and solving for u yields $u = K_{GD}x$, with K_{GD} as in (2.5). Given $x \in \mathbb{R}^n$, then $u = K_{GD}x$ is the control action that yields the least increment of V along the trajectories of system (2.1). Therefore, given any function $q : \mathbb{R}^n \to \mathbb{R}^m$, then $\Delta V(x, K_{GD}x) \leq \Delta V(x, q(x))$ for all $x \in \mathbb{R}^n$. Since V is a CLF for system (2.1), then a feedback $u = q(x)$ exists such that $\Delta V(x, q(x)) < 0$, for all nonzero $x \in \mathbb{R}^n$. Hence,

$$\Delta V(x, K_{GD}x) \leq \Delta V(x, q(x)) < 0 \tag{2.7}$$

for all nonzero $x \in \mathbb{R}^n$. Using K_{GD} from (2.5), and (2.3), yields

$$\Delta V(x, K_{GD}x) = -x^T Q x \tag{2.8}$$

with Q as defined in (2.5). Combining (2.7) and (2.8), it follows that $-x^T Q x < 0$, for all nonzero $x \in \mathbb{R}^n$, whence $Q > 0$ follows.

Sufficiency. Consider the static feedback $u = K_{GD}x$, with K_{GD} as in (2.5), and $\Delta V(x,u)$ in (2.3). Note that $\Delta V(x, K_{GD}x) = -x^T Q x < 0$, for all nonzero $x \in \mathbb{R}^n$ because $Q > 0$. Hence, V is a CLF for system (2.1). □

2.3 Quadratic Stabilisation via Quantised Feedback

We next analyse quadratic stabilisation of system (2.1) when the control is a static feedback based on a *quantised* measurement of the state. Note that to achieve quadratic stabilisation of the open-loop unstable system (2.1) by means of a quantised static feedback $u = q(x)$, a quantiser q with an infinite number of levels is needed. We employ the following quantiser definition.

Definition 2.5 (Quantiser) *A* quantiser q *is a discrete-range function* $q : \mathbb{R}^r \to \mathbb{R}^s$ *of the form*

$$q(x) = u_i \text{ if and only if } x \in \mathcal{R}_i, \quad \text{for } i \in \mathbb{Z}. \tag{2.9}$$

The sets $\mathcal{R}_i$ *are called the* quantisation regions *of* q *and* u_i *is called the* value *or* level *of* q *corresponding to* $\mathcal{R}_i$. *The sets* $\mathcal{R}_i$, $i \in \mathbb{Z}$, *satisfy*

$$\bigcup_{i \in \mathbb{Z}} \mathcal{R}_i = \mathbb{R}^r, \text{ and } \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \text{ whenever } i \neq j. \tag{2.10}$$

Since we will characterise quantisers that quadratically stabilise system (2.1), we employ the following definition.

Definition 2.6 (QS Quantiser) *Consider a CLF V of the form* (2.2) *and its increment along the trajectories of system* (2.1)*, $\Delta V(x,u)$ in* (2.3)*. A quantiser $q:\mathbb{R}^n \to \mathbb{R}^m$ that satisfies $q(0)=0$ and*

$$\Delta V(x,q(x)) < 0, \quad \textit{for all } x \in \mathbb{R}^n \setminus \{0\}, \tag{2.11}$$

is called quadratically stabilising *(QS) with respect to V. We say that a quantiser q is just 'QS' instead of 'QS with respect to V' when the CLF V is clear from the context.*

2.4 Quantisation Density

The density of quantisation, as introduced by Elia and Mitter (2001), is aimed at quantifying the efficiency of a quantiser in the use of its levels. It is well-known that any quantiser that quadratically stabilises a given open-loop-unstable discrete-time LTI system necessarily has an infinite number of levels, which become increasingly closer near the origin. Therefore, the density of quantisation has been defined to provide a finite value for quantisers having these features (for example, for logarithmic quantisers).

The concept of quantisation density, as introduced in Elia and Mitter (2001), applies to symmetric quantisers with scalar levels, that is, quantisers $q:\mathbb{R}^r \to \mathbb{R}$ that satisfy $q(x) = -q(-x)$ for all $x \in \mathbb{R}^r$. A generalisation of this concept to quantisers with two-dimensional levels ($q:\mathbb{R}^r \to \mathbb{R}^2$) appears in Elia and Frazzoli (2002) and Elia (2002). We next provide a straightforward generalisation to quantisers with levels of arbitrary dimension.

Definition 2.7 (Quantisation Density) *Given a quantiser $q:\mathbb{R}^r \to \mathbb{R}^s$, let $\mathcal{U}(q)$ denote the range of q, that is,*

$$\mathcal{U}(q) \triangleq \{u \in \mathbb{R}^s : u = q(x) \textit{ for some } x \in \mathbb{R}^r\}. \tag{2.12}$$

For $\epsilon \in (0,1]$, let $C^s(\epsilon)$ be the following region in $\mathbb{R}^s$:

$$C^s(\epsilon) \triangleq \{u \in \mathbb{R}^s : \epsilon \le \|u\|_2 \le 1/\epsilon\}. \tag{2.13}$$

The density of q, denoted $\eta(q)$, is defined as follows:

$$\eta(q) \triangleq \limsup_{\epsilon \to 0} \frac{\#[\mathcal{U}(q) \cap C^s(\epsilon)]}{-2\ln\epsilon}, \tag{2.14}$$

where $\#[\cdot]$ denotes the number of elements (cardinality) of a set.

The measure of density in Definition 2.7 coincides with the one given in Elia and Mitter (2001) when the output of the quantiser q is a scalar (that is, $s=1$) and q satisfies $q(x) = -q(-x)$. If a quantiser q has these features, then note that

$$\#[\mathcal{U}(q) \cap C^1(\epsilon)] = 2\#[\mathcal{U}(q) \cap C^1_+(\epsilon)], \tag{2.15}$$

where $C^1_+(\epsilon)$ is the set defined as

$$C^1_+(\epsilon) \triangleq \{u \in \mathbb{R} : \epsilon \leq u \leq 1/\epsilon\}.$$

Then, from (2.14) and (2.15), we have

$$\eta(q) = \limsup_{\epsilon \to 0} \frac{\#[\mathcal{U}(q) \cap C^1(\epsilon)]}{-2 \ln \epsilon} = \limsup_{\epsilon \to 0} \frac{\#[\mathcal{U}(q) \cap C^1_+(\epsilon)]}{-\ln \epsilon}.$$

The right-hand side of the expression above is precisely the density of q as defined in Elia and Mitter (2001). The density of quantisation of Definition 2.7 is also equal to one half the density defined in Elia and Frazzoli (2002) and Elia (2002) for quantisers having two-dimensional levels.

According to (2.14), the density of a quantiser with a finite number of levels is zero, since then the numerator of the right-hand side of (2.14) is bounded and the denominator grows without bound as $\epsilon \to 0$. Also, a quantiser with radially uniformly spaced values has infinite density, since in this case the numerator grows linearly as $\epsilon \to 0$ while the denominator grows logarithmically. Note also that the density of a quantiser q is infinite if, for some $0 < \epsilon \leq 1$, q has an infinite number of levels in the set $C^s(\epsilon)$.

The quantisation density measure of Definition 2.7 is finite for quantisers having radially logarithmically spaced values. We will verify this statement in Theorem 2.12. We require the following two preliminary results.

Lemma 2.8 *Let $0 < \rho < 1$ and let $\{\epsilon^i_n\}_{n=-1}^{\infty}$ for $i = 1, \ldots, N$ be N sequences, each one satisfying*

$$\epsilon^i_{-1} = 1, \quad \rho \leq \epsilon^i_0 \leq 1, \quad \textit{and} \quad \epsilon^i_n = \rho^n \epsilon^i_0, \quad \forall n \in \mathbb{Z}_+. \tag{2.16}$$

Let $f_i : (0, 1] \to \mathbb{Z}_{+,0}$ for $i = 1, \ldots, N$ be N functions, each one satisfying, for $n \in \mathbb{Z}_{+,0}$,

$$f_i(\epsilon) = n \quad \textit{if and only if} \quad \epsilon^i_n < \epsilon \leq \epsilon^i_{n-1}. \tag{2.17}$$

Define $f : (0, 1] \to \mathbb{Z}_{+,0}$ by

$$f(\epsilon) = \sum_{i=1}^{N} f_i(\epsilon). \tag{2.18}$$

Consider $\{\epsilon^i_0\}_{i=1}^N$ and sort its N elements into nonincreasing order to obtain $\{\zeta_i\}_{i=1}^N$ satisfying

$$\zeta_{i+1} \leq \zeta_i \quad \textit{for} \quad i = 1, \ldots, N-1. \tag{2.19}$$

Define the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ as

$$\epsilon_n = \begin{cases} 1 & \textit{if } n = -1, \\ \rho^{\lfloor \frac{n}{N} \rfloor} \zeta_{(n \bmod N)+1} & \textit{if } n \in \mathbb{Z}_{+,0}, \end{cases} \tag{2.20}$$

where $\lfloor b \rfloor$ denotes the greatest integer not greater than b, and $n \bmod N$ denotes the remainder of dividing n by N. Then,

i) $\epsilon_n \leq \epsilon_{n-1}$ *for all* $n \in \mathbb{Z}_{+,0}$.

ii) $\lim_{n\to\infty} \epsilon_n = 0$.

iii) If $n \in \mathbb{Z}_{+,0}$, *then*

$$f(\epsilon) = n \quad \textit{if and only if} \quad \epsilon_n < \epsilon \leq \epsilon_{n-1}.$$

Proof. i) If $n = 0$, from (2.20) we have $\epsilon_0 = \zeta_1$ and $\epsilon_{-1} = 1$. Since, by definition, $\zeta_1 = \epsilon_0^i$ for some $i \in \{1, \ldots, N\}$ and $\epsilon_0^i \leq 1$ from (2.16), then $\epsilon_0 = \zeta_1 \leq 1 = \epsilon_{-1}$, establishing i) for $n = 0$.

If $n = kN$, with $k \in \mathbb{Z}_+$, then from (2.20) we have $\epsilon_n = \rho^k \zeta_1$ and $\epsilon_{n-1} = \rho^{k-1}\zeta_N$. By definition of ζ_i, and using (2.16), we have $\zeta_1 \leq 1$ and $\rho \leq \zeta_N$. Multiplying the former inequality by $\rho^k > 0$, and the latter by $\rho^{k-1} > 0$, yields $\epsilon_n = \rho^k\zeta_1 \leq \rho^k \leq \rho^{k-1}\zeta_N = \epsilon_{n-1}$, establishing that $\epsilon_n \leq \epsilon_{n-1}$ for $n = kN$ with $k \in \mathbb{Z}_+$. Note that this already establishes i) if $N = 1$.

If $n \neq kN$ for all $k \in \mathbb{Z}_{+,0}$, then $\lfloor \frac{n}{N} \rfloor = \lfloor \frac{n-1}{N} \rfloor$ and

$$n \bmod N > (n-1) \bmod N. \tag{2.21}$$

From (2.20), then

$$\epsilon_n = \rho^r \zeta_{(n \bmod N)+1} \quad \text{and} \quad \epsilon_{n-1} = \rho^r \zeta_{((n-1) \bmod N)+1}, \tag{2.22}$$

where $r \triangleq \lfloor \frac{n}{N} \rfloor = \lfloor \frac{n-1}{N} \rfloor$. Since $\rho > 0$, from (2.22), (2.21) and (2.19), it follows that $\epsilon_n \leq \epsilon_{n-1}$ whenever $n \neq kN$ for all $k \in \mathbb{Z}_{+,0}$. We have thus established i).

ii) This follows straightforwardly from (2.20) and since $0 < \rho < 1$.

iii) Using i) it then follows that $\epsilon_n < \epsilon \leq \epsilon_{n-1}$ if and only if the following N sets of inequalities hold:

$$\epsilon_{n+j} < \epsilon \leq \epsilon_{n-N+j}, \quad \text{for } j = 0, \ldots, N-1, \tag{2.23}$$

where we define $\epsilon_k \triangleq 1$ for $k = -2, \ldots, -N$ if $N > 1$. Using (2.20), note that

$$\epsilon_{n+j} = \rho^{\lfloor \frac{n+j}{N} \rfloor} \zeta_{((n+j) \bmod N)+1}, \quad \text{and} \tag{2.24}$$

$$\epsilon_{n-N+j} = \rho^{\lfloor \frac{n+j}{N} \rfloor - 1} \zeta_{((n+j) \bmod N)+1}. \tag{2.25}$$

Recall that ζ_i for $i = 1, \ldots, N$ are obtained by sorting the values $\{\epsilon_0^i\}_{i=1}^N$ and hence

$$\zeta_i = \epsilon_0^{k(i)} \quad \text{for } i = 1, \ldots, N, \tag{2.26}$$

where the function $k : \{1, \ldots, N\} \to \{1, \ldots, N\}$ is bijective (the function k is a permutation). Therefore, combining (2.23)–(2.26), we have that $\epsilon_n < \epsilon \leq \epsilon_{n-1}$ if and only if

$$\rho^{\lfloor \frac{n+j}{N} \rfloor} \epsilon_0^{k(((n+j) \bmod N)+1)} < \epsilon \leq \rho^{\lfloor \frac{n+j}{N} \rfloor - 1} \epsilon_0^{k(((n+j) \bmod N)+1)} \tag{2.27}$$

for $j = 0, \ldots, N-1$. From (2.16), (2.17) and (2.27), then $\epsilon_n < \epsilon \leq \epsilon_{n-1}$ if and only if

$$f_{\bar{k}_n(j)}(\epsilon) = \left\lfloor \frac{n+j}{N} \right\rfloor \quad \text{for } j = 0, \ldots, N-1, \tag{2.28}$$

where we have defined $\bar{k}_n(j) = k(((n+j) \bmod N)+1)$. Note that the function $\bar{k}_n : \{0, \ldots, N-1\} \to \{1, \ldots, N\}$ is also bijective.

Sufficiency. Recalling (2.18) and using (2.28), then $\epsilon_n < \epsilon \leq \epsilon_{n-1}$ implies that

$$f(\epsilon) = \sum_{i=1}^{N} f_i(\epsilon) = \sum_{j=0}^{N-1} f_{\bar{k}_n(j)}(\epsilon) = \sum_{j=0}^{N-1} \left\lfloor \frac{n+j}{N} \right\rfloor = n. \tag{2.29}$$

Necessity. Proving necessity in iii) is equivalent to proving that provided $n \in \mathbb{Z}_{+,0}$, if $\epsilon \notin (\epsilon_n, \epsilon_{n-1}]$, then $f(\epsilon) \neq n$. Since $\epsilon \in (\epsilon_n, \epsilon_{n-1}]$ if and only if (2.28) holds, then $\epsilon \notin (\epsilon_n, \epsilon_{n-1}]$ is equivalent to

$$f_{\bar{k}_n(j)}(\epsilon) \neq \left\lfloor \frac{n+j}{N} \right\rfloor \quad \text{for some } j \in \{0, \ldots, N-1\}. \tag{2.30}$$

Note that the functions f_i for $i = 1, \ldots, N$ are nonincreasing, and that we already know that $f(\epsilon) = \sum_{i=1}^{N} f_i(\epsilon) = n$ for all $\epsilon_n < \epsilon \leq \epsilon_{n-1}$. We also know, from (2.28) and (2.30), that if $\epsilon \notin (\epsilon_n, \epsilon_{n-1}]$, then at least one of the values $f_i(\epsilon)$ will differ from its value for $\epsilon \in (\epsilon_n, \epsilon_{n-1}]$. Note then that $f(\epsilon)$ must also differ from its value for $\epsilon \in (\epsilon_n, \epsilon_{n-1}]$, because the functions f_i are all nonincreasing. Therefore, we have shown that if $\epsilon \notin (\epsilon_n, \epsilon_{n-1}]$, then $f(\epsilon) \neq n$, concluding the proof. □

Lemma 2.9 *Let $0 < \rho < 1$ and let $u \in \mathbb{R}^s$ satisfy*

$$\rho < \|u\|_2 \leq 1. \tag{2.31}$$

Define the set

$$\mathcal{U} \triangleq \{\rho^j u : j \in \mathbb{Z}\}, \tag{2.32}$$

the function

$$f : (0, 1] \to \mathbb{Z}_{+,0}, \qquad f(\epsilon) \triangleq \#[\mathcal{U} \cap C^s(\epsilon)], \tag{2.33}$$

where $C^s(\epsilon)$ is the set defined in (2.13)*, the quantities*

$$\phi \triangleq \min\{\|u\|_2, \rho/\|u\|_2\} \quad \textit{and} \quad \psi \triangleq \max\{\|u\|_2, \rho/\|u\|_2\}, \tag{2.34}$$

and the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$,

$$\epsilon_n = \begin{cases} 1 & \text{if } n = -1, \\ \rho^{n/2}\psi & \text{if } n \in \mathbb{Z}_{+,0} \text{ and is even,} \\ \rho^{(n-1)/2}\phi & \text{if } n \in \mathbb{Z}_{+,0} \text{ and is odd.} \end{cases} \tag{2.35}$$

Then, the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ is nonincreasing, satisfies $\lim_{n\to\infty} \epsilon_n = 0$ and if $n \in \mathbb{Z}_{+,0}$, then

$$f(\epsilon) = n \quad \textit{if and only if} \quad \epsilon_n < \epsilon \leq \epsilon_{n-1}.$$

Proof. From (2.32), we can write

$$\mathcal{U} = \mathcal{U}_+ \uplus \mathcal{U}_-, \tag{2.36}$$

where

$$\mathcal{U}_+ \triangleq \{\rho^j u : j \in \mathbb{Z}_{+,0}\} \quad \text{and} \quad \mathcal{U}_- \triangleq \{\rho^{-j} u : j \in \mathbb{Z}_+\}, \tag{2.37}$$

and $\uplus$ denotes disjoint union. From (2.33), then

$$f = f_1 + f_2, \tag{2.38}$$

where $f_1 : (0,1] \to \mathbb{Z}_{+,0}$ and $f_2 : (0,1] \to \mathbb{Z}_{+,0}$ are defined by

$$f_1(\epsilon) \triangleq \#[\mathcal{U}_+ \cap C^s(\epsilon)] \quad \text{and} \quad f_2(\epsilon) \triangleq \#[\mathcal{U}_- \cap C^s(\epsilon)]. \tag{2.39}$$

For each $\epsilon \in (0,1]$, the integers $f_1(\epsilon)$ and $f_2(\epsilon)$ are the number of elements of $\mathcal{U}_+$ and $\mathcal{U}_-$, respectively, that are contained in $C^s(\epsilon)$. Define the sequences $\{\epsilon_n^1\}_{n=-1}^{\infty}$ and $\{\epsilon_n^2\}_{n=-1}^{\infty}$ by

$$\epsilon_n^1 \triangleq \rho^n \|u\|_2 \quad \text{and} \quad \epsilon_n^2 \triangleq \frac{\rho^{n+1}}{\|u\|_2}, \text{ if } n \in \mathbb{Z}_{+,0},$$
$$\epsilon_{-1}^1 \triangleq \epsilon_{-1}^2 \triangleq 1.$$

Note that $\epsilon_0^1 = \|u\|_2$ satisfies (2.31). Operating on (2.31) yields $\rho \leq \frac{\rho}{\|u\|_2} = \epsilon_0^2 < 1$. Note then that (2.16) is satisfied for $i = 1, 2$. From (2.13), (2.37) and (2.39), then if $n \in \mathbb{Z}_{+,0}$,

$$f_1(\epsilon) = n \quad \text{if and only if} \quad \epsilon_n^1 < \epsilon \leq \epsilon_{n-1}^1, \tag{2.40}$$

$$f_2(\epsilon) = n \quad \text{if and only if} \quad \epsilon_n^2 < \epsilon \leq \epsilon_{n-1}^2. \tag{2.41}$$

Then, (2.17) is satisfied for $i = 1, 2$. Defining $\zeta_1 \triangleq \psi$ and $\zeta_2 \triangleq \phi$ and recalling (2.34), it follows that $\{\zeta_i\}_{i=1}^2$ is obtained by sorting the values $\epsilon_0^1 = \|u\|_2$ and $\epsilon_0^2 = \rho/\|u\|_2$ into nonincreasing order. Moreover, for $n \in \mathbb{Z}_{+,0}$ we can rewrite (2.35) as

$$\epsilon_n = \begin{cases} 1 & \text{if } n = -1, \\ \rho^{\lfloor \frac{n}{2} \rfloor} \zeta_{(n \bmod 2)+1} & \text{if } n \in \mathbb{Z}_{+,0}. \end{cases} \tag{2.42}$$

Therefore, Lemma 2.8 proves the result. □

Example 2.10 *To gain some insight into the functions f, f_1 and f_2 defined in Lemma 2.9 and its proof, consider $\rho = 0.7$, $u = [0.9 \;\; 0]^T \in \mathbb{R}^2$ and the set $\mathcal{U}$ defined in* (2.32)*. We have $\|u\|_2 = 0.9$ and $\rho/\|u\|_2 \approx 0.778$. From* (2.34) *then $\phi = 0.778$ and $\psi = 0.9$. Figure 2.1 depicts the functions f_1 and f_2 defined in* (2.39)*. Note that, for $n \in \mathbb{Z}_+$, $f_1(\epsilon) = n$ if and only if $\rho^n \|u\|_2 < \epsilon \leq \rho^{n-1} \|u\|_2$ and $f_2(\epsilon) = n$ if and only if $\rho^{n+1}/\|u\|_2 < \epsilon \leq \rho^n/\|u\|_2$. Also, $f_1(\epsilon) = 0$ if and only if $\|u\|_2 < \epsilon \leq 1$ and $f_2(\epsilon) = 0$ if and only if $\rho/\|u\|_2 < \epsilon \leq 1$.*

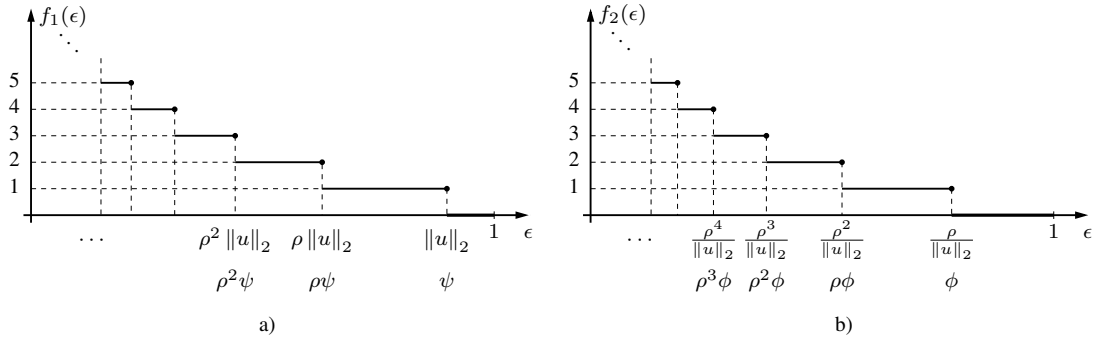


Figure 2.1: The functions f_1 and f_2 defined in (2.39) with $u = [0.9\ \ 0]^T$ and $\rho = 0.7$. a) f_1. b) f_2.

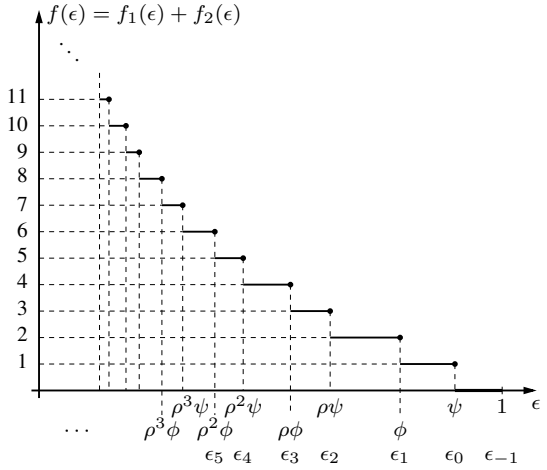


Figure 2.2: The function f defined in (2.33), with $u = [0.9\ \ 0]^T$ and $\rho = 0.7$.

Figure 2.2 depicts the function f defined in (2.33)*, which satisfies $f = f_1 + f_2$, and the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ defined in* (2.35)*. We can verify in Figure 2.2 that $f(\epsilon) = n$ if and only if $\epsilon_n < \epsilon \leq \epsilon_{n-1}$, whenever $n \in \mathbb{Z}_{+,0}$.*

Example 2.11 *As another example of the application of Lemma 2.9, consider $\rho = 0.7$, $u = [\ \sqrt{2}/2\ \sqrt{2}/2\ ]^T$ and the set $\mathcal{U}$ defined in* (2.32)*. We have $\|u\|_2 = 1$ and $\rho/\|u\|_2 = \rho = 0.7$. From* (2.34) *then $\phi = 0.7$ and $\psi = 1$. Figure 2.3 depicts the functions f_1 and f_2 defined in* (2.39)*. Figure 2.4 depicts the function f defined in* (2.33)*. The sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ defined in* (2.35) *is $\epsilon_{-1} = \epsilon_0 = 1$, $\epsilon_1 = \epsilon_2 = 0.7$, $\epsilon_3 = \epsilon_4 = 0.7^2$, ... Lemma 2.9 states that, for $n \in \mathbb{Z}_{+,0}$, $f(\epsilon) = n$ if and only if $\epsilon_n < \epsilon \leq \epsilon_{n-1}$. Note that $f(\epsilon) \neq 2$ for all $\epsilon \in (0, 1]$. Note also that this fact does not invalidate the statement "$f(\epsilon) = 2$ if and only if $\epsilon_2 < \epsilon \leq \epsilon_1$", because $\epsilon_2 = \epsilon_1$ and hence $\epsilon_2 < \epsilon \leq \epsilon_1$ is never true.*

The following result provides the density of a quantiser having radially logarithmically spaced levels.

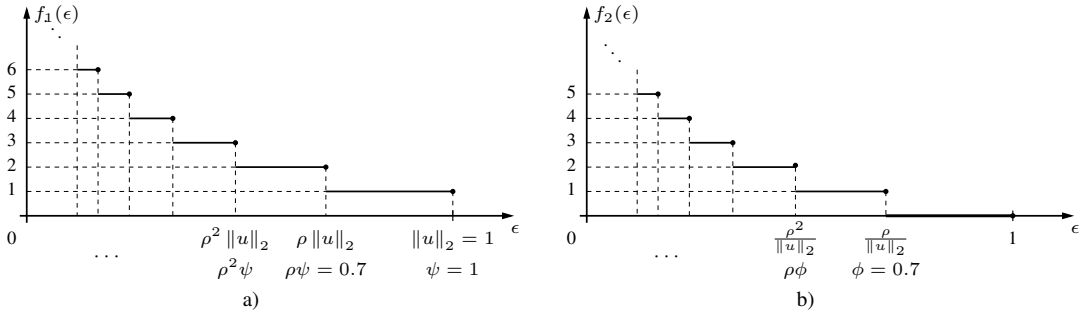


Figure 2.3: The functions f_1 and f_2 defined in (2.39) with $u = [\sqrt{2}/2 \ \sqrt{2}/2]^T$ and $\rho = 0.7$. a) f_1. b) f_2.

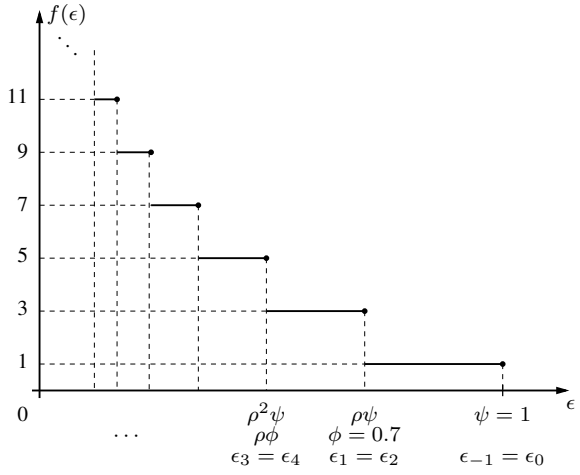


Figure 2.4: $f = f_1 + f_2$.

Theorem 2.12 *Let $q : \mathbb{R}^r \to \mathbb{R}^s$ be a quantiser and let the range of q, $\mathcal{U}(q)$, satisfy*

$$\mathcal{U}(q) = \biguplus_{i=1}^{N} \mathcal{U}_i \cup \{0\}, \tag{2.43}$$

where the sets $\mathcal{U}_i$, $i = 1, \ldots, N$, satisfy

$$\mathcal{U}_i = \{\rho^j u_i : j \in \mathbb{Z}\}, \tag{2.44}$$

with $0 < \rho < 1$ and $u_i \in \mathbb{R}^s \setminus \{0\}$. Then,

$$\eta(q) = \frac{N}{-\ln \rho}. \tag{2.45}$$

Proof. Note that replacing u_i by $\rho^k u_i$ in (2.44) yields identical $\mathcal{U}_i$, whenever $k \in \mathbb{Z}$ and for $i = 1, \ldots, N$. Therefore, without loss of generality we can assume that u_i, for $i = 1, \ldots, N$, satisfy

$$\rho < \|u_i\|_2 \leq 1. \tag{2.46}$$

Consider the set $C^s(\epsilon)$ defined in (2.13) and note that $0 \notin C^s(\epsilon)$ for $\epsilon \in (0,1]$. Then, since $\mathcal{U}(q)$ satisfies (2.43), it follows that $\mathcal{U}(q) \cap C^s(\epsilon) = \biguplus_{i=1}^{N} [\mathcal{U}_i \cap C^s(\epsilon)]$. Moreover, note that each $\mathcal{U}_i$, for $i = 1, \ldots, N$, satisfies

$$\begin{aligned} \mathcal{U}_i &= \mathcal{U}_i^+ \uplus \mathcal{U}_i^-, \quad \text{where} \\ \mathcal{U}_i^+ &= \{\rho^j u_i : j \in \mathbb{Z}_{+,0}\} \quad \text{and} \quad \mathcal{U}_i^- = \{\rho^{-j} u_i : j \in \mathbb{Z}_+\}. \end{aligned} \tag{2.47}$$

Therefore,

$$\mathcal{U}(q) \cap C^s(\epsilon) = \left(\biguplus_{i=1}^{N} [\mathcal{U}_i^+ \cap C^s(\epsilon)] \right) \uplus \left(\biguplus_{i=1}^{N} [\mathcal{U}_i^- \cap C^s(\epsilon)] \right)$$

and hence

$$\#[\mathcal{U}(q) \cap C^s(\epsilon)] = \left(\sum_{i=1}^{N} \#[\mathcal{U}_i^+ \cap C^s(\epsilon)] \right) + \left(\sum_{i=1}^{N} \#[\mathcal{U}_i^- \cap C^s(\epsilon)] \right). \tag{2.48}$$

Define, for $i = 1, \ldots, 2N$, the functions

$$f_i : (0,1] \to \mathbb{Z}_{+,0}, \qquad f_i(\epsilon) = \begin{cases} \#[\mathcal{U}_i^+ \cap C^s(\epsilon)] & \text{if } i \in \{1, \ldots, N\}, \\ \#[\mathcal{U}_{i-N}^- \cap C^s(\epsilon)] & \text{if } i \in \{N+1, \ldots, 2N\}, \end{cases} \tag{2.49}$$

and

$$f : (0,1] \to \mathbb{Z}_{+,0}, \qquad f(\epsilon) \triangleq \sum_{i=1}^{2N} f_i(\epsilon). \tag{2.50}$$

From (2.48)–(2.50), it follows that

$$\#[\mathcal{U}(q) \cap C^s(\epsilon)] = f(\epsilon) \tag{2.51}$$

For each $\epsilon \in (0,1]$, the quantities $f_i(\epsilon)$ indicate how many elements of $\mathcal{U}_i^+$ (for $i = 1, \ldots, N$) or of $\mathcal{U}_{i-N}^-$ (for $i = N+1, \ldots, 2N$) are contained in $C^s(\epsilon)$, and $f(\epsilon)$ indicates how many elements of $\mathcal{U}(q)$ are contained in $C^s(\epsilon)$.

Define the $2N$ sequences $\{\epsilon_n^i\}_{n=-1}^{\infty}$, for $i = 1, \ldots, 2N$, as

$$\epsilon_n^i = \begin{cases} 1 & \text{if } n = -1, \\ \rho^n \left\| u_i \right\|_2 & \text{if } n \in \mathbb{Z}_{+,0}, i \in \{1, \ldots, N\}, \\ \rho^{n+1} / \left\| u_{i-N} \right\|_2 & \text{if } n \in \mathbb{Z}_{+,0}, i \in \{N+1, \ldots, 2N\}. \end{cases} \tag{2.52}$$

From (2.46) and (2.52), then $\rho < \epsilon_0^i \leq 1$ for $i = 1, \ldots, N$. Operating on (2.46) yields $\rho \leq \rho / \left\| u_i \right\|_2 < 1$, and using (2.52) then $\rho \leq \epsilon_0^i < 1$ for $i = N+1, \ldots, 2N$. Then, note that (2.16) holds for $i = 1, \ldots, 2N$. From (2.13), (2.46), (2.47), (2.49) and (2.52), we have, for $i = 1, \ldots, 2N$,

$$f_i(\epsilon) = n \quad \text{if and only if} \quad \epsilon_n^i < \epsilon \leq \epsilon_{n-1}^i,$$

and hence (2.17) holds for $i = 1, \ldots, 2N$. Sort the $2N$ elements $\{\epsilon_0^i\}_{i=1}^{2N}$ into nonincreasing order to obtain $\{\zeta_i\}_{i=1}^{2N}$ satisfying $\zeta_{i+1} \leq \zeta_i$ for $i = 1, \ldots, 2N-1$. Define the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ as

$$\epsilon_n = \begin{cases} 1 & \text{if } n = -1, \\ \rho^{\left\lfloor \frac{n}{2N} \right\rfloor} \zeta_{(n \bmod 2N)+1} & \text{if } n \in \mathbb{Z}_{+,0}. \end{cases} \tag{2.53}$$

Then, Lemma 2.8 iii) shows that, for $n \in \mathbb{Z}_{+,0}$, $f(\epsilon) = n$ if and only if $\epsilon_n < \epsilon \leq \epsilon_{n-1}$.

From (2.14) and (2.51), we can write

$$\eta(q) = \limsup_{\epsilon \to 0} \frac{f(\epsilon)}{-2 \ln \epsilon}. \tag{2.54}$$

By definition of $\limsup$, and since f is defined only on the interval $(0, 1]$, we have

$$\limsup_{\epsilon \to 0} \frac{f(\epsilon)}{-2 \ln \epsilon} = \lim_{\epsilon \to 0^+} \sup_{x \in (0, \epsilon]} \frac{f(x)}{-2 \ln x}. \tag{2.55}$$

The sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ satisfies $\epsilon_{-1} = 1$ by definition. By Lemma 2.8 i) and ii), then $\epsilon_n \leq \epsilon_{n-1}$ for all $n \in \mathbb{Z}_{+,0}$ and $\lim_{n \to \infty} \epsilon_n = 0$. Then, given $\epsilon \in (0, 1]$, we can find $k \in \mathbb{Z}_{+,0}$ such that $\epsilon_k < \epsilon \leq \epsilon_{k-1}$. Hence

$$\sup_{x \in (0, \epsilon_k]} \frac{f(x)}{-2 \ln x} \leq \sup_{x \in (0, \epsilon]} \frac{f(x)}{-2 \ln x} \leq \sup_{x \in (0, \epsilon_{k-1}]} \frac{f(x)}{-2 \ln x} \tag{2.56}$$

and

$$\sup_{x \in (0, \epsilon_{k-1}]} \frac{f(x)}{-2 \ln x} = \sup_{n \geq k} \left(\sup_{x \in (\epsilon_n, \epsilon_{n-1}]} \frac{f(x)}{-2 \ln x} \right). \tag{2.57}$$

Since, by Lemma 2.8 iii), $f(x) = n$ if and only if $\epsilon_n < x \leq \epsilon_{n-1}$, then

$$\sup_{x \in (\epsilon_n, \epsilon_{n-1}]} \frac{f(x)}{-2 \ln x} = \sup_{x \in (\epsilon_n, \epsilon_{n-1}]} \frac{n}{-2 \ln x} = \frac{n}{-2 \ln \epsilon_{n-1}}. \tag{2.58}$$

Combining (2.57) and (2.58) yields

$$\sup_{x \in (0, \epsilon_{k-1}]} \frac{f(x)}{-2 \ln x} = \sup_{n \geq k} \frac{n}{-2 \ln \epsilon_{n-1}}. \tag{2.59}$$

Substituting (2.59) into (2.56) and taking limits yields

$$\lim_{k \to \infty} \sup_{n \geq k+1} \frac{n}{-2 \ln \epsilon_{n-1}} \leq \lim_{\epsilon \to 0^+} \sup_{x \in (0, \epsilon]} \frac{f(x)}{-2 \ln x} \leq \lim_{k \to \infty} \sup_{n \geq k} \frac{n}{-2 \ln \epsilon_{n-1}}$$

and hence, recalling (2.54) and (2.55), it follows that

$$\eta(q) = \limsup_{\epsilon \to 0} \frac{f(\epsilon)}{-2 \ln \epsilon} = \limsup_{n \to \infty} \frac{n}{-2 \ln \epsilon_{n-1}}. \tag{2.60}$$

From (2.53), we have, for $n \in \mathbb{Z}_{+,0}$,

$$\ln \epsilon_n = \ln \zeta_{(n \bmod 2N)+1} + \left\lfloor \frac{n}{2N} \right\rfloor \ln \rho.$$

Then, note that

$$\lim_{n \to \infty} \frac{n}{-2 \ln \epsilon_{n-1}} = \lim_{n \to \infty} \frac{n+1}{-2 \ln \epsilon_n} = \lim_{n \to \infty} \frac{n+1}{-2 \left(\ln \zeta_{(n \bmod 2N)+1} + \left\lfloor \frac{n}{2N} \right\rfloor \ln \rho \right)}. \tag{2.61}$$

Since $\ln \zeta_{(n \bmod 2N)+1}$ is bounded for all $n \in \mathbb{Z}_{+,0}$, it follows that

$$\lim_{n\to\infty} \frac{n+1}{-2\ln \epsilon_n} = \lim_{n\to\infty} \frac{n+1}{-2\left\lfloor \frac{n}{2N} \right\rfloor \ln \rho}.$$

Also, since $\left\lfloor \frac{n}{2N} \right\rfloor = \frac{n}{2N} + \Delta(n)$, where $|\Delta(n)| < 1$ for all $n \in \mathbb{Z}_{+,0}$, then

$$\lim_{n\to\infty} \frac{n+1}{-2\ln \epsilon_n} = \lim_{n\to\infty} \frac{n+1}{-2\left(\frac{n}{2N} + \Delta(n)\right) \ln \rho} = \lim_{n\to\infty} \frac{n+1}{-2\frac{n}{2N} \ln \rho} = \frac{N}{-\ln \rho}. \tag{2.62}$$

Combining (2.61) and (2.62) yields

$$\lim_{n\to\infty} \frac{n}{-2\ln \epsilon_{n-1}} = \frac{N}{-\ln \rho}. \tag{2.63}$$

Since $\lim_{n\to\infty} \frac{n}{-2\ln \epsilon_{n-1}}$ exists, then

$$\limsup_{n\to\infty} \frac{n}{-2\ln \epsilon_{n-1}} = \lim_{n\to\infty} \frac{n}{-2\ln \epsilon_{n-1}}.$$

The result then follows from (2.60) and (2.63). □

From expression (2.45), given by Theorem 2.12 for the density of a quantiser q whose range $\mathcal{U}(q)$ satisfies (2.43) and (2.44), it follows that the lower ρ is, the lower the density of the quantiser q (recall that $0 < \rho < 1$). Note that the lower ρ is, the more radially separated the values of q are. In this sense, we see that the density of quantisation of Definition 2.7 is related to the radial separation of the quantisation levels of a quantiser, and that lower densities correspond to greater separation.

Theorem 2.12 shows that a quantiser $q : \mathbb{R}^r \to \mathbb{R}^s$ with range $\mathcal{U}(q) = \{\rho^j u_1 : j \in \mathbb{Z}\} \cup \{0\}$, where $0 < \rho < 1$ and $u_1 \in \mathbb{R}^s \setminus \{0\}$ has a density $\eta(q) = -1/\ln \rho$. Moreover, the density of a quantiser whose range is a union of such sets is equal to the sum of the densities corresponding to each such set. Theorem 2.12 therefore shows that the density of a quantiser is *additive* over disjoint sets of the form $\{\rho^j u_i : j \in \mathbb{Z}\}$. In particular, when $N = 2$ and $u_2 = -u_1$, application of Theorem 2.12 yields $\eta(q) = 2/-\ln \rho$. This value of the density is the one employed in Elia and Mitter (2001), and corresponds to symmetric scalar logarithmic quantisers.

We next provide an example of the calculations performed in the proof of Theorem 2.12, which derived the density of a quantiser with radially logarithmically separated levels.

Example 2.13 *Consider a quantiser $q : \mathbb{R}^n \to \mathbb{R}^2$ with range $\mathcal{U}(q) = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \{0\}$, where $\mathcal{U}_i = \{\rho^k u_i : k \in \mathbb{Z}\}$, $i = 1, 2$, $\rho = 0.7$, $u_1 = [0.9 \;\; 0]^T$, and $u_2 = [\sqrt{2}/2 \;\; \sqrt{2}/2]^T$. Figure 2.5 a) shows the range of q, $\mathcal{U}(q)$. Note that $\mathcal{U}_1$ and $\mathcal{U}_2$ are disjoint. To find the density of this quantiser, we need to evaluate* (2.14)*. We have*

$$\begin{aligned}\#[\mathcal{U}(q) \cap C^2(\epsilon)] &= \sum_{i=1}^{2} \#[\mathcal{U}_i \cap C^2(\epsilon)] \\ &= \left(\sum_{i=1}^{2} \#[\mathcal{U}_i^+ \cap C^2(\epsilon)]\right) + \left(\sum_{i=1}^{2} \#[\mathcal{U}_i^- \cap C^2(\epsilon)]\right) = \sum_{i=1}^{4} f_i(\epsilon),\end{aligned}$$

where $\mathcal{U}_i^+$ and $\mathcal{U}_i^-$, for $i = 1, 2$, are the sets defined in (2.47)*, and f_i, for $i = 1, \ldots, 4$ are the functions defined in* (2.49)*. Figures 2.5 b) and c) show how to determine the value $\#[\mathcal{U}(q) \cap C^2(\epsilon)]$, for $\epsilon = 0.7$. Figure 2.5 b) shows the set $\mathcal{U}(q) = \mathcal{U}_1^+ \uplus \mathcal{U}_2^+ \uplus \mathcal{U}_1^- \uplus \mathcal{U}_2^- \uplus \{0\}$ and the set $C^2(0.7)$, and Figure 2.5 c) shows the set*

$$\mathcal{U}(q) \cap C^2(0.7) = [\mathcal{U}_1^+ \cap C^2(0.7)] \uplus [\mathcal{U}_2^+ \cap C^2(0.7)] \uplus [\mathcal{U}_1^- \cap C^2(0.7)] \uplus [\mathcal{U}_2^- \cap C^2(0.7)].$$

Note that functions f_1 and f_3 in this example coincide with f_1 and f_2 of Example 2.10, respectively, and f_2 and f_4 in this example coincide with f_1 and f_2 of Example 2.11, respectively. From Figure 2.5 c), we

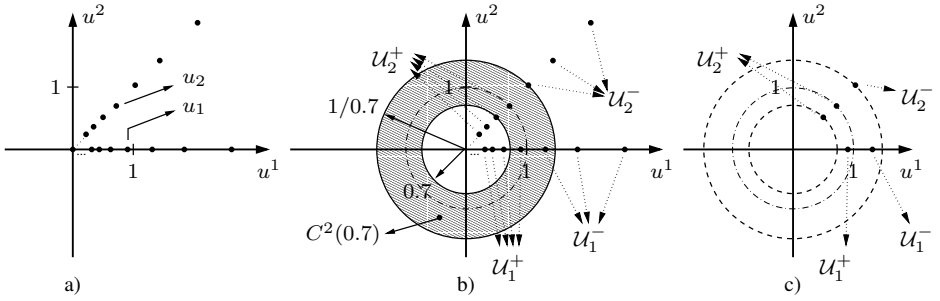


Figure 2.5: a) $\mathcal{U}(q)$. b) $\mathcal{U}_1^+, \mathcal{U}_2^+, \mathcal{U}_1^-, \mathcal{U}_2^-$ and $C^2(0.7)$. c) $\mathcal{U}(q) \cap C^2(0.7)$.

see that

$$\begin{aligned}
\#[\mathcal{U}_1^+ \cap C^2(0.7)] &= f_1(0.7) = 1, \\
\#[\mathcal{U}_2^+ \cap C^2(0.7)] &= f_2(0.7) = 2, \\
\#[\mathcal{U}_1^- \cap C^2(0.7)] &= f_3(0.7) = 1, \quad \textit{and} \\
\#[\mathcal{U}_2^- \cap C^2(0.7)] &= f_4(0.7) = 1.
\end{aligned}$$

We have $\|u_1\|_2 = 0.9$, $\|u_2\|_2 = 1$, $\rho/\|u_1\|_2 = 7/9 \approx 0.778$ and $\rho/\|u_2\|_2 = 0.7$. From (2.52)*, we have $\epsilon_0^1 = 0.9$, $\epsilon_0^2 = 1$, $\epsilon_0^3 = 0.778$ and $\epsilon_0^4 = 0.7$. Sorting these four values into nonincreasing order yields $\zeta_1 = 1$, $\zeta_2 = 0.9$, $\zeta_3 = 0.778$ and $\zeta_4 = 0.7$. Figure 2.6 a) depicts the function $f(\epsilon)$ defined in* (2.50)*, which satisfies* (2.51)*, and the sequence $\{\epsilon_n\}_{n=-1}^{\infty}$ defined in* (2.53)*. Note that $f(\epsilon) = n$ if and only if $\epsilon_n < \epsilon \leq \epsilon_{n-1}$, whenever $n \in \mathbb{Z}_+$. Note that $f(\epsilon) \neq 4$ whenever $\epsilon \in (0, 1]$. Note also that this fact does not invalidate the statement "$f(\epsilon) = 4$ if and only if $\epsilon_4 < \epsilon \leq \epsilon_3$", because $\epsilon_3 = \epsilon_4$ and hence $\epsilon_4 < \epsilon \leq \epsilon_3$ is never true. Figure 2.6 b) shows the function $f(\epsilon)/-2\ln\epsilon$. Note that this function is increasing on any interval $(\epsilon_n, \epsilon_{n-1}]$ and hence $\sup_{x \in (\epsilon_n, \epsilon_{n-1}]} \frac{f(x)}{-2\ln x} = \frac{f(\epsilon_{n-1})}{-2\ln \epsilon_{n-1}} = \frac{n}{-2\ln \epsilon_{n-1}}$, verifying* (2.58)*. According to Theorem 2.12, the density of the quantiser q in this example is $\eta(q) = 2/-\ln 0.7 \approx 5.607$. This value is also depicted in Figure 2.6 b).*

Remark 2.14 *A quantiser $q : \mathbb{R}^r \to \mathbb{R}^s$, whose range, $\mathcal{U}(q)$, is the cartesian product of the ranges of s scalar logarithmic quantisers, has infinite density. This follows since $\mathcal{U}(q)$ will have an infinite number*

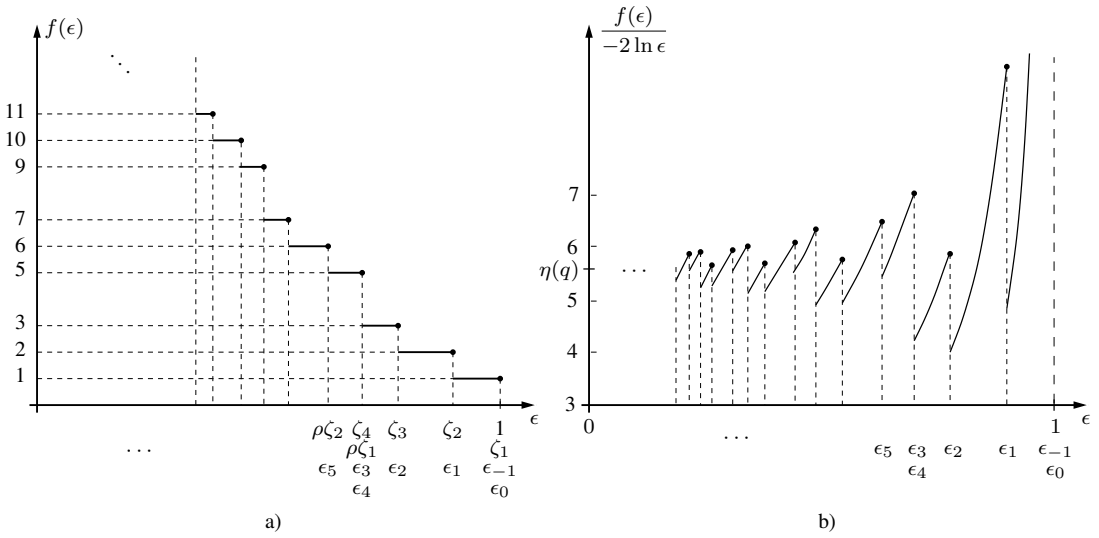


Figure 2.6: a) The function $f(\epsilon)$. b) $f(\epsilon)/-2\ln\epsilon$.

of elements in the set $C^s(\epsilon)$, for some value of ϵ satisfying $0 < \epsilon \leq 1$. Figure 2.7 shows the range, $\mathcal{U}(q)$, of a quantiser $q : \mathbb{R}^r \to \mathbb{R}^2$, having the property that

$$\mathcal{U}(q) \supset \{\rho_1^k a : k \in \mathbb{Z}\} \times \{\rho_2^k b : k \in \mathbb{Z}\}.$$

Figure 2.7 also shows the set $C^s(\epsilon)$ for some value of ϵ satisfying $0 < \epsilon \leq 1$. Note that $C^s(\epsilon)$ contains an infinite number of elements of $\mathcal{U}(q)$ and hence $\eta(q) = \infty$.

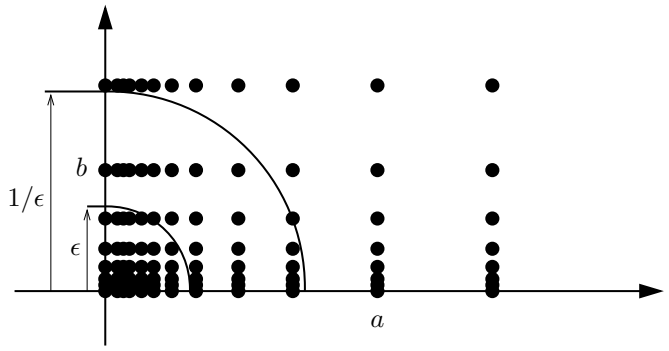


Figure 2.7: Quantiser with range corresponding to the cartesian product of the ranges of 2 scalar logarithmic quantisers. Only positive quadrant shown.

We next provide another result related to quantisation density. One useful feature of linear systems is that they may be easily analysed in different coordinate frames. However, when a linear system is connected to a quantiser, the resulting system is in general not linear. It would then be useful to know whether quantisation density changes under linear transformations. The following result shows that

the density of a quantiser is preserved under a linear transformation, provided the transformation is one-to-one.

Lemma 2.15 *Let $\bar{q} : \mathbb{R}^r \to \mathbb{R}^p$ be a quantiser, let $W \in \mathbb{R}^{s\times p}$ be a matrix having linearly independent columns and let $q : \mathbb{R}^r \to \mathbb{R}^s$ be defined by $q(x) = W\bar{q}(x)$, for all $x \in \mathbb{R}^r$. Then, $\eta(q) = \eta(\bar{q})$.*

Proof. Let $\mathcal{U}(\cdot)$ denote the range of a quantiser, and note that $\mathcal{U}(q) \subset \mathbb{R}^s$ and $\mathcal{U}(\bar{q}) \subset \mathbb{R}^p$. Let $C^s(\epsilon)$ and $C^p(\epsilon)$ be defined as in (2.13). Define the functions

$$f : (0,1] \to \mathbb{Z}_{+,0} \cup \{\infty\}, \qquad f(\epsilon) \triangleq \#[\mathcal{U}(q) \cap C^s(\epsilon)] \tag{2.64}$$

$$\bar{f} : (0,1] \to \mathbb{Z}_{+,0} \cup \{\infty\}, \qquad \bar{f}(\epsilon) \triangleq \#[\mathcal{U}(\bar{q}) \cap C^p(\epsilon)]. \tag{2.65}$$

By definition of q, we have $\mathcal{U}(q) = W\mathcal{U}(\bar{q})$ and hence

$$\begin{aligned} f(\epsilon) &= \#[\mathcal{U}(q) \cap C^s(\epsilon)] = \#[W\mathcal{U}(\bar{q}) \cap C^s(\epsilon)] \\ &= \#\left\{W\bar{u} : \bar{u} = \bar{q}(x),\ x \in \mathbb{R}^r,\ \epsilon \le \|W\bar{u}\|_2 \le 1/\epsilon\right\} \\ &= \#\left\{\bar{u} : \bar{u} = \bar{q}(x),\ x \in \mathbb{R}^r,\ \epsilon \le \|W\bar{u}\|_2 \le 1/\epsilon\right\}, \end{aligned} \tag{2.66}$$

where the last equality above follows since W has linearly independent columns. Note that we may bound $\|W\bar{u}\|_2$ by

$$\alpha_m \|\bar{u}\|_2 \le \|W\bar{u}\|_2 \le \alpha_M \|\bar{u}\|_2, \tag{2.67}$$

where $0 < \alpha_m \le \alpha_M$. Define

$$\alpha_1 \triangleq \min\left\{\alpha_m, \frac{1}{\alpha_M}\right\} \qquad \text{and} \qquad \alpha_2 \triangleq \max\left\{\alpha_M, \frac{1}{\alpha_m}\right\}, \tag{2.68}$$

and note that $0 < \alpha_1 \le 1 \le \alpha_2$. From (2.65), we have, for any $\epsilon \in (0, 1/\alpha_2]$,

$$\bar{f}(\alpha_2\epsilon) = \#\left\{\bar{u} : \bar{u} = \bar{q}(x),\ x \in \mathbb{R}^r,\ \alpha_2\epsilon \le \|\bar{u}\|_2 \le \frac{1}{\alpha_2\epsilon}\right\}. \tag{2.69}$$

We have

$$\alpha_2\epsilon \le \|\bar{u}\|_2 \implies \alpha_2\alpha_m\epsilon \le \alpha_m\|\bar{u}\|_2 \overset{(2.67)}{\implies} \alpha_2\alpha_m\epsilon \le \|W\bar{u}\|_2 \overset{(2.68)}{\implies} \epsilon \le \|W\bar{u}\|_2 \tag{2.70}$$

and

$$\|\bar{u}\|_2 \le \frac{1}{\alpha_2\epsilon} \implies \alpha_2\|\bar{u}\|_2 \le \frac{1}{\epsilon} \overset{(2.68)}{\implies} \alpha_M\|\bar{u}\|_2 \le \frac{1}{\epsilon} \overset{(2.67)}{\implies} \|W\bar{u}\|_2 \le \frac{1}{\epsilon}. \tag{2.71}$$

Combining (2.70) and (2.71) yields

$$\alpha_2\epsilon \le \|\bar{u}\|_2 \le \frac{1}{\alpha_2\epsilon} \implies \epsilon \le \|W\bar{u}\|_2 \le \frac{1}{\epsilon}. \tag{2.72}$$

From (2.66), (2.69) and (2.72), it follows that $\bar{f}(\alpha_2\epsilon) \le f(\epsilon)$. In a similar manner, we can show that $f(\epsilon) \le \bar{f}(\alpha_1\epsilon)$, establishing that

$$\bar{f}(\alpha_2\epsilon) \le f(\epsilon) \le \bar{f}(\alpha_1\epsilon), \quad \text{for all } \epsilon \in (0, 1/\alpha_2]. \tag{2.73}$$

From (2.73), and recalling that $\alpha_2 \geq 1$ and hence $1/\alpha_2 \leq 1$, then

$$\frac{\bar{f}(\alpha_2\epsilon)}{-2\ln\epsilon} \leq \frac{f(\epsilon)}{-2\ln\epsilon} \leq \frac{\bar{f}(\alpha_1\epsilon)}{-2\ln\epsilon}, \quad \text{for all } \epsilon \in (0, 1/\alpha_2].$$

Therefore,

$$\limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_2\epsilon)}{-2\ln\epsilon} \leq \limsup_{\epsilon\to 0} \frac{f(\epsilon)}{-2\ln\epsilon} \leq \limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_1\epsilon)}{-2\ln\epsilon}. \tag{2.74}$$

Note that $\ln\epsilon = \ln(\alpha\epsilon) - \ln\alpha$ whenever $\epsilon > 0$ and $\alpha > 0$. From (2.64) and recalling the definition of quantisation density in (2.14), it follows that

$$\limsup_{\epsilon\to 0} \frac{f(\epsilon)}{-2\ln\epsilon} = \eta(q).$$

We can thus rewrite (2.74) as

$$\limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_2\epsilon)}{-2\ln(\alpha_2\epsilon) + \ln\alpha_2} \leq \eta(q) \leq \limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_1\epsilon)}{-2\ln(\alpha_1\epsilon) + \ln\alpha_1}. \tag{2.75}$$

Since $[-\ln(\alpha\epsilon)] \to \infty$ as $\epsilon \to 0^+$, for any $\alpha > 0$, then it follows from (2.75) that

$$\limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_2\epsilon)}{-2\ln(\alpha_2\epsilon)} \leq \eta(q) \leq \limsup_{\epsilon\to 0} \frac{\bar{f}(\alpha_1\epsilon)}{-2\ln(\alpha_1\epsilon)}. \tag{2.76}$$

From (2.65) and the definition of quantisation density, then (2.76) implies that

$$\eta(\bar{q}) \leq \eta(q) \leq \eta(\bar{q}), \tag{2.77}$$

whence $\eta(q) = \eta(\bar{q})$. □

Lemma 2.15 shows that if two quantisers are related via a one-to-one linear transformation, then their densities are equal. This result will be repeatedly used in the sequel.

2.5 Infimum Quantisation Density

As we have previously mentioned, Elia and Mitter (2001) pose and solve the problem of finding the infimum quantisation density required to quadratically stabilise a given single-input system. We next consider this problem in the context of multiple-input systems. Subsequently, we derive a first result regarding infimum quantisation density in this context.

We thus consider the following problem:

Problem 2.16 *Given system* (2.1) *and a CLF* V *of the form* (2.2)*, solve*

$$\eta^\star = \inf \eta(q), \quad \textit{subject to} \tag{2.78}$$

$$q \textit{ is a quantiser and is QS with respect to } V, \tag{2.79}$$

where $\eta(q)$ *is the density of* q*, as defined in* (2.14)*.*

Theorem 2.17 *Let $q : \mathbb{R}^n \to \mathbb{R}^m$ be a QS quantiser for system* (2.1) *with respect to a CLF V of the form* (2.2)*. Consider the matrices L and M defined in* (2.4) *and the matrix Q defined in* (2.5)*. Define*

$$K \triangleq (B^T P B)^{-1/2} M^T, \tag{2.80}$$

and find the following singular value decomposition[1] *of $KQ^{-1/2}$:*

$$KQ^{-1/2} = S_1 \Sigma S_2^T \tag{2.81}$$

where

$$S_1 \in \mathbb{R}^{m\times m}, \quad S_2 \in \mathbb{R}^{n\times m}, \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m), \quad \textit{and} \quad S_1^T S_1 = \mathrm{I}_m = S_2^T S_2. \tag{2.82}$$

Define the quantiser $\bar{q} : \mathbb{R}^n \to \mathbb{R}^m$ by

$$\bar{q}(x) = q\left(Q^{-1/2} S_2 S_2^T Q^{1/2} x\right), \quad \textit{for all } x \in \mathbb{R}^n. \tag{2.83}$$

Then, $\bar{q}$ is QS and $\eta(\bar{q}) \leq \eta(q)$.

Proof. We begin by showing that $\bar{q}$ is QS. Since q is QS by assumption, then $q(0) = 0$. Then, from (2.83) it follows that $\bar{q}(0) = 0$. Consider the increment of V, as defined in (2.3). Using (2.3)–(2.5) and (2.80), we can write $\Delta V(x, u)$ as

$$\Delta V(x,u) = \left[Kx + (B^T P B)^{1/2} u\right]^T \left[Kx + (B^T P B)^{1/2} u\right] - x^T Q x. \tag{2.84}$$

We have

$$\begin{aligned}\Delta V(Q^{-1/2} S_2 S_2^T Q^{1/2} x, u) = \\ \left[KQ^{-1/2} S_2 S_2^T Q^{1/2} x + (B^T P B)^{1/2} u\right]^T \left[KQ^{-1/2} S_2 S_2^T Q^{1/2} x + (B^T P B)^{1/2} u\right] \\ - x^T Q^{1/2} S_2 S_2^T Q^{-1/2} Q Q^{-1/2} S_2 S_2^T Q^{1/2} x.\end{aligned} \tag{2.85}$$

Using (2.81) and (2.82) in (2.85), and simplifying, yields

$$\begin{aligned}\Delta V(Q^{-1/2} S_2 S_2^T Q^{1/2} x, u) = \\ \left[Kx + (B^T P B)^{1/2} u\right]^T \left[Kx + (B^T P B)^{1/2} u\right] - x^T Q^{1/2} S_2 S_2^T Q^{1/2} x.\end{aligned} \tag{2.86}$$

Note that the matrices $Q^{1/2} S_2 S_2^T Q^{1/2}$ and Q can be written as

$$Q^{1/2} S_2 S_2^T Q^{1/2} = Q^{1/2} \bar{S}_2 \begin{bmatrix} \mathrm{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \bar{S}_2^T Q^{1/2}, \quad \text{and}$$

$$Q = Q^{1/2} \bar{S}_2 \mathrm{I}_n \bar{S}_2^T Q^{1/2},$$

[1] This decomposition was proposed in Kao and Venkatesh (2002).

where $\bar{S}_2 \in \mathbb{R}^{n\times n}$ satisfies $\bar{S}_2\bar{S}_2^T = \mathrm{I}_n$ and $S_2^T = [\mathrm{I}_m \ \mathbf{0}_{m\times n-m}]\bar{S}_2^T$. Hence, $Q \geq Q^{1/2}S_2S_2^TQ^{1/2}$ and it follows from (2.84) and (2.86) that

$$\Delta V(x,u) \leq \Delta V(Q^{-1/2}S_2S_2^TQ^{1/2}x, u), \quad \text{for all } x \in \mathbb{R}^n \text{ and } u \in \mathbb{R}^m. \tag{2.87}$$

By (2.83), we have

$$\Delta V(x, \bar{q}(x)) = \Delta V\left(x, q(Q^{-1/2}S_2S_2^TQ^{1/2}x)\right) \tag{2.88}$$

and from (2.87), with $u = q(Q^{-1/2}S_2S_2^TQ^{1/2}x)$, then

$$\Delta V\left(x, q(Q^{-1/2}S_2S_2^TQ^{1/2}x)\right) \leq \Delta V\left(Q^{-1/2}S_2S_2^TQ^{1/2}x, q(Q^{-1/2}S_2S_2^TQ^{1/2}x)\right). \tag{2.89}$$

Since q is QS by assumption, then $\Delta V(x, q(x)) < 0$ for all nonzero $x \in \mathbb{R}^n$. In particular,

$$\Delta V\left(Q^{-1/2}S_2S_2^TQ^{1/2}x, q(Q^{-1/2}S_2S_2^TQ^{1/2}x)\right) < 0, \tag{2.90}$$

for all $x \in \mathbb{R}^n$ satisfying $Q^{-1/2}S_2S_2^TQ^{1/2}x \neq 0$. Note that $Q^{-1/2}S_2S_2^TQ^{1/2}x \neq 0$ if and only if $S_2^TQ^{1/2}x \neq 0$. Combining (2.88)–(2.90), it follows that

$$\Delta V(x, \bar{q}(x)) < 0, \quad \text{for all } x \in \mathbb{R}^n \text{ such that } S_2^TQ^{1/2}x \neq 0. \tag{2.91}$$

If $S_2^TQ^{1/2}x = 0$, then from (2.83) and since q is QS, we have $\bar{q}(x) = q(0) = 0$. From (2.84), it follows that

$$\begin{aligned}
\Delta V(x,0) &= x^TK^TKx - x^TQx = x^TQ^{1/2}Q^{-1/2}K^TKQ^{-1/2}Q^{1/2}x - x^TQx \\
&\overset{(2.81)}{=} x^TQ^{1/2}S_2\Sigma S_1^TS_1\Sigma S_2^TQ^{1/2}x - x^TQx \\
&= -x^TQx, \quad \text{whenever } S_2^TQ^{1/2}x = 0.
\end{aligned} \tag{2.92}$$

Therefore, we have

$$\Delta V(x, \bar{q}(x)) = \Delta V(x,0) = -x^TQx < 0, \quad \text{whenever } S_2^TQ^{1/2}x = 0 \text{ and } x \neq 0. \tag{2.93}$$

Combining (2.91) and (2.93) yields

$$\Delta V(x, \bar{q}(x)) < 0 \quad \text{for all nonzero } x \in \mathbb{R}^n. \tag{2.94}$$

We have thus established that $\bar{q}$ is QS.

We next show that $\eta(\bar{q}) \leq \eta(q)$. Let $\mathcal{U}(\cdot)$ denote the range of a quantiser [recall (2.12)]. From (2.83), note that $\mathcal{U}(\bar{q}) \subseteq \mathcal{U}(q)$. Therefore, it follows that

$$\#[\mathcal{U}(\bar{q}) \cap C^m(\epsilon)] \leq \#[\mathcal{U}(q) \cap C^m(\epsilon)], \tag{2.95}$$

for all $\epsilon \in (0,1]$, where $C^m(\epsilon)$ is the set defined in (2.13). From (2.14), then $\eta(\bar{q}) \leq \eta(q)$. This concludes the proof. □

Theorem 2.17 shows that, given *any* QS quantiser q, we can construct a QS quantiser $\bar{q}$ with a specific structure that is also QS and whose density is not greater than that of q. The key structural difference between an arbitrary QS quantiser q and a quantiser $\bar{q}$ constructed from q according to (2.83) is that the matrix $Q^{-1/2}S_2S_2^TQ^{1/2}$ has rank m [recall (2.82)] where $m \leq n$ (this inequality follows because the system input matrix B has full column rank). Figure 2.8 shows the quantiser $\bar{q}$ constructed from a given quantiser q according to (2.83). Note that the quantiser $\bar{q}$ can be written as

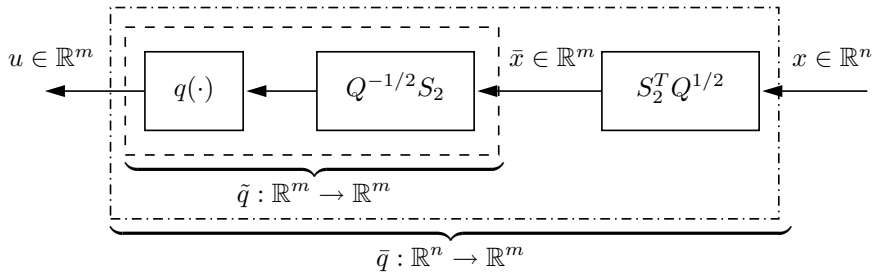


Figure 2.8: Structure of the quantiser $\bar{q}$ in Theorem 2.17.

$$\bar{q}(x) = \tilde{q}(S_2^TQ^{1/2}x), \tag{2.96}$$

where $\tilde{q} : \mathbb{R}^m \to \mathbb{R}^m$.

The result of Theorem 2.17 then implies that the search for the infimum density in Problem 2.16 can be performed exclusively over quantisers $\bar{q}$ having the specific structure (2.96), yielding the same result, $\eta^\star$. Note that the ranges of $\bar{q}$ and $\tilde{q}$ coincide, and hence $\eta(\bar{q}) = \eta(\tilde{q})$. We can thus recast Problem 2.16 as follows.

Problem 2.18 *Given system* (2.1) *and a CLF* V *of the form* (2.2)*, solve*

$$\eta^\star = \inf \eta(\tilde{q}), \quad \textit{subject to} \tag{2.97}$$

$$\tilde{q} : \mathbb{R}^m \to \mathbb{R}^m, \textit{ and the quantiser } \bar{q} : \mathbb{R}^n \to \mathbb{R}^m \textit{ defined by } (2.96) \textit{ is QS with respect to } V, \tag{2.98}$$

where $\eta(\tilde{q})$ *is the density of* $\tilde{q}$*, as defined in* (2.14)*,* Q *was defined in* (2.5)*, and with* S_2 *as in Theorem 2.17.*

We then immediately have the following result.

Theorem 2.19 *The infimum density* $\eta^\star$ *of Problem 2.18 coincides with that of Problem 2.16.*

Consequently, the search for the infimum density $\eta^\star$, which has to be performed over quantisers $q : \mathbb{R}^n \to \mathbb{R}^m$ is reduced to a search over quantisers $\tilde{q} : \mathbb{R}^m \to \mathbb{R}^m$ ($m \leq n$). We have thus provided a first result regarding infimum density over all quantisers that are QS with respect to a *given* CLF.

Remark 2.20 *It is interesting to particularise the result of Theorem 2.17 to single-input systems. In this case, the matrix* K *in* (2.80) *satisfies* $K \in \mathbb{R}^{1\times n}$ *and hence the matrices in the singular value*

decomposition (2.81) *are* $S_1 = 1$, $\Sigma = \sigma \in \mathbb{R}$ *and* $S_2^T = KQ^{-1/2}/\sigma$. *Therefore, it follows that* $S_2^T Q^{1/2} = K/\sigma$ *and from* (2.80) *and* (2.5)*, then* $S_2^T Q^{1/2} = \alpha_1 M^T = \alpha_2 K_{GD}$, *with* $\alpha_1 \in \mathbb{R}$ *and* $\alpha_2 \in \mathbb{R}$, *and where* K_{GD} *was defined in* (2.5). *Then, according to Theorem 2.19, the search for an infimum density quantiser can be performed exclusively over quantisers* $\bar{q} : \mathbb{R}^n \to \mathbb{R}$ *of the form* $\bar{q}(x) = q_s(K_{GD}x)$, *where* $q_s : \mathbb{R} \to \mathbb{R}$ *(scalar quantiser). Note that this structure for the quantisers over which the search must be performed is precisely the one in the first part of the proof of Theorem 1 of Elia and Mitter (2001).*

Remark 2.21 *Kao and Venkatesh (2002) claim that a quantiser that optimises density for a multiple-input system and with respect to a given CLF, needs to have levels only in a minimum-dimension subspace of the input space. However, careful inspection of the results in Kao and Venkatesh (2002) reveals that the definition of quantisation density employed in that paper, in general, coincides neither with that of Elia and Mitter (2001) for single-input systems nor with that of Elia and Frazzoli (2002) for two-input systems. In addition, the quantisation density defined in Kao and Venkatesh (2002), in general, is not even a scalar multiple of that in Elia and Mitter (2001). Therefore, we cannot employ the results of Kao and Venkatesh (2002) in the current context.*

2.6 Chapter Summary

We have briefly reviewed quadratic stabilisation of linear discrete-time systems. We have also reviewed the concept of quantisation density and provided a straightforward generalisation of this concept to quantisers having levels of arbitrary dimension. We have derived several new results regarding quantisation density for multiple-input systems. In particular, Theorem 2.12 derived the density of a multivariable quantiser having radially logarithmically spaced levels and Lemma 2.15 established the invariance of the density of a quantiser under a linear one-to-one transformation.

We have also posed the problem of optimising quantisation density over all quantisers that quadratically stabilise a multiple-input system with respect to a given CLF. We have then derived an important novel result (encompassing Theorem 2.17 and Theorem 2.19) that implies that the search for the infimum density can be performed exclusively over quantisers having a specific structure. Finally, in the case of single-input systems, we have shown that the latter result coincides with a result of Elia and Mitter (2001).

Chapter 3

Geometric Approach to Quadratic Stabilisation with Quantisers

3.1 Overview

In Chapter 2, we have shown that, when searching for the infimum quantisation density over all quantisers that are QS with respect to a given CLF, we need only consider quantisers having a specific form. More precisely, given a system of the form

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

and a CLF V of the form

$$V(x) = x^T P x, \quad \text{where } P = P^T > 0, \tag{3.2}$$

$A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, A is unstable, B has full column rank and the pair (A, B) is stabilisable, then we need only consider quantisers $q : \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$q(x) = \tilde{q}(S_2^T Q^{1/2} x), \tag{3.3}$$

where $\tilde{q} : \mathbb{R}^m \to \mathbb{R}^m$ is a quantiser, and $S_2 \in \mathbb{R}^{n\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are matrices constructed from the system and CLF matrices A, B and P.

The derivation of the structure (3.3), which restricts the quantisers that we need to consider, was performed without dealing with specific details on the construction of QS quantisers. However, to derive additional results regarding quantisation density for multiple-input systems, we will require further insight into specific features of QS quantisers. This chapter will therefore be concerned with characterising QS quantisers.

As a first step toward the characterisation of all QS quantisers of the form (3.3), we will focus on the simplest (in some sense) of the quantiser structures that can be put into the form (3.3). We will thus

consider quantisers $\tilde{q} : \mathbb{R}^m \to \mathbb{R}^m$ of the specific form

$$\tilde{q}(\bar{x}) = W\mathring{q}(\bar{D}^T\bar{x}), \tag{3.4}$$

where $W \in \mathbb{R}^{m\times p}$ and $\bar{D} \in \mathbb{R}^{m\times p}$ have linearly independent columns, $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$ is a quantiser, and the dimension p is as low as possible. Note that the quantiser $\mathring{q}$ is the only nonlinear element in the setting that we consider. Therefore, requiring the dimension p to be as low as possible constrains the nonlinear operation $\mathring{q}$ to occur between spaces of minimum dimension. In this sense, we may then regard the structure that we impose on $\tilde{q}$ to be the simplest possible.

Combining (3.3) and (3.4), we can write

$$q(x) = W\mathring{q}(D^Tx), \tag{3.5}$$

where $D \in \mathbb{R}^{n\times p}$ satisfies

$$D^T = \bar{D}^T S_2^T Q^{1/2}. \tag{3.6}$$

Figure 3.1 shows the setting that we consider. The requirement that the dimension p in the scheme of Figure 3.1 be as low as possible will constitute a key feature in the derivations of this chapter.

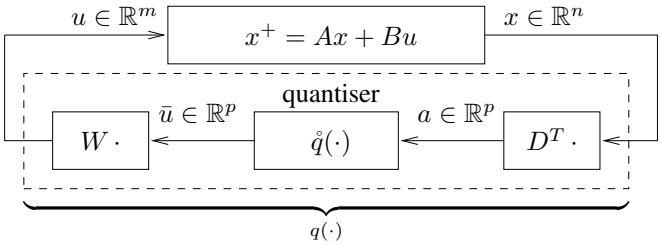


Figure 3.1: The quantised feedback considered: $u = q(x) = W\mathring{q}(D^Tx)$.

The first part of this chapter (§3.2–§3.4) is concerned with the derivation of necessary and sufficient conditions for a quantiser q of the form (3.5) to be QS. In §3.2 we derive the lowest possible value for the dimension p in the scheme of Figure 3.1. In §3.3, we give a geometric interpretation to the fact that a quantiser is QS, and we characterise a QS quantiser in terms of its quantisation regions and values. Our derivations are based on explicit geometric considerations and will thus provide a geometric approach to quadratic stabilisation by means of quantisers of the specific form considered. In §3.4, we derive necessary and sufficient conditions for a quantiser q of the form (3.5) to be QS.

In the second part of this chapter (§3.5), we explicitly construct QS quantisers having finite quantisation density. We will utilise the necessary and sufficient conditions derived in §3.4 to establish that the constructed quantisers are QS. We will also employ the results of Chapter 2 to derive an explicit expression for the density of the quantisers constructed. We provide a summary of the results of this chapter in §3.6.

3.2 Lowest Quantiser Dimension

In this section, we characterise the lowest value for the dimension p in the scheme of Figure 3.1 so that there exists a quantiser $q(x) = W\mathring{q}(D^T x)$ that is QS with respect to a given quadratic CLF V.

We recall from Chapter 2 the expression for the increment of the CLF V in (3.2) along the trajectories of system (3.1):

$$\Delta V(x,u) \triangleq V(Ax+Bu) - V(x) = x^T L x + 2x^T M u + u^T B^T P B u, \tag{3.7}$$

where

$$L \triangleq A^T P A - P, \quad M \triangleq A^T P B. \tag{3.8}$$

To derive the lowest value for p, we require the following preliminary result.

Lemma 3.1 *Consider system (3.1) and a CLF V of the form* (3.2)*. Let the feedback $u = WD^T x$, where $W \in \mathbb{R}^{m\times p}$ and $D \in \mathbb{R}^{n\times p}$ both have linearly independent columns, quadratically stabilise system* (3.1) *with respect to V. Then, there exists a quantiser $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$ such that the quantised feedback $u = W\mathring{q}(D^T x)$ quadratically stabilises system* (3.1) *with respect to V.*

Proof. Note that for any given $\alpha > 0$, we can always build a quantiser $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$ satisfying $\|a - \mathring{q}(a)\|_2 < \alpha \|a\|_2$, for all nonzero $a \in \mathbb{R}^p$. (This can be achieved, for example, with $\mathring{q}$ consisting of componentwise scalar logarithmic quantisers, each of these scalar quantisers having relative errors as small as desired.) Let $K = WD^T$, and consider ΔV, the increment of V, defined in (3.7). By assumption, $\Delta V(x, Kx) < 0$, for all $x \in \mathbb{R}^n \setminus \{0\}$. Note that $\Delta V(x, Kx) \leq -\beta \|x\|_2^2$, for some $\beta > 0$. Define $\epsilon(x) \triangleq Kx - W\mathring{q}(D^T x)$. We have

$$\|\epsilon(x)\|_2 = \left\|Kx - W\mathring{q}(D^T x)\right\|_2 = \left\|W[D^T x - \mathring{q}(D^T x)]\right\|_2 \leq \|W\|_2 \left\|D^T\right\|_2 \|x\|_2 \alpha.$$

Using (3.7) and (3.8), we then have

$$\begin{aligned}
\Delta V(x, W\mathring{q}(D^T x)) &= \Delta V(x, Kx - \epsilon(x)) \\
&= \Delta V(x, Kx) - 2x^T (M + K^T B^T P B)\epsilon(x) + \epsilon(x)^T B^T P B \epsilon(x) \\
&\leq -\beta \|x\|_2^2 + 2\|x\|_2 \left\|M + K^T B^T P B\right\|_2 \|\epsilon(x)\|_2 + \left\|B^T P B\right\|_2 \|\epsilon(x)\|_2^2 \\
&\leq \Big(-\beta + 2\left\|M + K^T B^T P B\right\|_2 \|W\|_2 \|D\|_2 \alpha + \\
&\qquad \left\|B^T P B\right\|_2 \|W\|_2^2 \|D\|_2^2 \alpha^2\Big) \|x\|_2^2 .
\end{aligned}$$

Thus, $\Delta V(x, W\mathring{q}(D^T x)) < 0$ if $\alpha > 0$ is chosen small enough. Hence, the result follows. □

We next find the lowest value for the dimension p in the scheme of Figure 3.1 so that a quantiser $q(x) = W\mathring{q}(D^T x)$ is QS with respect to a given quadratic CLF V. Parts of this proof follow directly from results in Kao and Venkatesh (2002), where the minimum dimension of a subspace of the input

space that is necessary for quadratic stabilisation is derived. However, we will provide an alternative characterisation in terms of the number of positive eigenvalues of the matrix L in (3.8), and we also directly derive this result considering the setting of Figure 3.1.

Theorem 3.2 *Consider system* (3.1) *and a quadratic CLF V of the form* (3.2)*. Let ℓ be the number of positive eigenvalues of the matrix L defined in (3.8), and suppose that L is invertible. Then, ℓ is the lowest value of p for which there exist matrices $W \in \mathbb{R}^{m\times p}$ and $D \in \mathbb{R}^{n\times p}$ with linearly independent columns, and a quantiser $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$ such that the quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ defined by $q(x) = W\mathring{q}(D^T x)$ is QS with respect to V.*

Proof. We begin by proving that if q is QS with respect to V, then $p \geq \ell$. Note that, since $L = L^T \in \mathbb{R}^{n\times n}$ is invertible and has ℓ positive eigenvalues, then L has $n - \ell$ negative eigenvalues. Since $q(0) = W\mathring{q}(D^T 0) = 0$ and W has linearly independent columns, then $\mathring{q}(0) = 0$. Let $\mathcal{P} = \{x \in \mathbb{R}^n : D^T x = 0\}$ and consider the increment of V, ΔV, defined in (3.7). Note that $\Delta V(x, 0) < 0$ for all $x \in \mathcal{P} \setminus \{0\}$. From (3.7), we have $\Delta V(x, 0) = x^T Lx$ and hence $x^T Lx < 0$ for all nonzero vectors in a subspace of dimension $n - p$, because $D \in \mathbb{R}^{n\times p}$ has linearly independent columns. Since L has $n - \ell$ negative eigenvalues, then $n - p \leq n - \ell$ (see Horn and Johnson, 1985, §4.3.23, p. 192), whence $p \geq \ell$.

We next prove that $p = \ell$ is a valid choice. Define

$$K \triangleq (B^T PB)^{-1/2} M^T, \tag{3.9}$$

with M as in (3.8). Consider the matrix Q defined in (2.5), repeated here for convenience:

$$Q \triangleq M(B^T PB)^{-1} M^T - L. \tag{3.10}$$

By Lemma 2.4, and since V is a CLF, then $Q > 0$. Consider the decomposition proposed in Kao and Venkatesh (2002):

$$KQ^{-1/2} = S_1 \Sigma S_2^T, \tag{3.11}$$

where $S_1 \in \mathbb{R}^{m\times m}$, $S_2 \in \mathbb{R}^{n\times m}$, $S_1^T S_1 = \mathrm{I}_m = S_2^T S_2$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$ and σ_i are the singular values of $KQ^{-1/2}$, arranged in decreasing order. Let s be the number of singular values of $KQ^{-1/2}$ that are greater than or equal to 1. Furthermore, let

$$G \triangleq -(B^T PB)^{-1/2} S_1 \begin{bmatrix} \Sigma_{1:s} & \mathbf{0}_{s\times m-s} \\ \mathbf{0}_{m-s\times s} & \mathbf{0}_{m-s\times m-s} \end{bmatrix} S_2^T Q^{1/2}, \tag{3.12}$$

where $\Sigma_{1:s} \triangleq \mathrm{diag}(\sigma_1, \ldots, \sigma_s)$ contains the singular values of $KQ^{-1/2}$ that are greater than or equal to 1. We next prove that $u = Gx$ is quadratically stabilising with respect to V. Using (3.7)–(3.10), we can write $\Delta V(x, u)$ as

$$\Delta V(x, u) = [Kx + (B^T PB)^{1/2} u]^T [Kx + (B^T PB)^{1/2} u] - x^T Qx. \tag{3.13}$$

We have

$$\begin{aligned} Kx + (B^TPB)^{1/2}Gx &= \left[KQ^{-1/2} + (B^TPB)^{1/2}GQ^{-1/2}\right] Q^{1/2}x \\ &= \left[S_1\Sigma S_2^T - S_1 \begin{bmatrix} \Sigma_{1:s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} S_2^T\right] Q^{1/2}x \qquad (3.14) \\ &= S_1 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{s+1:m} \end{bmatrix} S_2^T Q^{1/2}x, \qquad (3.15) \end{aligned}$$

where in (3.14) we have used (3.11) and (3.12), and in (3.15) we have defined

$$\Sigma_{s+1:m} \triangleq \mathrm{diag}(\sigma_{s+1}, \ldots, \sigma_m).$$

From (3.13) and (3.15), we have

$$\begin{aligned} \Delta V(x, Gx) &= x^TQ^{1/2}S_2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{s+1:m} \end{bmatrix} S_1^TS_1 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{s+1:m} \end{bmatrix} S_2^TQ^{1/2}x - x^TQ^{1/2}Q^{1/2}x \\ &= x^TQ^{1/2} \left(S_2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{s+1:m}^2 \end{bmatrix} S_2^T - \mathrm{I}_n \right) Q^{1/2}x, \qquad (3.16) \end{aligned}$$

where $\Sigma_{s+1:m}^2 = \mathrm{diag}(\sigma_{s+1}^2, \ldots, \sigma_m^2)$ and we have used the fact that $S_1^TS_1 = \mathrm{I}_m$. The expression between round brackets in (3.16) is a negative definite matrix since σ_i, for $i = s+1, \ldots, m$, are the singular values of $KQ^{-1/2}$ that are less than one. Therefore, $\Delta V(x, Gx) < 0$, for all $x \in \mathbb{R}^n \setminus \{0\}$, showing that V is a Lyapunov function for $x(k+1) = (A + BG)x(k)$.

The final step is to prove that $\mathrm{rank}(G) = \ell$. By (3.12), the rank of G is s, where s is the number of singular values greater than or equal to one of the matrix $KQ^{-1/2}$ in (3.11). The squared singular values of $KQ^{-1/2}$ are the eigenvalues of $KQ^{-1}K^T$. From (3.10) and (3.9), we have $Q = K^TK - L$. Since L is invertible, using a matrix inversion formula yields

$$Q^{-1} = -[L^{-1} + L^{-1}K^T(\mathrm{I} - KL^{-1}K^T)^{-1}KL^{-1}], \qquad (3.17)$$

whence

$$KQ^{-1}K^T = -(\mathrm{I} - KL^{-1}K^T)^{-1}KL^{-1}K^T = (F - \mathrm{I})^{-1}F, \qquad (3.18)$$

where we have defined $F \triangleq KL^{-1}K^T$. Let v be an eigenvector of $KQ^{-1}K^T$ with eigenvalue σ^2. Then, using (3.18), we have

$$(F - \mathrm{I})^{-1}Fv = \sigma^2 v \quad \Longleftrightarrow \quad Fv = (F - \mathrm{I})\sigma^2 v \quad \Longleftrightarrow \quad \sigma^2 v = (\sigma^2 - 1)Fv. \qquad (3.19)$$

From (3.19), it follows that $\sigma^2 \neq 1$, since otherwise we would obtain $v = 0$, which would be a contradiction since v is an eigenvector. This proves that $KQ^{-1/2}$ has no singular values equal to one,

and hence s is the number of singular values of $KQ^{-1/2}$ strictly greater than one. Then, from (3.19), we have

$$(F - \mathrm{I})^{-1} F v = \sigma^2 v \quad \Longleftrightarrow \quad Fv = \frac{\sigma^2}{\sigma^2 - 1} v \tag{3.20}$$

and hence v is also an eigenvector of F corresponding to the eigenvalue $\sigma^2/(\sigma^2 - 1)$. Since $\sigma^2 \geq 0$ and $\sigma^2 \neq 1$, it then follows that the number of singular values greater than one of $KQ^{-1/2}$, s, equals the number of positive eigenvalues of $F = KL^{-1}K^T$. Hence s is less than or equal to the number of positive eigenvalues of L^{-1}, which is the same as that of L. Therefore $s \leq \ell$. In addition, since G has rank s, we may write $G = WD^T$, where $W \in \mathbb{R}^{m \times s}$ and $D \in \mathbb{R}^{n \times s}$ both have linearly independent columns. Then, since the feedback $u = Gx$ quadratically stabilises system (3.1) with respect to V, Lemma 3.1 shows that there exists a quantised feedback $u = W\mathring{q}(D^T x)$ with $\mathring{q} : \mathbb{R}^s \to \mathbb{R}^s$ that also achieves the same goal. By the first part of this proof, then $s \geq \ell$. We thus have $s \leq \ell$ and $s \geq \ell$, whence $s = \ell$, showing that the lowest value of p is indeed ℓ. □

Remark 3.3 *Theorem 3.2 gives the lowest value of the dimension p for the scheme of Figure 3.1 to be quadratically stable* with respect to a given CLF V. *This lowest value is equal to the number of positive eigenvalues of the matrix L defined in (3.8). Note that L necessarily has at least one nonnegative eigenvalue. This follows since, for $V(x) = x^T P x$, the open-loop system $x(k+1) = Ax(k)$ gives $V(Ax) - V(x) = x^T L x$. Thus, if all the eigenvalues of L were negative, the open-loop system would be stable, contradicting the assumption that A has at least one eigenvalue outside or on the unit circle.*

Remark 3.4 *The fact that a QS quantiser $q(x) = W\mathring{q}(D^T x)$ exists, where $W \in \mathbb{R}^{m \times \ell}$ has linearly independent columns and ℓ is the number of positive eigenvalues of L, implies that the number of inputs, m, is greater than or equal to the number of positive eigenvalues of L.*

Remark 3.5 *In the proof of Theorem 3.2, it was shown that if the number of positive eigenvalues of the matrix L defined in (3.8) is ℓ, then a linear feedback with rank ℓ that quadratically stabilises system (3.1) with respect to the given CLF V always exists (provided L is invertible). Moreover, one such feedback is $u = Gx$, with G as in* (3.12).

3.3 Geometric Approach

The main aim of this section is to give a geometric interpretation to the fact that a quantiser q is QS and to characterise a QS quantiser in terms of its quantisation regions and values. The geometric interpretation is given in §3.3.1, and the characterisation in §3.3.2.

3.3.1 Geometric Interpretation

The geometric interpretation that we will obtain is based on the analysis of conditions that the quantisation regions and values of the QS quantiser q must necessarily satisfy. We will utilise the following definition.

Definition 3.6 (QS Pair) *Consider V and ΔV as in (3.2) and (3.7), respectively, let $\mathcal{R} \subset \mathbb{R}^n$ and $u \in \mathbb{R}^m$. We say that the pair $(u, \mathcal{R})$ is QS (with respect to V) if $u = 0$ when $0 \in \mathcal{R}$ and*

$$\Delta V(x, u) < 0, \quad \textit{for all } x \in \mathcal{R} \setminus \{0\}. \tag{3.21}$$

The following lemma shows the significance of Definition 3.6 in the current context.

Lemma 3.7 *Let $q : \mathbb{R}^n \to \mathbb{R}^m$ be a quantiser, let $\mathcal{R}_i$ and u_i, for all $i \in \mathbb{Z}$, denote its quantisation regions and corresponding values, respectively. Then, q is QS if and only if $(u_i, \mathcal{R}_i)$ is QS, for all $i \in \mathbb{Z}$.*

We can give a geometric interpretation to (3.21) as follows. Define the sets[1]

$$X(u) \triangleq \{x \in \mathbb{R}^n : \Delta V(x, u) < 0\}, \text{ and for future reference,} \tag{3.22}$$

$$X_0(u) \triangleq \{x \in \mathbb{R}^n : \Delta V(x, u) \leq 0\}. \tag{3.23}$$

Using (3.22), we obtain the following geometric characterisation of a QS pair.

Lemma 3.8 *Consider V, ΔV and $X(u)$ as in (3.2), (3.7) and (3.22), respectively, let $\mathcal{R} \subset \mathbb{R}^n$ and $u \in \mathbb{R}^m$. Then, the pair $(u, \mathcal{R})$ is QS if and only if $u = 0$ when $0 \in \mathcal{R}$ and $\mathcal{R} \setminus \{0\} \subseteq X(u)$.*

Using Lemma 3.8, we readily derive the following characterisation of a QS quantiser in terms of its quantisation regions and values.

Lemma 3.9 *Let $q : \mathbb{R}^n \to \mathbb{R}^m$ be a quantiser, let $\mathcal{R}_i$ and u_i, for all $i \in \mathbb{Z}$, denote its quantisation regions and corresponding values, respectively. Then, q is QS if and only if $u_i = 0$ when $0 \in \mathcal{R}_i$ and*

$$\mathcal{R}_i \setminus \{0\} \subseteq X(u_i) \quad \textit{for all } i \in \mathbb{Z}. \tag{3.24}$$

The approach that we follow is based on the analysis of the set inclusion condition (3.24). This analysis is carried out by exploiting the geometry of the sets $X(u)$ and by considering the constraints imposed on the quantisation regions and values of q by the fact that q has the form $q(x) = W\mathring{q}(D^T x)$.

To analyse the sets $X(u)$ defined in (3.22), recall (3.7) and note that the geometry of $X(u)$ depends on the matrix L defined in (3.8). Since L is symmetric, we can decompose it as

$$L = U^T \Lambda U, \text{ where } UU^T = \mathrm{I}_n \text{ and } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{3.25}$$

[1] A similar idea is used by Ishii and Francis (2002b) in the context of continuous-time systems with switching control.

and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the eigenvalues of L (Horn and Johnson, 1985). In the sequel, we assume that L is invertible. For future reference, note that

$$L^{-1} = U^T \Lambda^{-1} U, \text{ where } \Lambda^{-1} = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_n). \tag{3.26}$$

It is now useful to consider the following affine transformation:

$$T_u(x) \triangleq U(x + L^{-1}Mu), \tag{3.27}$$

where L and M are defined in (3.8). Note that, from (3.25), the transformation T_u is invertible and

$$T_u^{-1}(\tilde{x}) = U^T \tilde{x} - L^{-1}Mu. \tag{3.28}$$

Next, consider the sets $X(u)$ and $X_0(u)$, defined in (3.22) and (3.23), respectively. Let $\tilde{X}(u)$ and $\tilde{X}_0(u)$ be the images under T_u of $X(u)$ and $X_0(u)$, respectively, that is, $\tilde{X}(u) \triangleq T_u(X(u))$ and $\tilde{X}_0(u) \triangleq T_u(X_0(u))$. Using (3.7), (3.25), (3.27) and (3.28), we have

$$\tilde{X}(u) = \{\tilde{x} \in \mathbb{R}^n : \tilde{x}^T \Lambda \tilde{x} + u^T H u < 0\}, \tag{3.29}$$

$$\tilde{X}_0(u) = \{\tilde{x} \in \mathbb{R}^n : \tilde{x}^T \Lambda \tilde{x} + u^T H u \leq 0\}, \tag{3.30}$$

where
$$H \triangleq B^T P B - M^T L^{-1} M. \tag{3.31}$$

Consider next a QS quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ satisfying $q(x) = W\mathring{q}(D^T x)$, where $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$, and both $W \in \mathbb{R}^{m \times p}$ and $D \in \mathbb{R}^{n \times p}$ have linearly independent columns. Let $\mathcal{R}_i$ and u_i, for all $i \in \mathbb{Z}$, denote the quantisation regions and corresponding values of q. Note that the regions $\mathcal{R}_i$ satisfy

$$\mathcal{R}_i = \bigcup_{a \in \mathcal{A}_i} \{x \in \mathbb{R}^n : D^T x = a\}, \tag{3.32}$$

for some $\mathcal{A}_i \subset \mathbb{R}^p$. Recall Lemma 3.9 and condition (3.24) where now $\mathcal{R}_i$ has the form (3.32). By Theorem 3.2, we know that the lowest possible value for p is ℓ, where ℓ is the number of positive eigenvalues of the matrix L defined in (3.8). Since we are interested in the case where the dimension p is as low as possible, we take $p = \ell$. Consider the sets (hyperplanes) $\mathcal{P}(a) \triangleq \{x \in \mathbb{R}^n : D^T x = a\}$, for all $a \in \mathbb{R}^\ell$. Note that, from (3.32), $\mathcal{R}_i = \bigcup_{a \in \mathcal{A}_i} \mathcal{P}(a)$. Hence, $\mathcal{R}_i \setminus \{0\} \subseteq X(u_i)$ if and only if $\mathcal{P}(a) \setminus \{0\} \subseteq X(u_i)$, for all $a \in \mathcal{A}_i$. Since T_u [defined in (3.27)] is invertible, then $\mathcal{R}_i \setminus \{0\} \subseteq X(u_i)$ if and only if

$$T_{u_i}(\mathcal{P}(a)) \setminus \{T_{u_i}(0)\} \subseteq \tilde{X}(u_i), \quad \text{for all } a \in \mathcal{A}_i, \tag{3.33}$$

where $\tilde{X}(u_i) = T_{u_i}(X(u_i))$ was defined in (3.29). Thus, condition (3.24) can be equivalently analysed by considering (3.33) for all $i \in \mathbb{Z}$. Note that, since T_u is an invertible affine transformation, it transforms hyperplanes into hyperplanes and hence $T_{u_i}(\mathcal{P}(a))$ is a hyperplane. The analysis of (3.33) then involves the problem of finding conditions for a whole hyperplane or for a hyperplane minus one point to be contained in $\tilde{X}(u)$. The following theorem gives the solution to this problem.

Theorem 3.10 *Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$, where $\lambda_1, \ldots, \lambda_{n-\ell} < 0$ and $\lambda_{n-\ell+1}, \ldots, \lambda_n > 0$. Let $\tilde{D} \in \mathbb{R}^{n\times \ell}$ have linearly independent columns, let $u \in \mathbb{R}^m$ and $\tilde{a} \in \mathbb{R}^\ell$. Let $\tilde{\mathcal{P}} \triangleq \{\tilde{x} \in \mathbb{R}^n : \tilde{D}^T\tilde{x} = \tilde{a}\}$ and let $\tilde{X}(u)$ and $\tilde{X}_0(u)$ satisfy (3.29) and (3.30), respectively, where $H = H^T \in \mathbb{R}^{m\times m}$. Then,*

1. *$\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$ if and only if*

$$\tilde{D}^T\Lambda^{-1}\tilde{D} > 0, \quad \textit{and} \tag{3.34}$$

$$\tilde{a} = 0. \tag{3.35}$$

2. *$\tilde{\mathcal{P}} \subset \tilde{X}(u)$ only if $u \neq 0$ and*

$$\tilde{D}^T\Lambda^{-1}\tilde{D} \geq 0. \tag{3.36}$$

3. *If $\tilde{D}^T\Lambda^{-1}\tilde{D} > 0$ and $u \neq 0$, then $\tilde{\mathcal{P}} \subset \tilde{X}(u)$ if and only if*

$$\tilde{a}^T(\tilde{D}^T\Lambda^{-1}\tilde{D})^{-1}\tilde{a} < -u^THu. \tag{3.37}$$

4. *If $\tilde{D}^T\Lambda^{-1}\tilde{D} > 0$ and $u \neq 0$, then $\tilde{\mathcal{P}} \subset \tilde{X}_0(u)$ if and only if*

$$\tilde{a}^T(\tilde{D}^T\Lambda^{-1}\tilde{D})^{-1}\tilde{a} \leq -u^THu. \tag{3.38}$$

Proof. See Appendix A. □

In summary, the derivations of this section were as follows. Lemma 3.9 gave a geometric interpretation of a QS quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ as a set inclusion condition in terms of its quantisation regions and values. We then considered the geometry of the sets $X(u)$ and the constraints imposed on the quantisation regions of q by the fact that q has the form $q(x) = W\mathring{q}(D^Tx)$. By utilising the transformation T_u, this led to the derivation of conditions for a hyperplane to be contained in $\tilde{X}(u) = T_u(X(u))$, which was performed in Theorem 3.10.

3.3.2 Characterisation of QS Pairs

By Lemma 3.7, a quantiser is QS if and only if all its quantisation value/region pairs are QS. Hence, we next provide a characterisation of QS pairs. This characterisation will be used in §3.4 to derive necessary and sufficient conditions on W, D and $\mathring{q}$ so that $q(x) = W\mathring{q}(D^Tx)$ is QS. We first establish the following preliminary result.

Lemma 3.11 *Let $D \in \mathbb{R}^{n\times\ell}$ have linearly independent columns, where ℓ is the number of positive eigenvalues of the matrix L defined in (3.8), and define $\mathcal{P} \triangleq \{x \in \mathbb{R}^n : D^Tx = 0\}$. Then, $(u, \mathcal{P})$ is QS if and only if $u = 0$ and $D^TL^{-1}D > 0$.*

Proof. Defining $\tilde{\mathcal{P}} \triangleq T_0(\mathcal{P})$, with T_0 as defined in (3.27), we obtain

$$\tilde{\mathcal{P}} = \{\tilde{x} \in \mathbb{R}^n : D^T U^T \tilde{x} = 0\}. \tag{3.39}$$

Necessity. Note that $0 \in \mathcal{P}$ and since $(u, \mathcal{P})$ is QS, then Definition 3.6 implies that $u = 0$. By Lemma 3.8, the pair $(0, \mathcal{P})$ is QS if and only if $\mathcal{P} \setminus \{0\} \subset X(0)$. Since the transformation T_u is invertible, $\mathcal{P} \setminus \{0\} \subset X(0)$ if and only if $\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$, with $\tilde{X}$ as defined in (3.29) and since $T_0(0) = 0$. From Theorem 3.10 part 1 and (3.39), then $D^T U^T \Lambda^{-1} U D > 0$ which using (3.26) yields $D^T L^{-1} D > 0$.

Sufficiency. By (3.26), $D^T L^{-1} D = D^T U^T \Lambda^{-1} U D$. Then Theorem 3.10 part 1 and (3.39) show that $\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$. Since T_u is invertible and $T_0(0) = 0$, then $\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$ if and only if $\mathcal{P} \setminus \{0\} \subset X(0)$. Using Lemma 3.8, we establish that $(u, \mathcal{P})$ is QS. □

The main result of this section is the following theorem, which provides a characterisation of QS pairs. This result will be used in the derivation of necessary and sufficient conditions in §3.4.

Theorem 3.12 (Characterisation of QS Pairs) *Let $D \in \mathbb{R}^{n \times \ell}$ have linearly independent columns, where ℓ is the number of positive eigenvalues of the matrix L defined in (3.8). Suppose that $D^T L^{-1} D > 0$, let $\mathcal{R}$ be a nonempty region that satisfies*

$$\mathcal{R} = \bigcup_{a \in \mathcal{A}} \{x \in \mathbb{R}^n : D^T x = a\} \tag{3.40}$$

for some $\mathcal{A} \subset \mathbb{R}^\ell$, and let $u \in \mathbb{R}^m$. Then, $(u, \mathcal{R})$ is QS [with respect to the CLF V given in (3.2)*] if and only if one of the following statements holds:*

1) $u = 0$ and $\mathcal{A} = \{0\}$.

2) $u \neq 0$ and $(a + D^T L^{-1} M u)^T (D^T L^{-1} D)^{-1} (a + D^T L^{-1} M u) < -u^T H u$, for all $a \in \mathcal{A}$, where M and H were defined in (3.8) and (3.31), respectively.

Proof. Define $\mathcal{P}(a) \triangleq \{x \in \mathbb{R}^n : D^T x = a\}$ and note that $\mathcal{R} = \bigcup_{a \in \mathcal{A}} \mathcal{P}(a)$.

Necessity. Since $(u, \mathcal{R})$ is QS and $\mathcal{R} = \bigcup_{a \in \mathcal{A}} \mathcal{P}(a)$, then note that $(u, \mathcal{P}(a))$ is QS for all $a \in \mathcal{A}$. Suppose that $u = 0$. Then, by Lemma 3.8, $\mathcal{P}(a) \setminus \{0\} \subset X(0)$, which happens if and only if $T_0(\mathcal{P}(a)) \setminus \{0\} \subset \tilde{X}(0)$, since T_0 is invertible and $T_0(0) = 0$. From (3.27) and (3.28), we have $T_0(\mathcal{P}(a)) = \{\tilde{x} \in \mathbb{R}^n : D^T U^T \tilde{x} = a\}$. Then, Theorem 3.10 part 1, implies that the only possible value of a is $a = 0$. This establishes 1). Next, suppose that $u \neq 0$. Since $(u, \mathcal{P}(a))$ is QS for all $a \in \mathcal{A}$, then Definition 3.6 implies that $0 \notin \mathcal{P}(a)$ for all $a \in \mathcal{A}$. Consequently, from Lemma 3.8 it follows that $\mathcal{P}(a) \subset X(u)$ for all $a \in \mathcal{A}$, which happens if and only if $T_u(\mathcal{P}(a)) \subset \tilde{X}(u)$. From (3.27) and (3.28), we have

$$T_u(\mathcal{P}(a)) = \{\tilde{x} \in \mathbb{R}^n : D^T U^T \tilde{x} = a + D^T L^{-1} M u\}. \tag{3.41}$$

Using Theorem 3.10 part 3, and (3.26), then 2) follows.

Sufficiency. If 1) is true, then $\mathcal{R} = \mathcal{P}(0)$ and, since $D^T L^{-1} D > 0$ by assumption, Lemma 3.11 shows that $(0, \mathcal{P}(0))$ is QS. If 2) is true, then consider (3.41). By Theorem 3.10 part 3, then $T_u(\mathcal{P}(a)) \subset \tilde{X}(u)$ and hence $\mathcal{P}(a) \subset X(u)$, for all $a \in \mathcal{A}$. This proves that $(u, \mathcal{P}(a))$ is QS for all $a \in \mathcal{A}$ and hence $(u, \mathcal{R})$ is QS, completing the proof. □

3.4 Necessary and Sufficient Conditions

In this section, we present a key result. Specifically, we apply the geometric characterisation developed in §3.3 to derive necessary and sufficient conditions for a quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ defined by $q(x) = W\mathring{q}(D^T x)$ to be QS, where $W \in \mathbb{R}^{m \times \ell}$, $D \in \mathbb{R}^{n \times \ell}$, $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ and ℓ is the number of positive eigenvalues of the matrix L defined in (3.8). We require the following preliminary result.

Lemma 3.13 *Let $D \in \mathbb{R}^{n \times \ell}$ have linearly independent columns, where ℓ is the number of positive eigenvalues of the matrix L defined in (3.8). Suppose that $D^T L^{-1} D > 0$. Then,*

$$M^T L^{-1} D (D^T L^{-1} D)^{-1} D^T L^{-1} M > -H, \tag{3.42}$$

where M and H were defined in (3.8) and (3.31), respectively.

Proof. Let $\mathcal{P} = \{x \in \mathbb{R}^n : D^T x = 0\}$ and note that $0 \in \mathcal{P}$. Fix any nonzero $u \in \mathbb{R}^m$. We have $0 \notin X_0(u) = \{x \in \mathbb{R}^n : \Delta V(x, u) \leq 0\}$, since, by (3.7), $\Delta V(0, u) = u^T B^T P B u > 0$ because $P > 0$, B has full column rank and $u \neq 0$. Hence $\mathcal{P} \not\subset X_0(u)$. Consider the transformation T_u defined in (3.27), define $\tilde{\mathcal{P}} \triangleq T_u(\mathcal{P})$ and consider the set $\tilde{X}_0(u)$ defined in (3.30). Since T_u is invertible, then $\mathcal{P} \not\subset X_0(u)$ if and only if $\tilde{\mathcal{P}} \not\subset \tilde{X}_0(u)$. Using (3.27) and (3.28), we have $\tilde{\mathcal{P}} = \{\tilde{x} \in \mathbb{R}^n : D^T U^T \tilde{x} = D^T L^{-1} M u\}$ and defining

$$\tilde{D} \triangleq UD \text{ and } \tilde{a} \triangleq D^T L^{-1} M u, \tag{3.43}$$

we can write $\tilde{\mathcal{P}} = \{\tilde{x} \in \mathbb{R}^n : \tilde{D}^T \tilde{x} = \tilde{a}\}$. From (3.26) and (3.43), we have $\tilde{D}^T \Lambda^{-1} \tilde{D} = D^T L^{-1} D$ and by assumption $D^T L^{-1} D > 0$ and $u \neq 0$. Then, Theorem 3.10 part 4, states that $\tilde{\mathcal{P}} \subset \tilde{X}_0(u)$ if and only if (3.38) holds. Since $\tilde{\mathcal{P}} \not\subset \tilde{X}_0(u)$, then (3.38) cannot be true. Therefore, using (3.43), it follows that

$$u^T M^T L^{-1} D (D^T L^{-1} D)^{-1} D^T L^{-1} M u > -u^T H u,$$

for all nonzero $u \in \mathbb{R}^m$. Then, the result follows. □

We next derive conditions on the matrices W and D, and then proceed to derive conditions on $\mathring{q}$.

Theorem 3.14 (Necessary and Sufficient Conditions on W and D) *Let $D \in \mathbb{R}^{n \times \ell}$ and $W \in \mathbb{R}^{m \times \ell}$ have linearly independent columns, where ℓ is the number of positive eigenvalues of the matrix L,*

defined in (3.8). Then, there exists a quantiser $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ *such that the quantiser* $q : \mathbb{R}^n \to \mathbb{R}^m$ *defined by*

$$q(x) = W\mathring{q}(D^T x), \quad \textit{for all } x \in \mathbb{R}^n, \tag{3.44}$$

is QS if and only if

i) $D^T L^{-1} D > 0$*, and*

ii) $J \triangleq -W^T H W > 0$*, where* H *was defined in (3.31).*

Proof. Define $S \triangleq D^T L^{-1} M W$, with M as in (3.8).

Necessity. Let $\mathcal{R}$ be the quantisation region of q that contains the origin, let u be the value of q corresponding to $\mathcal{R}$ and define $\mathcal{P} \triangleq \{x \in \mathbb{R}^n : D^T x = 0\}$. Since q is QS, then Lemma 3.7 shows that $(u, \mathcal{R})$ is QS. Since $\mathcal{P} \subseteq \mathcal{R}$, then note, straightforwardly from Definition 3.6, that $(u, \mathcal{P})$ also is QS. We thus have that $(u, \mathcal{P})$ is QS and Lemma 3.11 proves i).

By i), we have $D^T L^{-1} D > 0$ and hence $D^T L^{-1} D$ is invertible. Define $R \triangleq (D^T L^{-1} D)^{-1}$ and note that $R > 0$. Since q is QS and satisfies (3.44), then by Lemma 3.7 and Theorem 3.12 it follows that for any $x \in \mathbb{R}^n$ satisfying $D^T x \neq 0$, there exists $u \in \mathbb{R}^m$ satisfying

$$(a + D^T L^{-1} M u)^T R (a + D^T L^{-1} M u) < -u^T H u, \tag{3.45}$$

with $a = D^T x$. Since D has linearly independent columns and $u = Ww$ for $w \in \mathbb{R}^\ell$, it then follows from (3.45) that for any nonzero $a \in \mathbb{R}^\ell$, there exists $w \in \mathbb{R}^\ell$ satisfying

$$(a + Sw)^T R (a + Sw) < w^T J w, \tag{3.46}$$

where we have used the definitions of S and J. Operating on (3.46) yields

$$a^T R a + 2a^T R S w + w^T (S^T R S - J) w < 0. \tag{3.47}$$

By Lemma 3.13, (3.42) holds. Since W has linearly independent columns, then premultiplying (3.42) by W^T and postmultiplying by W yields $S^T R S > J$, whence $S^T R S - J > 0$. Hence given a nonzero $a \in \mathbb{R}^\ell$ there exists $w \in \mathbb{R}^\ell$ satisfying (3.47) if and only if

$$\min_{w \in \mathbb{R}^\ell} a^T R a + 2a^T R S w + w^T (S^T R S - J) w < 0. \tag{3.48}$$

The minimum on the left-hand side of (3.48) can be straightforwardly calculated as:

$$a^T (R - RS(S^T R S - J)^{-1} S^T R) a. \tag{3.49}$$

It then follows that $(R - RS(S^T R S - J)^{-1} S^T R)$ must be negative definite and hence its inverse also must be. Calculating its inverse using a matrix inversion formula yields

$$R^{-1} - S J^{-1} S^T < 0. \tag{3.50}$$

From (3.50) and since $R > 0$, we have

$$0 < R^{-1} < SJ^{-1}S^T. \tag{3.51}$$

Since, $S \in \mathbb{R}^{\ell\times\ell}$ and $J \in \mathbb{R}^{\ell\times\ell}$, then (3.51) shows that $J^{-1} > 0$, whence $J > 0$, establishing ii). This concludes the necessity part of the proof.

Sufficiency. Since i) is true, Lemma 3.13 establishes (3.42). Since W has linearly independent columns, premultiplying (3.42) by W^T and postmultiplying by W yields

$$S^T(D^TL^{-1}D)^{-1}S > J, \tag{3.52}$$

where we have used the definitions of S and J. Since $J > 0$ by ii), it follows from (3.52) that $S \in \mathbb{R}^{\ell\times\ell}$ is nonsingular.

Next, consider the feedback $u = -WS^{-1}D^Tx$. This feedback quadratically stabilises with respect to the given CLF $V(x) = x^TPx$. To see this, consider, for all $u \in \mathrm{Im}(W)$, the hyperplanes

$$\begin{aligned}\mathcal{P}(u) &= \{x \in \mathbb{R}^n : u = -WS^{-1}D^Tx\} = \{x \in \mathbb{R}^n : \bar{u} = -S^{-1}D^Tx\}\\ &= \bigcup_{a\in\{\bar{u}\}} \{x \in \mathbb{R}^n : -S^{-1}D^Tx = a\},\end{aligned} \tag{3.53}$$

where $\bar{u} \in \mathbb{R}^\ell$ is a point that satisfies $u = W\bar{u}$. Note that, for each $u \in \mathrm{Im}(W)$, the point $\bar{u}$ that satisfies $u = W\bar{u}$ is unique because W has linearly independent columns. In particular, $u = 0$ if and only if $\bar{u} = 0$. Let $\tilde{D} \triangleq -DS^{-T}$ and note that $\tilde{D}^TL^{-1}\tilde{D} > 0$ since $D^TL^{-1}D > 0$ and S is nonsingular. Then, using (3.53), Theorem 3.12 part 1) shows that $(0, \mathcal{P}(0))$ is QS. Moreover, we have

$$\tilde{D}^TL^{-1}Mu = \tilde{D}^TL^{-1}MW\bar{u} = -S^{-1}S\bar{u} = -\bar{u}, \text{ and} \tag{3.54}$$

$$-u^THu = -\bar{u}^TW^THW\bar{u} = \bar{u}^TJ\bar{u} \tag{3.55}$$

where we have used the definitions of S, $\tilde{D}$ and J. Using (3.53)–(3.55) and the assumption that $J > 0$, Theorem 3.12 part 2) then shows that the pairs $(u, \mathcal{P}(u))$ are QS, for all nonzero $u \in \mathrm{Im}(W)$. By Definition 3.6, we have $\Delta V(x, u) < 0$ for all $x \in \mathcal{P}(u) \setminus \{0\}$, for all $u \in \mathrm{Im}(W)$. Equivalently, $\Delta V(x, -WS^{-1}D^Tx) < 0$, for all $x \in \mathbb{R}^n \setminus \{0\}$. Hence, the feedback $u = -WS^{-1}D^Tx$ is quadratically stabilising with respect to the given CLF $V(x) = x^TPx$ and has rank ℓ, since both W and D have linearly independent columns. Then, Lemma 3.1 shows that a QS quantiser of the form $q(x) = WS^{-1}\mathring{q}'(D^Tx)$ exists, where $\mathring{q}' : \mathbb{R}^\ell \to \mathbb{R}^\ell$. The result then follows by defining $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ by $\mathring{q}(a) = S^{-1}\mathring{q}'(a)$. □

Remark 3.15 *Note that the necessary and sufficient conditions of Theorem 3.14 have been derived without imposing the form* (3.6) *on the matrix D. Therefore, if we intend to minimise quantisation density, we may have to consider matrices D that, in addition to satisfying the conditions of Theorem 3.14, also satisfy* (3.6).

Remark 3.16 *By the first part of the sufficiency proof of Theorem 3.14, note that conditions i)–ii) of this theorem imply that the matrix $S = D^T L^{-1} MW$ is nonsingular.*

Theorem 3.17 (Necessary and Sufficient Conditions on $\mathring{q}$) *Let $W \in \mathbb{R}^{m\times \ell}$ and $D \in \mathbb{R}^{n\times \ell}$ have linearly independent columns, where ℓ is the number of positive eigenvalues of the matrix L defined in (3.8). Suppose that conditions i)—ii) of Theorem 3.14 are satisfied. Let $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ be a quantiser with quantisation regions denoted by $\mathcal{A}_i$ and corresponding values $\bar{u}_i$, for all $i \in \mathbb{Z}$. Then, the quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ satisfying*

$$q(x) = W\mathring{q}(D^T x), \quad \textit{for all } x \in \mathbb{R}^n, \tag{3.56}$$

is QS if and only if

i) $\bar{u}_j = 0$ *and* $\mathcal{A}_j = \{0\}$ *for some* $j \in \mathbb{Z}$*, and*

ii) $\bar{u}_i \neq 0$ *and*

$$(a + S\bar{u}_i)^T (D^T L^{-1} D)^{-1} (a + S\bar{u}_i) < \bar{u}_i^T J \bar{u}_i, \tag{3.57}$$

for all $a \in \mathcal{A}_i$, for all $i \in \mathbb{Z}$, $i \neq j$.

Proof. Note that, by (3.56), $\bar{u} \in \mathbb{R}^\ell$ is a value of $\mathring{q}$ if and only if $W\bar{u} \in \mathbb{R}^m$ is a value of q. Also, $\mathcal{A}$ is a quantisation region of $\mathring{q}$ if and only if $\bigcup_{a\in\mathcal{A}}\{x \in \mathbb{R}^n : D^T x = a\}$ is a quantisation region of q. Then, let $\mathcal{R}_i$ and u_i be the quantisation regions and values of q that are in correspondence with $\mathcal{A}_i$ and $\bar{u}_i$, for all $i \in \mathbb{Z}$. By Lemma 3.7, q is QS if and only if $(u_i, \mathcal{R}_i)$ are QS, for all $i \in \mathbb{Z}$.

Necessity. Since q is QS, then $q(0) = 0$ and hence $u_j = 0$ is a quantisation value of q, for some $j \in \mathbb{Z}$. Then, $W\bar{u}_j = u_j = 0$, whence $\bar{u}_j = 0$ since W has linearly independent columns. From Theorem 3.12 and since $(0, \mathcal{R}_j)$ is QS, we have $\mathcal{A}_j = \{0\}$, proving i). For any integer $i \neq j$, we have $u_i \neq 0$, whence $\bar{u}_i \neq 0$. Then, Theorem 3.12 part 2) and the fact that $u_i = W\bar{u}_i$ establish (3.57), for all $a \in \mathcal{A}_i$, proving ii).

Sufficiency. Using i), ii) and the fact that $u_i = W\bar{u}_i$ for all $i \in \mathbb{Z}$ in Theorem 3.12 shows that $(u_i, \mathcal{R}_i)$ is QS, for all $i \in \mathbb{Z}$. Then, Lemma 3.7 shows that q is QS. □

Remark 3.18 *For a fixed nonzero $\bar{u}_i \in \mathbb{R}^\ell$, and since $D^T L^{-1} D > 0$ and $J > 0$, all $a \in \mathbb{R}^\ell$ that satisfy (3.57) are contained in an ellipsoid centred at $-S\bar{u}_i$ and whose size depends on $\bar{u}_i^T J \bar{u}_i$. Moreover, since $S^T (D^T L^{-1} D)^{-1} S > J$, then none of these ellipsoids contain the point $a = 0$.*

Theorems 3.14 and 3.17 give necessary and sufficient conditions for the quantised feedback scheme of Figure 3.1 to be quadratically stable with respect to the CLF V defined in (3.2). Theorem 3.14 gives the necessary and sufficient conditions on the matrices W and D, and Theorem 3.17 the conditions on $\mathring{q}$. Theorem 3.17 states that, provided W and D satisfy the necessary and sufficient conditions given by

Theorem 3.14, each quantisation region of $\mathring{q}$ has to be contained in an ellipsoid (see Remark 3.18). This provides a novel geometric characterisation of the quantised feedback laws $u = W\mathring{q}(D^T x)$, where $\mathring{q}$ has the lowest possible dimension, that quadratically stabilise system (3.1) with respect to a given quadratic CLF.

The results derived can be used both for testing whether a given quantised feedback of the form considered stabilises quadratically with respect to a given CLF and for designing a quadratically stabilising quantised feedback. Note that the conditions derived are valid irrespective of the structure of the reduced-dimension quantiser $\mathring{q}$, that is, $\mathring{q}$ can result from the independent quantisation of the ℓ components of $D^T x$ or can be an intrinsically multivariable quantiser.

3.5 Stabilising Finite-density Multivariable Quantiser Design

In this section, we combine the geometric approach of the first part of this chapter and the results of Chapter 2 to design finite-density multivariable QS quantisers.

Specifically, we will design a quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ that satisfies $q(x) = W\mathring{q}(D^T x)$ for all $x \in \mathbb{R}^n$, where $W \in \mathbb{R}^{m\times\ell}$, $D \in \mathbb{R}^{n\times\ell}$, $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ is a quantiser and ℓ is the number of positive eigenvalues of L. The quantiser q will be QS (with respect to the given CLF V) and will have finite quantisation density.

In §3.5.1, we present the specific quantiser $\mathring{q}$ involved in the design of q. In §3.5.2, we give sufficient conditions for q to be QS, and compute its density. We shall henceforth refer to $\mathring{q}$ as the "reduced-dimension quantiser".

3.5.1 Reduced-dimension Quantiser

We next explain the construction of a specific quantiser $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$, where p is arbitrary. Since this construction is somewhat involved, we begin by illustrating the construction when $p = 2$ in Figure 3.2. The quantisation regions of $\mathring{q}$ have square form when $p = 2$ (cubic when $p = 3$, hypercubic when $p > 3$). The centre of each square (cube, hypercube) is its corresponding quantisation value. The construction of the quantiser $\mathring{q}$ involves a parameter, C, which is an odd integer that satisfies $\mathrm{C} \geq 3$. The quantiser $\mathring{q}$ can be constructed for any such value of C ($\mathrm{C} = 5$ in Figure 3.2). The quantiser $\mathring{q}$ also involves the functions $\mathcal{I} : \mathbb{R} \to \frac{2}{\mathrm{C}-1}\mathbb{Z}$, and $j : \mathbb{R}_+ \to \mathbb{Z}$, defined by

$$\mathcal{I}(b) = \frac{2}{\mathrm{C}-1}\left\lceil \frac{\mathrm{C}-1}{2}|b| - \frac{1}{2}\right\rceil \operatorname{sgn}(b), \tag{3.58}$$

$$j(b) = \left\lfloor \frac{\ln \dfrac{\mathrm{C}}{(\mathrm{C}-1)b}}{\ln \dfrac{1}{\rho}} \right\rfloor, \tag{3.59}$$

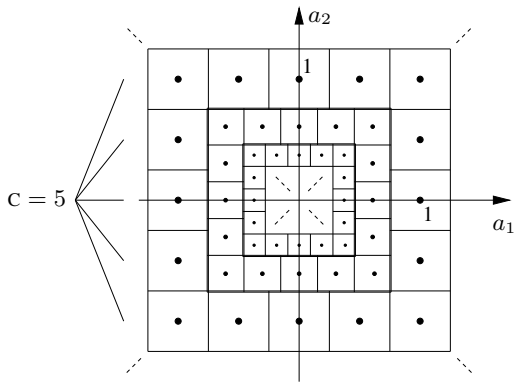


Figure 3.2: Reduced-dimension quantiser $\mathring{q} : \mathbb{R}^2 \to \mathbb{R}^2$. Each solid-line square is a quantisation region of $\mathring{q}$. The centre of each square (represented by a circle) is the corresponding quantisation value.

where $\lceil b \rceil$ denotes the least integer not less than b, $\lfloor b \rfloor$ denotes the greatest integer not greater than b and

$$\rho \triangleq \frac{\mathrm{c} - 2}{\mathrm{c}}. \tag{3.60}$$

For a vector $a \in \mathbb{R}^p$, we denote, with a slight abuse of notation, by $\mathcal{I}(a)$ the vector $[\mathcal{I}(a_1) \cdots \mathcal{I}(a_p)]^T$. The reduced dimension quantiser $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$ is then constructed as follows:

$$\mathring{q}(a) = \begin{cases} 0 & \text{if } a = 0, \quad (3.61a) \\ \rho^{j(\|a\|_\infty)} \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right) & \text{if } a \neq 0. \quad (3.61b) \end{cases}$$

This quantiser has the form depicted in Figure 3.2 when $p = 2$ and $\mathrm{c} = 5$. To gain additional insight into the quantiser $\mathring{q}$ defined by (3.61), we next derive several results that will later allow us to give a simple interpretation of $\mathring{q}$.

Figure 3.3 depicts the function $\mathcal{I} : \mathbb{R} \to \frac{2}{\mathrm{c}-1}\mathbb{Z}$. Note that $\mathcal{I}$ is a uniform scalar quantiser and satisfies

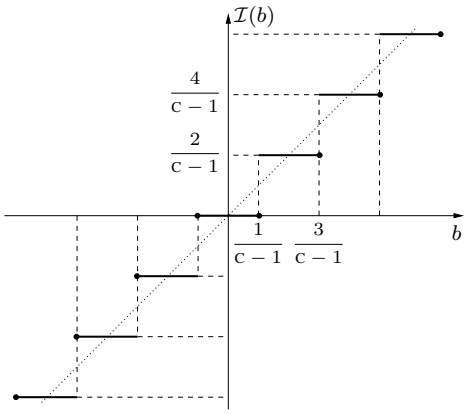


Figure 3.3: The function $\mathcal{I} : \mathbb{R} \to \frac{2}{\mathrm{c}-1}\mathbb{Z}$.

$\mathcal{I}(b) = -\mathcal{I}(-b)$, for all $b \in \mathbb{R}$.

The following lemma derives properties of $\mathcal{I}$ and j.

Lemma 3.19 *Let* $\mathrm{c} \geq 3$ *be an odd integer, let* $b \in \mathbb{R}_+$ *and* $a \in \mathbb{R}^p$*, consider* ρ *as defined in* (3.60) *and the functions* $\mathcal{I}$ *and* j *defined in* (3.58) *and* (3.59)*, respectively. Then,*

i) $j(b) = 0 \quad \Leftrightarrow \quad \mathcal{I}(b) = 1 \quad \Leftrightarrow \quad \dfrac{\mathrm{c}-2}{\mathrm{c}-1} < b \leq \dfrac{\mathrm{c}}{\mathrm{c}-1}.$

ii) $j(\rho^k b) = j(b) + k$*, for all* $k \in \mathbb{Z}$*.*

iii) $\|\mathcal{I}(a)\|_\infty = \mathcal{I}(\|a\|_\infty)$*.*

iv) If $a \neq 0$*, then* $\left\|\mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right)\right\|_\infty = 1$*.*

Proof. i) From (3.59) and (3.60), we have

$$j(b) = 0 \quad \Leftrightarrow \quad 0 \leq \frac{\ln \dfrac{\mathrm{c}}{(\mathrm{c}-1)b}}{\ln \dfrac{\mathrm{c}}{\mathrm{c}-2}} < 1.$$

Since $\mathrm{c} \geq 3$, then $\ln \dfrac{\mathrm{c}}{\mathrm{c}-2} > 0$ and it follows that

$$j(b) = 0 \quad \Leftrightarrow \quad 0 \leq \ln \frac{\mathrm{c}}{(\mathrm{c}-1)b} < \ln \frac{\mathrm{c}}{\mathrm{c}-2},$$

whence

$$j(b) = 0 \quad \Leftrightarrow \quad 1 \leq \frac{\mathrm{c}}{(\mathrm{c}-1)b} < \frac{\mathrm{c}}{\mathrm{c}-2}.$$

Hence,

$$j(b) = 0 \quad \Leftrightarrow \quad \frac{\mathrm{c}-2}{\mathrm{c}-1} < b \leq \frac{\mathrm{c}}{\mathrm{c}-1}.$$

From (3.58), note that

$$\mathcal{I}(b) = \frac{\mathrm{c}-1}{\mathrm{c}-1} = 1 \quad \Leftrightarrow \quad \frac{\mathrm{c}-2}{\mathrm{c}-1} < b \leq \frac{\mathrm{c}}{\mathrm{c}-1}.$$

We have thus established i).

ii) Note that

$$\ln \frac{\mathrm{c}}{(\mathrm{c}-1)\rho^k b} = \ln \frac{\mathrm{c}}{(\mathrm{c}-1)b} + k \ln \frac{1}{\rho}.$$

Then, using (3.59), we have, for $k \in \mathbb{Z}$,

$$j(\rho^k b) = \left\lfloor \frac{\ln \dfrac{\mathrm{c}}{(\mathrm{c}-1)\rho^k b}}{\ln \dfrac{1}{\rho}} \right\rfloor = \left\lfloor \frac{\ln \dfrac{\mathrm{c}}{(\mathrm{c}-1)b}}{\ln \dfrac{1}{\rho}} + k \right\rfloor = \left\lfloor \frac{\ln \dfrac{\mathrm{c}}{(\mathrm{c}-1)b}}{\ln \dfrac{1}{\rho}} \right\rfloor + k = j(b) + k,$$

which establishes ii).

iii) By definition of infinity norm and since $\mathcal{I}(a) = [\mathcal{I}(a_1) \ \cdots \ \mathcal{I}(a_p)]^T$, we have

$$\|\mathcal{I}(a)\|_\infty = \max_i |\mathcal{I}(a_i)|. \tag{3.62}$$

Since $\mathcal{I}:\mathbb{R}\to\mathbb{R}$ satisfies $\mathcal{I}(b)=-\mathcal{I}(-b)$, then $|\mathcal{I}(b)|=|\mathcal{I}(-b)|=|\mathcal{I}(|b|)|$, for all $b\in\mathbb{R}$. Moreover, since $\mathcal{I}(b)\geq 0$ whenever $b\geq 0$, then $|\mathcal{I}(b)|=\mathcal{I}(|b|)$, for all $b\in\mathbb{R}$. Therefore,

$$\max_i |\mathcal{I}(a_i)| = \max_i \mathcal{I}(|a_i|). \tag{3.63}$$

Since $\mathcal{I}:\mathbb{R}\to\mathbb{R}$ is nondecreasing, it follows that

$$\max_i \mathcal{I}(|a_i|) = \mathcal{I}(\max_i |a_i|) = \mathcal{I}(\|a\|_\infty). \tag{3.64}$$

Then iii) follows by combining (3.62)–(3.64).

iv) From (3.59)–(3.60), we have

$$\frac{\ln\left(\dfrac{\mathrm{C}}{(\mathrm{C}-1)\|a\|_\infty}\right)}{\ln\left(\dfrac{\mathrm{C}}{\mathrm{C}-2}\right)} - 1 < j(\|a\|_\infty) \leq \frac{\ln\left(\dfrac{\mathrm{C}}{(\mathrm{C}-1)\|a\|_\infty}\right)}{\ln\left(\dfrac{\mathrm{C}}{\mathrm{C}-2}\right)}, \tag{3.65}$$

whence, since $\mathrm{C}>\mathrm{C}-2>0$,

$$\ln\left(\frac{\mathrm{C}-2}{(\mathrm{C}-1)\|a\|_\infty}\right) < j(\|a\|_\infty)\ln\left(\frac{\mathrm{C}}{\mathrm{C}-2}\right) \leq \ln\left(\frac{\mathrm{C}}{(\mathrm{C}-1)\|a\|_\infty}\right). \tag{3.66}$$

Using (3.60) and operating on (3.66), we obtain

$$\frac{\mathrm{C}-2}{\mathrm{C}-1} < \|a\|_\infty \rho^{-j(\|a\|_\infty)} \leq \frac{\mathrm{C}}{\mathrm{C}-1}. \tag{3.67}$$

From i) and (3.67), we have

$$\mathcal{I}\left(\|a\|_\infty \rho^{-j(\|a\|_\infty)}\right) = 1.$$

From iii) and the equation above, then

$$\left\|\mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right)\right\|_\infty = \mathcal{I}\left(\|a\|_\infty \rho^{-j(\|a\|_\infty)}\right) = 1, \tag{3.68}$$

establishing iv) and concluding the proof. □

Taking advantage of Lemma 3.19, we next give a useful interpretation of $\mathring{q}$ by analysing its quantisation regions and corresponding values. Recall (3.58)–(3.61) and consider an arbitrary nonzero $a\in\mathbb{R}^p$. Then, $\|a\|_\infty>0$ and thus note that $j(\|a\|_\infty)$ is well defined. We first consider the case when a satisfies $j(\|a\|_\infty)=0$. Lemma 3.19 i) shows that $j(\|a\|_\infty)=0$ if and only if $\frac{\mathrm{C}-2}{\mathrm{C}-1}<\|a\|_\infty\leq\frac{\mathrm{C}}{\mathrm{C}-1}$. The set of all $a\in\mathbb{R}^p$ that satisfy $j(\|a\|_\infty)=0$ is thus a region contained between the two (hyper)cubical surfaces in $\mathbb{R}^p$ of equations $\|a\|_\infty=\frac{\mathrm{C}-2}{\mathrm{C}-1}$ and $\|a\|_\infty=\frac{\mathrm{C}}{\mathrm{C}-1}$. Recalling (3.60), note that $\frac{\mathrm{C}-2}{\mathrm{C}-1}=\rho\frac{\mathrm{C}}{\mathrm{C}-1}$. By (3.61b), if $j(\|a\|_\infty)=0$ then $\mathring{q}(a)=\mathcal{I}(a)$. Then, Lemma 3.19 iv) shows that $\|\mathcal{I}(a)\|_\infty=1$ and hence $\|\mathring{q}(a)\|_\infty=1$. Therefore, if $j(\|a\|_\infty)=0$ then $\mathring{q}(a)$ is located on the unit (hyper)cube. Note that, since $\|a\|_\infty\leq\frac{\mathrm{C}}{\mathrm{C}-1}$, then $|a_i|\leq\frac{\mathrm{C}}{\mathrm{C}-1}$, for $i=1,\ldots,p$. From (3.58) and Figure 3.3, then $\mathcal{I}(a_i)$ can only take one of C different values, namely $0,\pm\frac{2}{\mathrm{C}-1},\ldots,\pm\frac{\mathrm{C}-1}{\mathrm{C}-1}$ (recall that C is an odd integer). In addition,

$|\mathcal{I}(a_i)| = \mathcal{I}(|a_i|) = 1$ for at least one value of i in $\{1, \ldots, p\}$, because $\|\mathcal{I}(a)\|_\infty = \mathcal{I}(\|a\|_\infty) = 1$. Note then that the number of different values $\mathcal{I}(a)$ is finite whenever $j(\|a\|_\infty) = 0$. Figure 3.4 a) shows the region $j(\|a\|_\infty) = 0$, for $p = 2$. Figure 3.4 b) shows the quantisation regions and values of $\mathring{q}$, whenever a is such that $j(\|a\|_\infty) = 0$ and for $\mathrm{C} = 5$. Note that, if $\mathcal{I}(a_1) = 1$ and $j(\|a\|_\infty) = 0$, then $\mathcal{I}(a_2)$ only takes one of $\mathrm{C} = 5$ different values [see the shaded set in Figure 3.4 b)].

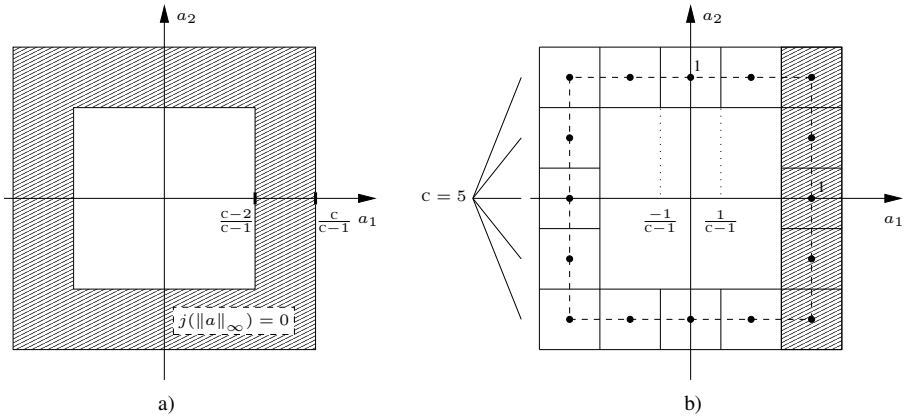


Figure 3.4: a) The set $\{a \in \mathbb{R}^2 : j(\|a\|_\infty) = 0\}$. b) Values (circles) and regions (solid-line squares) of $\mathring{q}$ for $a : j(\|a\|_\infty) = 0$, the unit (hyper)cube (dashed line) and the set $\{a \in \mathbb{R}^2 : \mathcal{I}(a_1) = 1, j(\|a\|_\infty) = 0\}$ (shaded).

We next analyse the quantisation regions and corresponding values of $\mathring{q}$ when $a \in \mathbb{R}^p$ is nonzero but otherwise arbitrary, that is, when it does not necessarily satisfy $j(\|a\|_\infty) = 0$. By Lemma 3.19 iv) and iii), we have

$$\left\|\mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right)\right\|_\infty = 1 = \mathcal{I}\left(\|a\|_\infty \rho^{-j(\|a\|_\infty)}\right),$$

and by i), then $j\left(\|a\|_\infty \rho^{-j(\|a\|_\infty)}\right) = 0$. Therefore, multiplying an arbitrary nonzero $a \in \mathbb{R}^p$ by $\rho^{-j(\|a\|_\infty)}$ yields a point $\tilde{a} = a\rho^{-j(\|a\|_\infty)}$ satisfying $j(\|\tilde{a}\|_\infty) = 0$. Hence, $\tilde{a}$ lies in the set depicted in Figure 3.4 a) and we have already analysed the possible values $\mathcal{I}(\tilde{a})$. To produce $\mathring{q}(a)$, recall from (3.61b) that $\mathcal{I}(\tilde{a})$ is multiplied by $\rho^{j(\|a\|_\infty)}$, where, by (3.59), j takes only integer values. Figure 3.5 depicts the functions $j(b)$ and $\rho^{j(b)}$ for $\mathrm{C} = 5$. Figure 3.6 a) shows the regions corresponding to $j(\|a\|_\infty) = 0$ (where $\mathring{q}(a) = \mathcal{I}(a)$), $j(\|a\|_\infty) = 1$ (where $\mathring{q}(a) = \rho\mathcal{I}(a\rho^{-1})$) and $j(\|a\|_\infty) = 2$ (where $\mathring{q}(a) = \rho^2\mathcal{I}(a\rho^{-2})$). Finally, Figure 3.6 b) depicts the quantisation regions and values of $\mathring{q}$. The quantisation regions are depicted by the solid-line squares and their corresponding values are the centres of the squares, shown as circles. This figure coincides with Figure 3.2, as expected.

We have now gained some insight into the quantiser $\mathring{q}$ defined in (3.61). Before proceeding with the design of a QS quantiser for system (3.1), we require two additional results.

Lemma 3.20 *Let* $\mathrm{C} \geq 3$ *be an odd integer and let* $\mathcal{U}(\mathring{q})$ *be the range of the quantiser* $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$

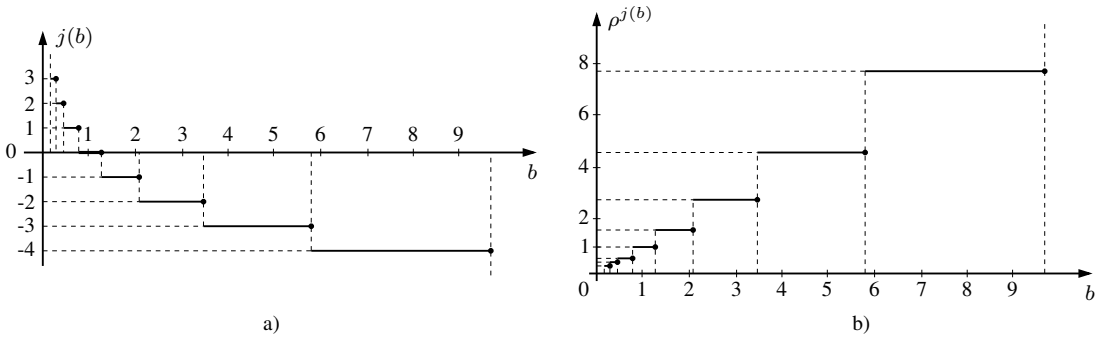


Figure 3.5: The functions a) $j(b)$ and b) $\rho^{j(b)}$ for $\mathrm{C} = 5$.

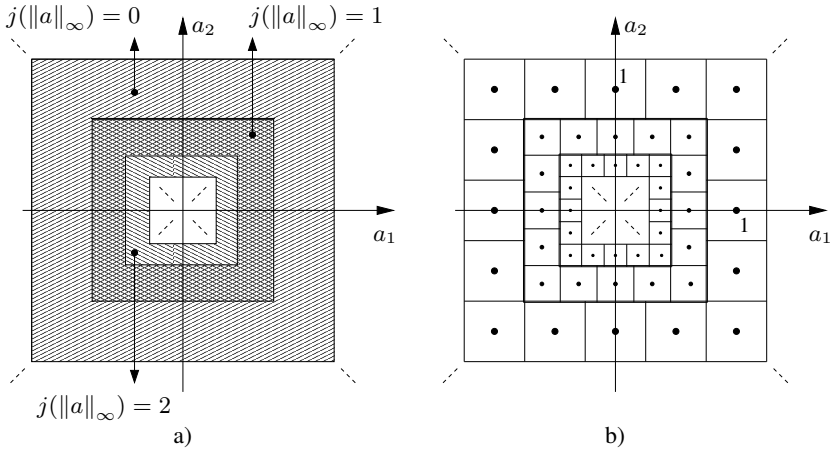


Figure 3.6: a) Regions where $j(\|a\|_\infty) = 0$, $j(\|a\|_\infty) = 1$ and $j(\|a\|_\infty) = 2$. b) Quantisation regions (solid-line squares) and values (circles) of $\mathring{q}$.

defined in (3.61)*. Then,*

$$\mathcal{U}(\mathring{q}) = \biguplus_{i=1}^{\mathrm{C}^p-(\mathrm{C}-2)^p} \{\rho^k \bar{u}_i : k \in \mathbb{Z}\} \cup \{0\}, \tag{3.69}$$

where ρ was defined in (3.60) *and $\bar{u}_i \neq 0$, for $i = 1, \ldots, \mathrm{C}^p - (\mathrm{C}-2)^p$.*

Proof. From (3.61), we have

$$\mathcal{U}(\mathring{q}) = \left\{\rho^{j(\|a\|_\infty)} \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right) : a \in \mathbb{R}^p \setminus \{0\}\right\} \cup \{0\}. \tag{3.70}$$

Fix a nonzero $\bar{a} \in \mathbb{R}^p$. Note that $\left\|\rho^k \bar{a}\right\|_\infty = \rho^k \left\|\bar{a}\right\|_\infty$. By Lemma 3.19 ii), then $j(\left\|\rho^k \bar{a}\right\|_\infty) = j(\rho^k \left\|\bar{a}\right\|_\infty) = j(\left\|\bar{a}\right\|_\infty) + k$. Note then that $\rho^k \bar{a} \rho^{-j(\|\rho^k \bar{a}\|_\infty)} = \bar{a}\rho^{-j(\|\bar{a}\|_\infty)}$ and $\rho^{j(\left\|\rho^k \bar{a}\right\|_\infty)} = \rho^k \rho^{j(\|\bar{a}\|_\infty)}$, for all $k \in \mathbb{Z}$. Therefore, we can rewrite (3.70) as

$$\mathcal{U}(\mathring{q}) = \left\{\rho^k \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right) : k \in \mathbb{Z}, a \in \mathbb{R}^p \setminus \{0\}\right\} \cup \{0\}.$$

From Lemma 3.19 iv), we know that $\left\|\mathcal{I}(a\rho^{-j(\|a\|_\infty)})\right\|_\infty = 1$, for all nonzero $a \in \mathbb{R}^p$. In addition, given any $c \in \mathbb{R}^p$ such that $\left\|\mathcal{I}(c)\right\|_\infty = 1$, then $c \neq 0$ and there exists $a \in \mathbb{R}^p \setminus \{0\}$ such that

$\mathcal{I}(c) = \mathcal{I}(a\rho^{-j(\|a\|_\infty)})$. Indeed, using Lemma 3.19 iii), we have $\|\mathcal{I}(c)\|_\infty = \mathcal{I}(\|c\|_\infty) = 1$ and by Lemma 3.19 i), it follows that $j(\|c\|_\infty) = 0$. Therefore, $c\rho^{-j(\|c\|_\infty)} = c$, showing that we may take $a = c$. Hence, it follows that

$$\mathcal{U}(\mathring{q}) = \left\{\rho^k \mathcal{I}(a) : k \in \mathbb{Z}, a \in \mathbb{R}^p, \|\mathcal{I}(a)\|_\infty = 1\right\} \cup \{0\}. \tag{3.71}$$

Define the sets

$$C_= \triangleq \{\mathcal{I}(a) : a \in \mathbb{R}^p, \|\mathcal{I}(a)\|_\infty = 1\}, \tag{3.72}$$

$$C_\leq \triangleq \{\mathcal{I}(a) : a \in \mathbb{R}^p, \|\mathcal{I}(a)\|_\infty \leq 1\}, \tag{3.73}$$

$$C_< \triangleq \{\mathcal{I}(a) : a \in \mathbb{R}^p, \|\mathcal{I}(a)\|_\infty < 1\}. \tag{3.74}$$

Consider $C_\leq$ in (3.73). Note that $\|\mathcal{I}(a)\|_\infty \leq 1$ if and only if $|\mathcal{I}(a_i)| \leq 1$, for all $i = 1, \ldots, p$. Then,

$$C_\leq = \{\mathcal{I}(a) : a \in \mathbb{R}^p, |\mathcal{I}(a_i)| \leq 1 \text{ for all } i = 1, \ldots, p\}.$$

Define

$$E_\leq \triangleq \{\mathcal{I}(b) : b \in \mathbb{R}, |\mathcal{I}(b)| \leq 1\}.$$

From (3.58), we have $\#E_\leq = \mathrm{C}$ (see also Figure 3.3). Note then that $\#C_\leq = (\#E_\leq)^p = \mathrm{C}^p$. Also, define

$$E_< \triangleq \{\mathcal{I}(b) : b \in \mathbb{R}, |\mathcal{I}(b)| < 1\},$$

and note from (3.58) that $\#E_< = \mathrm{C} - 2$. Moreover, from (3.74) and using the same reasoning as for $C_\leq$, it follows that $\#C_< = (\#E_<)^p = (\mathrm{C} - 2)^p$. From (3.72)–(3.74), note that $C_= = C_\leq \setminus C_<$ and $C_< \subset C_\leq$. Therefore, $\#C_= = \#C_\leq - \#C_< = \mathrm{C}^p - (\mathrm{C} - 2)^p$ and using (3.71) we can write

$$\mathcal{U}(\mathring{q}) = \bigcup_{i=1}^{\mathrm{C}^p - (\mathrm{C}-2)^p} \{\rho^k \bar{u}_i : k \in \mathbb{Z}\} \cup \{0\}, \tag{3.75}$$

where $\bar{u}_i$, for $i = 1, \ldots, \mathrm{C}^p - (\mathrm{C} - 2)^p$, are the elements of $C_=$. Note from (3.72) that $\bar{u}_i \neq 0$, for $i = 1, \ldots, \mathrm{C}^p - (\mathrm{C} - 2)^p$. To prove that the sets in (3.75) are disjoint, suppose that $\rho^k \bar{u}_i = \rho^{k'} \bar{u}_{i'}$, where $\bar{u}_i$ and $\bar{u}_{i'}$ are two elements of $C_=$. Then, $\bar{u}_i = \rho^{k'-k} \bar{u}_{i'}$ and $\|\bar{u}_i\|_\infty = \rho^{k'-k} \|\bar{u}_{i'}\|_\infty$. Note that the $\bar{u}_i$, which are the elements of $C_=$, all satisfy $\|\bar{u}_i\|_\infty = 1$. Hence, $\rho^{k'-k} = 1$, whence $k' = k$ and then $\bar{u}_i = \bar{u}_{i'}$. Therefore, $\rho^k \bar{u}_i = \rho^{k'} \bar{u}_{i'}$ only if $\bar{u}_i = \bar{u}_{i'}$, showing that the sets in (3.75) are disjoint, and establishing (3.69). □

Lemma 3.20 shows that the range of $\mathring{q}$ is a finite disjoint union of radially logarithmically spaced values.

Lemma 3.21 *Let* $\mathrm{C} > 2$ *and consider the quantiser* $\mathring{q}$ *defined above in* (3.61)*. Then,*

$$\|a - \mathring{q}(a)\|_\infty \leq \frac{\rho^{j(\|a\|_\infty)}}{\mathrm{C} - 1}, \tag{3.76}$$

for all nonzero $a \in \mathbb{R}^p$.

Proof. From (3.58), note that for all $b \in \mathbb{R}$,

$$|b - \mathcal{I}(b)| \leq \frac{1}{\mathrm{c} - 1}. \tag{3.77}$$

From (3.61), we have

$$\begin{aligned}\|a - \mathring{q}(a)\|_\infty &= \left\|\rho^{j(\|a\|_\infty)} a\rho^{-j(\|a\|_\infty)} - \rho^{j(\|a\|_\infty)}\mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right)\right\|_\infty \\ &= \rho^{j(\|a\|_\infty)}\left\|a\rho^{-j(\|a\|_\infty)} - \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right)\right\|_\infty \\ &\leq \frac{\rho^{j(\|a\|_\infty)}}{\mathrm{c} - 1},\end{aligned}$$

where the last line above follows from (3.77). □

3.5.2 QS Quantiser with Finite Density

We next show how to design a QS quantiser with finite density employing the quantiser $\mathring{q}$ constructed in §3.5.1.

Theorem 3.22 *Let $D \in \mathbb{R}^{n\times\ell}$ have linearly independent columns and satisfy $D^T L^{-1} D = \mathrm{I}_\ell$, where L was defined in (3.8) and ℓ is the number of positive eigenvalues of L. Let $W \in \mathbb{R}^{m\times\ell}$ be such that $S \triangleq D^T L^{-1} M W = -\mathrm{I}_\ell$, with M as defined in (3.8), and such that $J \triangleq -W^T H W > 0$, where H was defined in (3.31). Let λ denote the smallest eigenvalue of the matrix J, and let c be an odd integer satisfying $\mathrm{c} \geq 3$ and*

$$\mathrm{c} > 1 + \sqrt{\ell/\lambda}. \tag{3.78}$$

Then, the quantiser q defined by

$$q(x) = W\mathring{q}\left(D^T x\right), \tag{3.79}$$

with $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ as defined in (3.61) *and* (3.58)–(3.60) *in §3.5.1 setting $p = \ell$, is QS and has a quantisation density given by*

$$\eta(q) = \frac{\mathrm{c}^\ell - (\mathrm{c}-2)^\ell}{\ln \mathrm{c} - \ln(\mathrm{c}-2)}. \tag{3.80}$$

Proof. We begin by proving that q is QS. Note that conditions i)—ii) of Theorem 3.14 are satisfied.

Let $\mathcal{R}_0$ denote the quantisation region of q that contains the origin. From (3.79) and (3.61a), we have $q(0) = 0$ and hence the value of q corresponding to $\mathcal{R}_0$ is $u = 0$. Note that since S is invertible, then the matrix $W \in \mathbb{R}^{m\times\ell}$ has full column rank. From (3.79), then $q(x) = 0$ if and only if $\mathring{q}(D^T x) = 0$ which by (3.61) happens if and only if $D^T x = 0$. Consequently, condition i) of Theorem 3.17 is satisfied. In addition, note that $\mathcal{A}_0 = \{0\}$ is a quantisation region of $\mathring{q}$ with corresponding value $\bar{u}_0 = 0$.

Let $\mathcal{A}_i$ and $\bar{u}_i$, for all $i \in \mathbb{Z} \setminus \{0\}$, denote the remaining quantisation regions and corresponding values of $\mathring{q}$. For all $a \in \mathcal{A}_i$, we have, since $D^T L^{-1} D = \mathrm{I}_\ell$ and $S = -\mathrm{I}_\ell$,

$$(a + S\bar{u}_i)^T (D^T L^{-1} D)^{-1} (a + S\bar{u}_i) = (a - \bar{u}_i)^T (a - \bar{u}_i) = \|a - \mathring{q}(a)\|_2^2, \tag{3.81}$$

for all $i \in \mathbb{Z} \setminus \{0\}$. In addition, using an inequality relating the two-norm and the infinity-norm, and applying Lemma 3.21, yields

$$\|a - \mathring{q}(a)\|_2^2 \leq \ell \, \|a - \mathring{q}(a)\|_\infty^2 \leq \frac{\ell}{(\mathrm{C}-1)^2} \rho^{2j(\|a\|_\infty)}. \tag{3.82}$$

From (3.78), it follows that $(\mathrm{C}-1)^2 > \ell/\lambda$. Hence, from (3.82), we have

$$\|a - \mathring{q}(a)\|_2^2 < \lambda \rho^{2j(\|a\|_\infty)}. \tag{3.83}$$

Also, since λ is the smallest eigenvalue of $J > 0$, then $\bar{u}_i^T J \bar{u}_i \geq \lambda \, \|\bar{u}_i\|_2^2$. Using $\bar{u}_i = \mathring{q}(a)$ for all $a \in \mathcal{A}_i$, where $\mathring{q}(a)$ is given by (3.61b), then

$$\bar{u}_i^T J \bar{u}_i \geq \lambda \rho^{2j(\|a\|_\infty)} \left\| \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right) \right\|_2^2 \geq \lambda \rho^{2j(\|a\|_\infty)} \left\| \mathcal{I}\left(a\rho^{-j(\|a\|_\infty)}\right) \right\|_\infty^2. \tag{3.84}$$

Using Lemma 3.19 iv) in (3.84), we obtain

$$\bar{u}_i^T J \bar{u}_i \geq \lambda \rho^{2j(\|a\|_\infty)}. \tag{3.85}$$

Combining (3.81), (3.83) and (3.85), yields

$$(a + S\bar{u}_i)^T (D^T L^{-1} D)^{-1} (a + S\bar{u}_i) < \bar{u}_i^T J \bar{u}_i, \tag{3.86}$$

for all $a \in \mathcal{A}_i$, for all $i \in \mathbb{Z} \setminus \{0\}$. Therefore, Theorem 3.17 shows that q is QS.

We next establish (3.80). Let $\bar{q} : \mathbb{R}^n \to \mathbb{R}^\ell$ be the quantiser defined by $\bar{q}(x) = \mathring{q}(D^T x)$. Since D has linearly independent columns, then $\mathcal{U}(\bar{q}) \triangleq \{\bar{q}(x) : x \in \mathbb{R}^n\}$ and $\mathcal{U}(\mathring{q}) \triangleq \{\mathring{q}(a) : a \in \mathbb{R}^\ell\}$ are equal. Therefore, $\eta(\bar{q}) = \eta(\mathring{q})$ because quantisation density depends only on the range of a quantiser [recall (2.12)–(2.14) in Chapter 2]. From (3.79) and since W has linearly independent columns, then Lemma 2.15 establishes that $\eta(q) = \eta(\mathring{q})$. Then, using Lemma 3.20 and Theorem 2.12, yields

$$\eta(q) = \frac{\mathrm{C}^\ell - (\mathrm{C}-2)^\ell}{-\ln \rho}.$$

Eq. (3.80) then follows by substituting (3.60) into the equation above. □

Remark 3.23 *The conditions imposed by Theorem 3.22 on the matrices W and D, in addition to conditions i)—ii) of Theorem 3.14, incur no loss of generality. Indeed, suppose that we would like to design a quantiser $\tilde{q}(x) = \tilde{W}\mathring{\tilde{q}}(\tilde{D}^T x)$, where $\tilde{W}$ and $\tilde{D}$ satisfy conditions i)—ii) of Theorem 3.14 but are otherwise arbitrary. Theorem 3.22 can then be employed to design a quantiser $\tilde{q}(x) = \tilde{W}\mathring{\tilde{q}}(\tilde{D}^T x)$ by using $D = \tilde{D}(\tilde{D}^T L^{-1} \tilde{D})^{-1/2}$ and $W = -\tilde{W}\tilde{S}^{-1}(\tilde{D}^T L^{-1} \tilde{D})^{1/2}$, where $\tilde{S} \triangleq \tilde{D}^T L^{-1} M \tilde{W}$. Note then that $D^T L^{-1} D = \mathrm{I}_\ell$ and $D^T L^{-1} M W = -\mathrm{I}_\ell$, and Theorem 3.22 yields a finite-density QS quantiser $q(x) = W\mathring{q}(D^T x) = \tilde{W}\mathring{\tilde{q}}(\tilde{D}^T x)$, where $\mathring{\tilde{q}}(a) = -\tilde{S}^{-1}(\tilde{D}^T L^{-1} \tilde{D})^{1/2} \mathring{q}((\tilde{D}^T L^{-1} \tilde{D})^{-1/2} a)$.*

Theorem 3.22 provides a means to design finite-density quantisers that quadratically stabilise a system with respect to the given CLF. Moreover, the densities of these quantisers are given explicitly by (3.80). Eq. (3.80) reveals that in order to minimise the density of a quantiser constructed according to Theorem 3.22 for a given CLF, the value of C needs to be as low as possible. This follows since ℓ is constrained to be equal to the number of positive eigenvalues of $L = A^T P A - P$, which does not change if both the system and the CLF are fixed. On the other hand, the parameter C needs to be high enough to satisfy both $\mathrm{C} \geq 3$ and (3.78). Therefore, the best choice for C in terms of quantisation density is the least odd integer that satisfies both $\mathrm{C} \geq 3$ and (3.78). The constraint $\mathrm{C} \geq 3$ is imposed by the geometry of the construction of §3.5.1. On the other hand, the constraint (3.78) is required to yield a QS quantiser. If $1 + \sqrt{\ell/\lambda} > 3$, then maximising λ may allow the use of lower values of C. Recall that λ is the smallest eigenvalue of $J = -W^T H W$ and hence different matrices W may yield different values of λ. When W and D satisfy the conditions of Theorem 3.22, pre- and post-multiplying (3.42) in Lemma 3.13 by W^T and W, respectively, yields $J < \mathrm{I}$, whence $\lambda < 1$. Therefore, it follows that if $\ell \geq 4$, then $1 + \sqrt{\ell/\lambda} > 3$ and if C satisfies (3.78) then C will automatically satisfy $\mathrm{C} \geq 3$. This implies that when the number ℓ of positive eigenvalues of L is greater than 4, then constructing $\mathring{q}$ according to §3.5.1 does not directly impose a lower limit on the achievable quantisation density.

We next summarise the steps involved in the suggested construction of a finite-density multivariable QS quantiser.

Finite-density QS Quantiser Construction

(a) Given the system and CLF matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $P \in \mathbb{R}^{n\times n}$ such that A is unstable, (A, B) is stabilisable and $P = P^T > 0$.

(b) Compute L and M from (3.8), and Q from (3.10).

(c) Verify that $V(x) = x^T P x$ is a CLF by checking that $Q > 0$ (Lemma 2.4)[2].

(d) Verify that L is nonsingular[3].

(e) Compute ℓ, the number of positive eigenvalues of L.

(f) Compute H from (3.31).

(g) Choose $\tilde{D} \in \mathbb{R}^{n\times \ell}$ and $\tilde{W} \in \mathbb{R}^{m\times \ell}$ satisfying $\tilde{D}^T L^{-1} \tilde{D} > 0$ and $\tilde{W}^T H \tilde{W} < 0$. These matrices can be computed, for example, by factoring the matrix G in the proof of Theorem 3.2 as $G = \tilde{W}\tilde{D}^T$ (see also the example in §3.5.3 below).

[2] If $Q \not> 0$ then choose another P and start again.

[3] If L is singular, then our method is not applicable. Choosing a different P may yield a nonsingular L. However, in some cases no choice of P will cause L to be nonsingular (for example, if $A = \mathrm{I}$). This is a limitation of the proposed method.

(h) Compute $D = \tilde{D}(\tilde{D}^T L^{-1}\tilde{D})^{-1/2}$ and $W = -\tilde{W}\tilde{S}^{-1}(\tilde{D}^T L^{-1}\tilde{D})^{1/2}$, where $\tilde{S} = \tilde{D}^T L^{-1} M\tilde{W}$. We now have $D^T L^{-1} D = \mathrm{I}_\ell$ and $D^T L^{-1} MW = -\mathrm{I}_\ell$ (recall Remark 3.23).

(i) Compute $J = -W^T HW$ and its smallest eigenvalue, λ.

(j) Choose an odd integer $\mathrm{C} \geq 3$ satisfying (3.78)[4].

(k) Consider the quantiser $\mathring{q} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ defined in (3.61) and (3.58)–(3.60) in § 3.5.1.

(l) The required finite-density QS quantiser $q : \mathbb{R}^n \to \mathbb{R}^m$ is defined by $q(x) = W\mathring{q}(D^T x)$. Its density is given by (3.80).

3.5.3 Example

Consider system (3.1) with matrices

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3.87}$$

and the quadratic function $V(x) = x^T Px$ with

$$P = \begin{bmatrix} 16.6561 & -7.2172 & -15.3227 & -0.1282 & -0.08183 \\ -7.2172 & 4.4423 & 7.2172 & 0.04159 & 0.02651 \\ -15.3227 & 7.2172 & 16.6561 & 0.1282 & 0.08183 \\ -0.1282 & 0.04159 & 0.1282 & 21.5906 & 10.7047 \\ -0.08183 & 0.02651 & 0.08183 & 10.7047 & 6.6030 \end{bmatrix}. \tag{3.88}$$

We would like to design a quantiser $q : \mathbb{R}^5 \to \mathbb{R}^4$ with finite density that renders V a Lyapunov function for the closed-loop system $x(k+1) = Ax(k) + Bq(x(k))$.

We hence follow steps (a)–(l) above. We readily check that the matrices A, B and P given above are such that A is unstable, (A, B) is stabilisable and $P = P^T > 0$ [step (a)]. We compute $L = A^T PA - P$, $M = A^T PB$ and $Q = M(B^T PB)^{-1} M^T - L$ [step (b)]. We check that $Q > 0$ and hence we verify that the given V is a CLF [step (c)]. We next check that L is nonsingular [step (d)] and compute the number of positive eigenvalues of L [step (e)]. This yields $\ell = 3$, which is the dimension required for the reduced-dimension quantiser. We compute $H = B^T PB - M^T L^{-1} M$ [step (f)].

The next step [step (g)] is to find matrices $\tilde{W} \in \mathbb{R}^{4\times 3}$ and $\tilde{D} \in \mathbb{R}^{5\times 3}$ satisfying $\tilde{D}^T L^{-1}\tilde{D} > 0$ and $\tilde{W}^T H\tilde{W} < 0$. We may find these matrices as follows. From Lemma 3.1, we know that if a

[4] In terms of quantisation density, the smaller C is, the better (recall discussion on page 58).

linear feedback $u = \tilde{W}\tilde{D}^T x$, with $\tilde{W}$ and $\tilde{D}$ having the required dimensions and linearly independent columns, quadratically stabilises the system with respect to V, then there exists a quantised feedback of the form $u = \tilde{W}\mathring{q}(\tilde{D}^T x)$ that is QS with respect to V. It then follows by Theorem 3.14 that such $\tilde{W}$ and $\tilde{D}$ satisfy $\tilde{D}^T L^{-1}\tilde{D} > 0$ and $\tilde{W}^T H\tilde{W} < 0$. We then recall Remark 3.5 and calculate the matrix G in the proof of Theorem 3.2. Since such G has rank equal to the number of positive eigenvalues of L, we may factor G as $G = \tilde{W}\tilde{D}^T$, where $\tilde{W}$ and $\tilde{D}$ have the required dimensions and properties. We hence compute $K = (B^T P B)^{-1/2} M^T$, we find the singular value decomposition $KQ^{-1/2} = S_1 \Sigma S_2^T$, and compute, from (3.12):

$$\tilde{W} = -(B^T P B)^{-1/2} S_1 \begin{bmatrix} \Sigma_{1:3} \\ \mathbf{0}_{1\times 3} \end{bmatrix}, \qquad \tilde{D}^T = \begin{bmatrix} \mathrm{I}_3 & \mathbf{0}_{3\times 1} \end{bmatrix} S_2^T Q^{1/2}. \tag{3.89}$$

This concludes step (g).

We proceed with step (h) and compute

$$D = \tilde{D}(\tilde{D}^T L^{-1}\tilde{D})^{-1/2} \quad \text{and} \quad W = -\tilde{W}\tilde{S}^{-1}(\tilde{D}^T L^{-1}\tilde{D})^{1/2}, \tag{3.90}$$

where $\tilde{S} \triangleq \tilde{D}^T L^{-1} M \tilde{W}$.

Computing $J = -W^T H W$ and calculating its smallest eigenvalue yields $\lambda \approx 0.01767$ [step (i)]. The least odd integer C that satisfies $\mathrm{C} \geq 3$ and (3.78), where $\ell = 3$, is $\mathrm{C} = 15$ [step (j)]. We then implement the quantised feedback $u(k) = q(x(k)) = W\mathring{q}(D^T x(k))$, with $\mathring{q}$ as described in (3.61), and (3.58)–(3.60) in §3.5.1 [steps (k) and (l)]. The design of a finite-density QS quantiser for the given system is now complete.

We simulated the resulting closed-loop system from the arbitrary initial condition

$$x(0) = \begin{bmatrix} -0.3911 & 0.4368 & 0.0111 & -0.2887 & -0.0789 \end{bmatrix}^T.$$

The simulation results are shown in Figures 3.7–3.9. Figures 3.7 and 3.8 show the evolution of the five components of the system state and of the three components of the output of the reduced-dimension quantiser, respectively. Figure 3.9 shows the evolution of $V(x) = x^T P x$. Note that $V(x(k))$ decreases as the time k increases, verifying that the quantiser $q(x) = W\mathring{q}(D^T x)$ is QS.

Remark 3.24 *At step (g), any matrices $\tilde{W} \in \mathbb{R}^{m\times\ell}$ and $\tilde{D} \in \mathbb{R}^{n\times\ell}$ satisfying $\tilde{D}^T L^{-1}\tilde{D} > 0$ and $\tilde{W}^T H\tilde{W} < 0$ can be chosen. Choosing these matrices by factoring the matrix G in the proof of Theorem 3.2, as suggested, yields a matrix D of the form* (3.6) *at step (h), as follows from* (3.89) *and* (3.90)*. This choice is advantageous in relation to quantisation density (recall Theorem 2.17 in Chapter 2).*

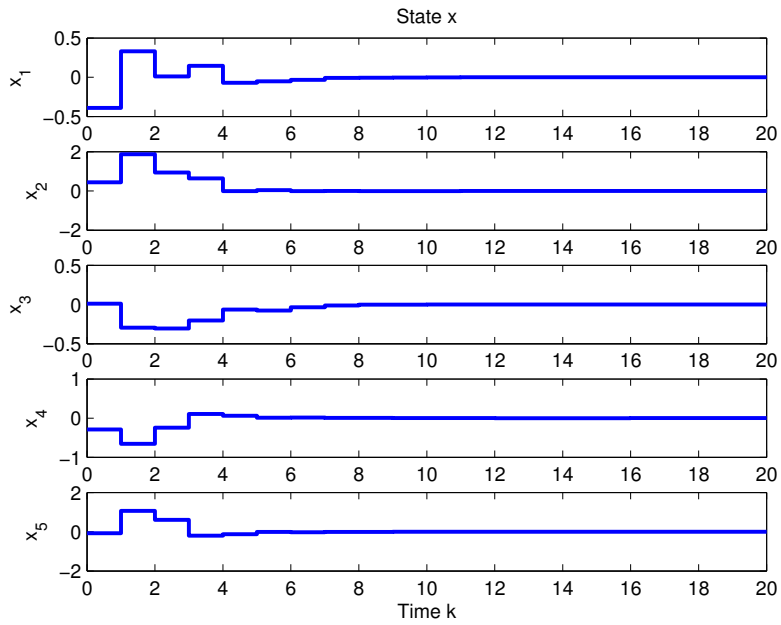


Figure 3.7: System state: x.

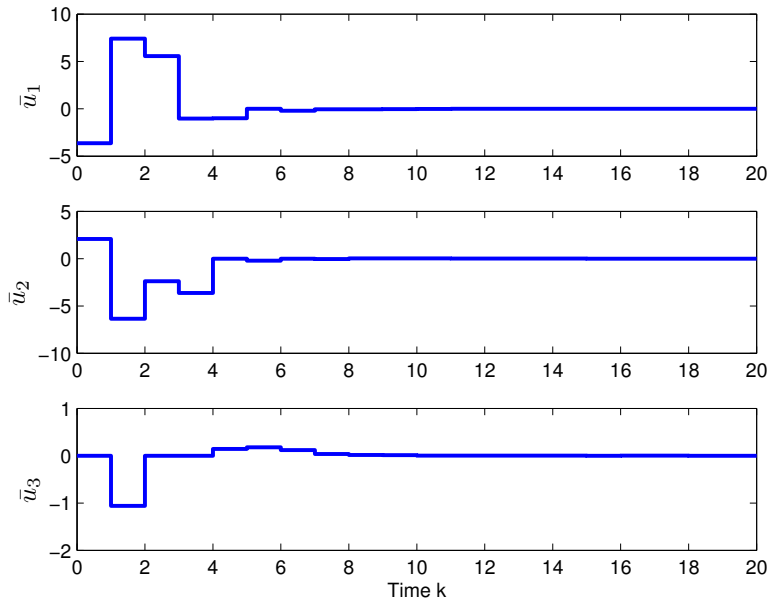


Figure 3.8: Reduced-dimension quantiser output: $\bar{u} = \mathring{q}(D^T x)$.

3.6 Chapter Summary

In this chapter, we have derived an explicit geometric characterisation of quadratically stabilising state feedback laws that are based on the use of multivariable quantisers of minimum dimension. This charac-

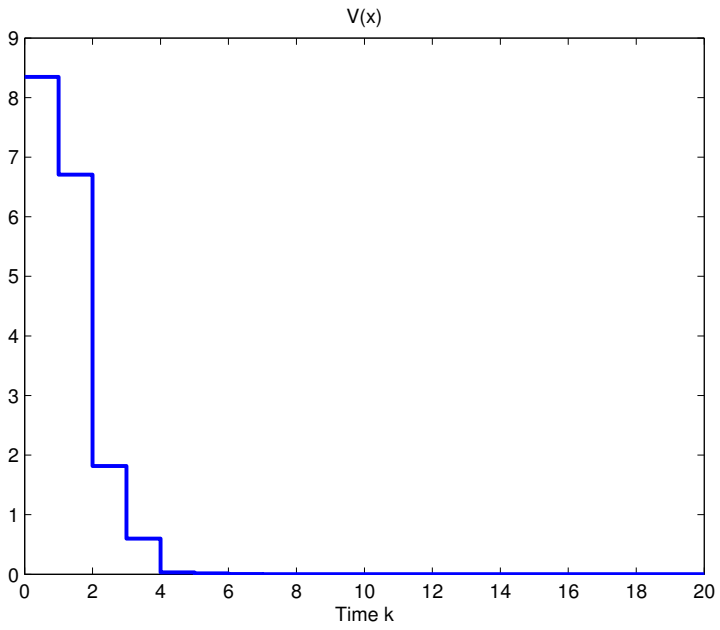


Figure 3.9: $V(x)$.

terisation consists in a set of necessary and sufficient conditions for a quantised static state feedback to render a given quadratic function a Lyapunov function for the closed-loop system. These necessary and sufficient conditions, derived in Theorems 3.14 and 3.17, provide a means to analyse and design such quantised feedback laws and were derived from a set inclusion condition that is necessarily satisfied by the quantisation regions and values of a quadratically stabilising quantiser.

We have then designed quantisers with finite density that are able to quadratically stabilise a multiple-input system with respect to a given CLF. We believe that this is the first method that has been proposed for explicitly constructing quantisers having *finite density* and that are able to quadratically stabilise systems having *an arbitrary number of inputs*. The design of such quantisers required results from Chapter 2, jointly with the necessary and sufficient conditions derived in this chapter in Theorems 3.14 and 3.17.

Chapter 4

State-space Approach to Quantiser "Coarseness" for Single-input Systems

4.1 Overview

In Chapters 2 and 3, we have analysed the problem of deriving a least dense quantiser over all quantisers that quadratically stabilise a given system. As we have seen, the density of a quantiser depends on the separation (in some sense) of its quantisation *values*. In the setting that we consider, the values of a quantiser constitute the control input values that are applied to the system, since we have $u = q(x)$. In this sense, we may say that the concept of quantisation density considered so far is input-space-based.

By contrast, in this chapter we explore a different notion of quantisation density, based on the separation of the quantisation *regions* of a quantiser, rather than of its values. The derivations of this chapter will focus exclusively on single-input systems. To avoid confusion with the standard quantisation density concept, we shall refer to the concept in this chapter as quantisation coarseness. The intuitive idea behind our exploration is that a quantiser is coarse (least dense) if its quantisation regions are as large as possible. We will introduce a novel type of quantisers, namely CAQS (Coarse-Almost-Quadratically-Stabilising) quantisers, and analyse different links between the proposed state-space approach and the input-space-based quantisation density concept considered in previous chapters.

The remainder of this chapter is organised as follows. In §4.2, we particularise the setting considered to single-input systems and define the type of quantised feedback that will be employed. In §4.3, we provide some preliminary results. The main results of the chapter are presented in §4.4 and §4.5, where we explore quantiser coarseness from a state-space standpoint, introduce CAQS quantisers and analyse the connections with the standard concept of quantisation density. We present a quantiser design example in §4.6 and a summary in §4.7.

4.2 Single-input Systems

We consider a single-input system of the form

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

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $x \in \mathbb{R}^n$ and $u \in \mathbb{R}$. As in previous chapters, we deal with quadratic stabilisation of the system by means of quantised static feedback $u = q(x)$, where $q : \mathbb{R}^n \to \mathbb{R}$. We will exclusively focus on quantisers $q : \mathbb{R}^n \to \mathbb{R}$ of the specific form

$$q(x) = \mathring{q}(d^T x), \tag{4.2}$$

where $\mathring{q} : \mathbb{R} \to \mathbb{R}$ is a scalar quantiser and $d \in \mathbb{R}^n$. Figure 4.1 shows the setting that we consider.

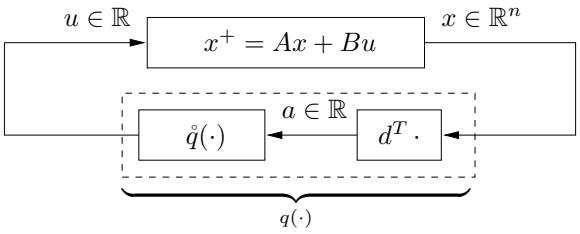


Figure 4.1: The quantised feedback considered: $u = q(x) = \mathring{q}(d^T x)$.

Remark 4.1 *When dealing with single-input systems, the quantised feeback structure $u = q(x) = \mathring{q}(d^T x)$ coincides with the structure considered in Chapter 3. For a single-input system, Chapter 3 considers quantised feedbacks $u = q(x) = W\mathring{q}(D^T x)$, with $W \in \mathbb{R}^{1\times p}$ and $D \in \mathbb{R}^{n\times p}$ both having linearly independent columns, $\mathring{q} : \mathbb{R}^p \to \mathbb{R}^p$, and p being as low as possible. For open-loop-unstable single-input systems, the minimum value of p is 1, and hence $W \in \mathbb{R}$, $D \in \mathbb{R}^n$ and $\mathring{q} : \mathbb{R} \to \mathbb{R}$. Note then that there is no loss of generality in selecting $W = 1$, hence yielding the structure $u = \mathring{q}(D^T x) = \mathring{q}(d^T x)$, which we consider in this chapter.*

Since q satisfies $q(x) = \mathring{q}(d^T x)$, for all $x \in \mathbb{R}^n$, the quantisation regions of q have a specific shape. We will refer to the quantisation regions of such q as *parallel-hyperplane* regions and to q as a *parallel-hyperplane* quantiser. This terminology is motivated by the fact that the quantisation regions of q are all limited by parallel hyperplanes. We will thus employ the following definitions.

Definition 4.2 (Parallel-hyperplane Region) *Let $d \in \mathbb{R}^n$, $d \neq 0$. A parallel-hyperplane region $\mathcal{R}$ with direction d is a set that satisfies*

$$\mathcal{R} = \bigcup_{a\in\mathcal{A}} \{x \in \mathbb{R}^n : d^T x = a\}, \tag{4.3}$$

for some $\mathcal{A} \subseteq \mathbb{R}$.

Remark 4.3 *The vector d and the set $\mathcal{A}$ that define a parallel-hyperplane region $\mathcal{R}$ are not unique. It is straightforward to check that if $\mathcal{R}$ has direction d, then it also has direction αd for any nonzero $\alpha \in \mathbb{R}$.*

Definition 4.4 (Parallel-hyperplane Quantiser) *Fix $d \in \mathbb{R}^n$, $d \neq 0$. A parallel-hyperplane quantiser q with direction d is a quantiser whose quantisation regions, $\mathcal{R}_i$, for all $i \in \mathbb{Z}$, are parallel-hyperplane regions with (the same) direction d.*

4.3 Preliminary Results

We next present some preliminary results that will be needed in the sequel. Since the quantiser structure that we consider in this chapter is a special case of that considered in the geometric approach of Chapter 3, we will utilise some of the results of that chapter. Recall that the geometric approach of Chapter 3 utilises a quadratic CLF

$$V(x) = x^T P x, \quad \text{where } P = P^T > 0 \tag{4.4}$$

and the matrices

$$L \triangleq A^T P A - P, \quad M \triangleq A^T P B \quad \text{and} \quad H \triangleq B^T P B - M^T L^{-1} M. \tag{4.5}$$

Other tools from Chapter 3 that we will employ in this chapter are: the decomposition of L as

$$L = U^T \Lambda U, \text{ where } UU^T = \mathrm{I}_n \text{ and } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{4.6}$$

where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the eigenvalues of L; the invertible transformation

$$T_u(x) \triangleq U(x + L^{-1} M u); \tag{4.7}$$

the increment of V along the trajectories of system (4.1),

$$\Delta V(x, u) \triangleq V(Ax + Bu) - V(x) = x^T L x + 2 x^T M u + u^T B^T P B u; \tag{4.8}$$

the sets

$$X(u) \triangleq \{x \in \mathbb{R}^n : \Delta V(x, u) < 0\}, \text{ and} \tag{4.9}$$

$$X_0(u) \triangleq \{x \in \mathbb{R}^n : \Delta V(x, u) \leq 0\}, \tag{4.10}$$

and their images through T_u:

$$\tilde{X}(u) = \{\tilde{x} \in \mathbb{R}^n : \tilde{x}^T \Lambda \tilde{x} + u^T H u < 0\}, \tag{4.11}$$

$$\tilde{X}_0(u) = \{\tilde{x} \in \mathbb{R}^n : \tilde{x}^T \Lambda \tilde{x} + u^T H u \leq 0\}. \tag{4.12}$$

Remark 4.5 *From Remark 3.3, we recall that, since we deal with open-loop unstable systems and we assume that L is invertible, then L necessarily has at least one positive eigenvalue. Then, particularising Remark 3.4 to single-input systems, we conclude that L has one positive and $n-1$ negative eigenvalues.*

The following result gives a property of H for single-input systems.

Lemma 4.6 *For a single-input system, the quantity H defined in (4.5) is a real number and satisfies $H < 0$.*

Proof. That H is a real number follows straightforwardly since the matrices in (4.5) have real entries and $B \in \mathbb{R}^{n\times 1}$. Let $u = Kx$, where $K \in \mathbb{R}^{1\times n}$, be quadratically stabilising with respect to the CLF V, that is, $\Delta V(x, Kx) < 0$ for all nonzero $x \in \mathbb{R}^n$. Define

$$\mathcal{P}(\bar{u}) \triangleq \{x \in \mathbb{R}^n : Kx = \bar{u}\} \tag{4.13}$$

and note that $\Delta V(x, 0) < 0$ for all nonzero $x \in \mathcal{P}(0)$. Therefore, we have $\mathcal{P}(0) \setminus \{0\} \subset X(0)$ [see (4.9)]. Transforming $\mathcal{P}(\bar{u})$ through $T_{\bar{u}}$, [see (4.7)], we obtain

$$\tilde{\mathcal{P}}(\bar{u}) \triangleq T_{\bar{u}}(\mathcal{P}(\bar{u})) = \{\tilde{x} \in \mathbb{R}^n : KU^T\tilde{x} = (1 + KL^{-1}M)\bar{u}\}. \tag{4.14}$$

Since $T_{\bar{u}}$ is invertible and $T_0(0) = 0$, then $\mathcal{P}(0) \setminus \{0\} \subset X(0)$ if and only if $\tilde{\mathcal{P}}(0) \setminus \{0\} \subset \tilde{X}(0)$. Then, applying Theorem 3.10 part 1 to (4.14) with $\bar{u} = 0$ proves that $KU^T\Lambda^{-1}UK^T > 0$. For any $\bar{u} \neq 0$, we have $\mathcal{P}(\bar{u}) \subset X(\bar{u})$ and hence $\tilde{\mathcal{P}}(\bar{u}) \subset \tilde{X}(\bar{u})$. Then, applying Theorem 3.10 part 3 to (4.14) with $\bar{u} \neq 0$ shows that

$$(1 + KL^{-1}M)^2\bar{u}^2(KU^T\Lambda^{-1}UK^T)^{-1} < -H\bar{u}^2.$$

Since $(1 + KL^{-1}M)^2 \geq 0$, $\bar{u}^2 > 0$ and $KU^T\Lambda^{-1}UK^T > 0$, it then follows that $0 < -H$, whence $H < 0$. □

For future reference, whenever $d, \tilde{d} \in \mathbb{R}^n$, we define

$$\gamma \triangleq \sqrt{-Hd^TL^{-1}d}, \quad \tilde{\gamma} \triangleq \sqrt{-H\ \tilde{d}^T\ \Lambda^{-1}\tilde{d}} \quad \text{and} \quad \beta \triangleq -d^TL^{-1}M, \tag{4.15}$$

with L, M and H as defined in (4.5) and with Λ as in (4.6).

In Chapter 3, Theorem 3.10 provided the main tool for the geometric approach developed. The following theorem gives additional results that will also serve as a main tool for our development in this chapter.

Theorem 4.7 *Let $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$, where $\lambda_1, \ldots, \lambda_{n-1} < 0$ and $\lambda_n > 0$. Let $\tilde{d} \in \mathbb{R}^n \setminus \{0\}$, $u \in \mathbb{R}$ and $\tilde{a} \in \mathbb{R}$, and define $\tilde{\mathcal{P}} \triangleq \{\tilde{x} \in \mathbb{R}^n : \tilde{d}^T\ \tilde{x} = \tilde{a}\}$. Let $\tilde{X}(u)$ and $\tilde{X}_0(u)$ be the sets defined in (4.11) and (4.12), respectively, where $H < 0$, and consider the quantity $\tilde{\gamma}$ defined in (4.15). Then,*

1. If $\tilde{d}^T\ \Lambda^{-1}\tilde{d} = 0$ and $u \neq 0$, then $\tilde{\mathcal{P}} \subset \tilde{X}(u)$ if and only if $\tilde{a} = 0$.

2. There exists $\tilde{p} \in \tilde{\mathcal{P}}$ such that $\tilde{p} \notin \tilde{X}(u)$ and $\tilde{\mathcal{P}} \setminus \{\tilde{p}\} \subset \tilde{X}(u)$ if and only if

$$\tilde{d}^T\ \Lambda^{-1}\tilde{d} > 0, \quad \text{and} \tag{4.16}$$

$$\tilde{a}^2 = \tilde{\gamma}^2u^2. \tag{4.17}$$

Also, if such a $\tilde{p}$ exists, then

$$\tilde{p}^T \Lambda \tilde{p} + Hu^2 = 0. \tag{4.18}$$

Proof. See Appendix A. □

The following is the last preliminary result that we require.

Lemma 4.8 *Let $d \in \mathbb{R}^n$ satisfy $d \neq 0$ and $d^T L^{-1} d \geq 0$, and consider γ and β as defined in* (4.15)*. Then, $|\beta| > \gamma$.*

Proof. Note that $H < 0$ by Lemma 4.6. Then, since $d^T L^{-1} d \geq 0$, we have $\gamma \in \mathbb{R}$ and the inequality $|\beta| > \gamma$ is well defined.

If $d^T L^{-1} d > 0$, then Lemma 3.13 shows that

$$\beta^2 (d^T L^{-1} d)^{-1} > -H,$$

whence $\beta^2 > \gamma^2$. Then, the result follows by taking square root.

If $d^T L^{-1} d = 0$, note that $\gamma = 0$ and consider $\mathcal{P}(0)$ as defined in (4.13), where $K \in \mathbb{R}^{1 \times n}$ satisfies $\Delta V(x, Kx) < 0$ for all nonzero $x \in \mathbb{R}^n$. Fix $0 \neq u \in \mathbb{R}$ and consider the set $X(u)$ in (4.9). Since $\Delta V(0, u) = B^T P B u^2 \geq 0$, then $0 \notin X(u)$. Note that $0 \in \mathcal{P}(0)$ and hence $\mathcal{P}(0) \not\subset X(u)$. Transform $\mathcal{P}(0)$ through T_u to obtain

$$T_u(\mathcal{P}(0)) = \{\tilde{x} \in \mathbb{R}^n : d^T U^T \tilde{x} = d^T L^{-1} M u\}. \tag{4.19}$$

Since T_u is invertible, then $T_u(\mathcal{P}(0)) \not\subset \tilde{X}(u)$. Note that $0 = d^T L^{-1} d = d^T U^T \Lambda^{-1} U d$ from (4.6). Then, applying Theorem 4.7 part 1, it follows from (4.19) and since $u \neq 0$, that $d^T L^{-1} M u \neq 0$. Using (4.15), then $-\beta u \neq 0$, whence $|\beta| > 0$. Since $\gamma = 0$, the result follows. □

4.4 Quantiser Coarseness

In this section, we explore the concept of coarseness of a quantiser from a state-space standpoint. Recall the setting of Figure 4.1. Given a quadratic CLF V, we seek parallel-hyperplane quantisers q that are QS with respect to V. In addition, we would like to impose the condition that the quantisation regions of q be as large as possible.

In §4.4.1, we follow this intuitive idea by defining coarse-QS pairs. We then show that defining a coarse-QS quantiser in a corresponding sense is not useful. This motivates us to relax the QS constraint by defining and characterising coarse-almost-QS (CAQS) pairs. In §4.4.2, we define and characterise CAQS quantisers, showing that CAQS quantisers have logarithmically spaced quantisation values. We then show how to build logarithmic QS quantisers whose spacing between quantisation values is arbitrarily close to that of a given CAQS quantiser. In §4.5, we analyse connections with quantisation density.

4.4.1 Coarse-QS and CAQS Pairs

We begin by defining coarse-QS pairs.

Definition 4.9 (Coarse-QS Pair) *Let $\mathcal{R}$ be a parallel-hyperplane region with direction d and let $u \in \mathbb{R}$. We say that the pair $(u, \mathcal{R})$ is* coarse-QS *if (a) the pair $(u, \mathcal{R})$ is QS and (b) whenever $\mathcal{R} \subseteq \mathcal{R}^\star$, where $\mathcal{R}^\star$ is a parallel-hyperplane region with direction d and $(u, \mathcal{R}^\star)$ is QS, then $\mathcal{R}^\star = \mathcal{R}$.*

The rationale of this definition is that a pair $(u, \mathcal{R})$ is coarse-QS if the region $\mathcal{R}$ is as large as possible, subject to the constraints that $\mathcal{R}$ be a parallel-hyperplane region with direction d and that the pair $(u, \mathcal{R})$ be QS. By extension, if we could find a parallel-hyperplane quantiser with direction d such that all of its pairs $(u, \mathcal{R})$ were coarse-QS, then it would be natural to define such a quantiser as a coarse-QS quantiser. However, one of the consequences of the following lemma is that such a quantiser does not exist.

Lemma 4.10 (Characterisation of Coarse-QS Pairs) *Let $d \in \mathbb{R}^n$ satisfy $d^T L^{-1} d > 0$, $\mathcal{R}$ satisfy (4.3) for some nonempty $\mathcal{A} \subseteq \mathbb{R}$, and $u \in \mathbb{R}$. Then, $(u, \mathcal{R})$ is coarse-QS if and only if one of the following statements holds:*

1) $u = 0$ and $\mathcal{A} = \{0\}$,

2) $u \neq 0$ and $\mathcal{A} = (\beta u - \gamma|u|, \beta u + \gamma|u|)$,

with L as defined in (4.5), and γ and β in (4.15).

Proof. By definition, a coarse-QS pair is QS. If $u = 0$, then Theorem 3.12 part 1) establishes that $(u, \mathcal{R})$ is QS if and only if 1) holds, whence it straightforwardly follows that $(u, \mathcal{R})$ is coarse-QS if and only if 1) holds. If $u \neq 0$, Theorem 3.12 part 2) establishes that $(u, \mathcal{R})$ is QS if and only if $(a + d^T L^{-1} M u)^2 < -H(d^T L^{-1} d) u^2$, for all $a \in \mathcal{A}$. Using (4.15) it follows that $(u, \mathcal{R})$ is QS if and only if $(a - \beta u)^2 < \gamma^2 u^2$ for all $a \in \mathcal{A}$. Note that $(a - \beta u)^2 < \gamma^2 u^2$ if and only if $\beta u - \gamma|u| < a < \beta u + \gamma|u|$. Therefore, we have that $(u, \mathcal{R})$ is QS if and only if $\mathcal{A} \subseteq (\beta u - \gamma|u|, \beta u + \gamma|u|)$. Hence, it straightforwardly follows that $(u, \mathcal{R})$ is coarse-QS if and only if 2) holds. □

We now verify that there can exist no parallel-hyperplane quantiser having all its quantisation value/region pairs coarse-QS. For a contradiction, suppose that q is a parallel-hyperplane quantiser with direction d all of whose quantisation value/region pairs are coarse-QS. Then, Lemma 3.7 shows that q is QS. Then, Theorem 3.14 shows that $D = d$ must necessarily satisfy $d^T L^{-1} d > 0$. Let $\mathcal{R}_i$ and u_i, for all $i \in \mathbb{Z}$ denote the quantisation regions and corresponding values of such a quantiser. From Definition 2.5, we require that

$$\bigcup_{i \in \mathbb{Z}} \mathcal{R}_i = \mathbb{R}^n \text{ and } \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \text{ whenever } i \neq j. \tag{4.20}$$

Since the $\mathcal{R}_i$ are parallel-hyperplane regions with direction d, they all satisfy

$$\mathcal{R}_i = \bigcup_{a \in \mathcal{A}_i} \{x \in \mathbb{R}^n : d^T x = a\}.$$

By Lemma 4.10, the sets $\mathcal{A}_i$ satisfy $\mathcal{A}_i = (\beta u_i - \gamma|u_i|, \beta u_i + \gamma|u_i|)$ whenever $u_i \neq 0$, and $\mathcal{A}_i = \{0\}$ for the unique $i \in \mathbb{Z}$ such that $u_i = 0$. To satisfy (4.20), the sets $\mathcal{A}_i$ must satisfy $\bigcup_{i\in\mathbb{Z}} \mathcal{A}_i = \mathbb{R}$ and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$, which is not possible, since the $\mathcal{A}_i$ are all bounded open intervals, except for the unique one satisfying $\mathcal{A}_i = \{0\}$.

This fact motivates the following definition.

Definition 4.11 (CAQS Pair) *Let $d \in \mathbb{R}^n$, let $\mathcal{R}$ be a parallel-hyperplane region with direction d and let $u \in \mathbb{R}$. We say that the pair $(u, \mathcal{R})$ is* coarse-almost-QS *(CAQS) if (a) there exists $p \in \mathcal{R}$, such that $(u, \mathcal{R} \setminus \{p\})$ is QS and (b) whenever $\mathcal{R} \subseteq \mathcal{R}^\star$, where $\mathcal{R}^\star$ is a parallel-hyperplane region with direction d such that there exists $p^\star \in \mathcal{R}^\star$ satisfying $(u, \mathcal{R}^\star \setminus \{p^\star\})$ is QS, then $\mathcal{R}^\star = \mathcal{R}$.*

If $(u, \mathcal{R})$ is CAQS, $\mathcal{R}$ may contain at most one point that makes the pair $(u, \mathcal{R})$ not QS. In addition, there exists no other region that contains $\mathcal{R}$ and has this property. We will see that Definition 4.11 does allow us to define a CAQS quantiser as a quantiser all of whose pairs $(u_i, \mathcal{R}_i)$ are CAQS. The following proposition is needed in the subsequent characterisation of CAQS pairs.

Proposition 4.12 *Let $u \in \mathbb{R}$, $\mathcal{R}$ be a parallel-hyperplane region, $0 \in \mathcal{R}$ and $(u, \mathcal{R})$ be CAQS. Then, $(u, \mathcal{R})$ is QS.*

Proof. For a contradiction, suppose that $(u, \mathcal{R})$ is CAQS but not QS. Then, there exists $p \in \mathcal{R}$, $p \neq 0$, such that $p \notin X(u)$ and $(u, \mathcal{R} \setminus \{p\})$ is QS. From Definition 3.6 and since $0 \in \mathcal{R} \setminus \{p\}$, we then have $u = 0$ and $\mathcal{R} \setminus \{p, 0\} \subset X(0)$. Since $\mathcal{R}$ is a parallel-hyperplane region, it satisfies (4.3) for some nonzero $d \in \mathbb{R}^n$ and some $\mathcal{A} \subseteq \mathbb{R}$. Note that $0 \in \mathcal{A}$ since $0 \in \mathcal{R}$. Define

$$\mathcal{P}(a) \triangleq \{x \in \mathbb{R}^n : d^T x = a\}$$

and note that $\mathcal{P}(0) \in \mathcal{R}$ and $\mathcal{P}(0) \setminus \{p, 0\} \subset X(0)$.

If $p \in \mathcal{P}(0)$, then $\delta p \in \mathcal{P}(0)$ for all $\delta \in \mathbb{R}$. Eqs. (4.9) and (4.8) show that $X(0) = \{x \in \mathbb{R}^n : x^T L x < 0\}$. Since $p \notin X(0)$, then $\delta p \notin X(0)$ for all $\delta \in \mathbb{R}$. Hence, $\mathcal{P}(0) \setminus \{p, 0\} \not\subset X(0)$, reaching a contradiction.

If $p \notin \mathcal{P}(0)$, then let $b = d^T p$, note that $b \neq 0$ and consider $\mathcal{P}(b)$. Note that $\mathcal{P}(b) \subset \mathcal{R}$ and, since $\mathcal{R} \setminus \{p, 0\} \subset X(0)$ and $0 \notin \mathcal{P}(b)$, then $\mathcal{P}(b) \setminus \{p\} \subset X(0)$. Since T_u is invertible, we have $\mathcal{P}(b) \setminus \{p\} \subset X(0)$ if and only if $T_0(\mathcal{P}(b)) \setminus \{T_0(p)\} \subset \tilde{X}(0)$. Using (4.7) yields

$$T_0(\mathcal{P}(b)) = \{\tilde{x} \in \mathbb{R}^n : d^T U^T \tilde{x} = b\}.$$

Then, (4.17) in Theorem 4.7 shows that $b = 0$. Contradiction.

Therefore, we have established that $(u, \mathcal{R})$ is QS. □

Lemma 4.13 (Characterisation of CAQS Pairs) *Let $d \in \mathbb{R}^n$ satisfy $d^T L^{-1} d > 0$, with L as in* (4.5)*, let $\mathcal{R}$ satisfy (4.3) for some nonempty $\mathcal{A} \subseteq \mathbb{R}$, and let $u \in \mathbb{R}$. Then, $(u, \mathcal{R})$ is CAQS if and only if one of the following statements holds:*

1) $u = 0$ and $\mathcal{A} = \{0\}$,

2) $u \neq 0$ and either $\mathcal{A} = [\beta u - \gamma|u|, \beta u + \gamma|u|)$ or $\mathcal{A} = (\beta u - \gamma|u|, \beta u + \gamma|u|]$,

where γ and β were defined in (4.15).

Proof. Necessity.

- If $0 \in \mathcal{R}$, then Proposition 4.12 shows that $(u, \mathcal{R})$ is QS and Theorem 3.12 establishes 1).
- If $0 \notin \mathcal{R}$, then consider the sets $\mathcal{P}(a) = \{x \in \mathbb{R}^n : d^T x = a\}$, where $a \in \mathbb{R}$. Since $\mathcal{A}$ is nonempty, there exists $0 \neq a_2 \in \mathcal{A}$ and hence $\mathcal{P}(a_2) \subseteq \mathcal{R}$, for any such a_2. Since $(u, \mathcal{R})$ is CAQS, then either

  (i) $(u, \mathcal{P}(a_2))$ is QS or

  (ii) there exists $p \in \mathcal{P}(a_2)$, $p \notin X(u)$, such that $(u, \mathcal{P}(a_2) \setminus \{p\})$ is QS.

  If (i) is true, then Theorem 3.12 shows that $u \neq 0$. If (ii) is true, then $\mathcal{P}(a_2) \setminus \{p\} \subset X(u)$ and hence $T_u(\mathcal{P}(a_2)) \setminus T_u(\{p\}) \subset \tilde{X}(u)$. Note that $T_u(p) \notin \tilde{X}(u)$, since $p \notin X(u)$. Using (4.7) and (4.15), we have

  $$T_u(\mathcal{P}(a)) = \{\tilde{x} \in \mathbb{R}^n : d^T U^T \tilde{x} = a - \beta u\}. \tag{4.21}$$

  Then, from Theorem 4.7 and (4.6), we prove that $(a_2 - \beta u)^2 = \tilde{\gamma}^2 u^2$. The latter implies that $u \neq 0$, since otherwise $a_2 = 0$. We have thus proved that if $0 \notin \mathcal{R}$, then $u \neq 0$.

  Recall that $\mathcal{R} = \bigcup_{a \in \mathcal{A}} \mathcal{P}(a)$ and that we have either (i) or (ii) above.

  a) From Theorem 3.12, (i) is true if and only if $\beta u - \gamma|u| < a_2 < \beta u + \gamma|u|$.

  b) Using (4.21) and Theorem 4.7, (ii) is true if and only if either $a_2 = \beta u - \gamma|u|$ or $a_2 = \beta u + \gamma|u|$.

  Therefore, 2) follows straightforwardly from a) and b) above and Definition 4.11.

This completes the necessity part of the proof.

Sufficiency. Consider the sets $\mathcal{P}(a)$ defined in the necessity part of the proof and let

$$\mathcal{R}^\star = \bigcup_{a \in \mathcal{A}^\star} \mathcal{P}(a),$$

where $\mathcal{A} \subseteq \mathcal{A}^\star$ and hence $\mathcal{R} \subseteq \mathcal{R}^\star$. Assume that there exists $p^\star \in \mathcal{R}^\star$ such that $\mathcal{R}^\star \setminus \{p^\star\} \subset X(u)$. Let $a^\star$ be such that $\mathcal{P}(a^\star) \subseteq \mathcal{R}^\star$. Then, either $(u, \mathcal{P}(a^\star))$ is QS or $(u, \mathcal{P}(a^\star) \setminus \{p^\star\})$ is QS and $p^\star \notin X(u)$. We next show that (a) and (b) in Definition 4.11 hold.

If 1) is true, then Theorem 3.12 shows that $(u, \mathcal{R})$ is QS. Then, given any $p \in \mathcal{R}$, $(u, \mathcal{R} \setminus \{p\})$ also is QS, establishing (a). Using (4.21) and Theorem 4.7, we obtain $a^\star = 0$, thus proving that $\mathcal{R}^\star = \mathcal{R}$ and establishing (b).

If 2) is true, let $a_2 \in \mathbb{R}$ satisfy $\mathcal{P}(a_2) \subseteq \mathcal{R}$. If $\beta u - \gamma|u| < a_2 < \beta u + \gamma|u|$, then Theorem 3.12 shows that $(u, \mathcal{P}(a_2))$ is QS. If $a_2 = \beta u - \gamma|u|$ or $a_2 = \beta u + \gamma|u|$, then using (4.21) and Theorem 4.7, it follows that there exists $p \in \mathcal{P}(a_2)$, $p \notin X(u)$ such that $(u, \mathcal{P}(a_2) \setminus \{p\})$ is QS. Since $\mathcal{R} = \bigcup_{a \in \mathcal{A}} \mathcal{P}(a)$, with $\mathcal{A}$ as in 2), then we see that there exists $p \in \mathcal{R}$ such that $(u, \mathcal{R} \setminus \{p\})$ is QS, establishing (a). Applying to $\mathcal{R}^\star$ the same reasoning that was applied to $\mathcal{R}$ in the necessity proof, we can prove that $\mathcal{A}^\star = \mathcal{A}$ and hence $\mathcal{R}^\star = \mathcal{R}$, which establishes (b).

This proves that $(u, \mathcal{R})$ is CAQS and concludes the sufficiency proof. □

4.4.2 CAQS Quantisers

In this section, we define and characterise CAQS quantisers. We show that CAQS quantisers are logarithmic and show how to construct logarithmic QS parallel-hyperplane quantisers having a logarithmic base that is arbitrarily close to that of a given CAQS quantiser. We shall henceforth consider only symmetric quantisers, that is, quantisers that satisfy $q(x) = -q(-x)$, for all x.

Definition 4.14 (CAQS Quantiser) *Let q be a parallel-hyperplane quantiser with direction $d \in \mathbb{R}^n$. Let $\mathcal{R}_i$ and u_i, for all $i \in \mathbb{Z}$, be its quantisation regions and corresponding quantisation values, respectively. We say that q is CAQS if $(u_i, \mathcal{R}_i)$ is CAQS, for all $i \in \mathbb{Z}$.*

The following proposition is needed to achieve the characterisation of CAQS quantisers in Theorem 4.16.

Proposition 4.15 *Let q be a parallel-hyperplane quantiser with direction $d \in \mathbb{R}^n$ and let q be CAQS. Then, $d^T L^{-1} d > 0$.*

Proof. Since q is parallel-hyperplane with direction d, then $d \neq 0$. Let $\mathcal{R}$ be the quantisation region of q that contains the origin, and let u be the corresponding value. By Definition 4.14, $(u, \mathcal{R})$ is CAQS. By Proposition 4.12, then $(u, \mathcal{R})$ is QS. Consider $\mathcal{P} \triangleq \{x \in \mathbb{R}^n : d^T x = 0\}$. Note that $\mathcal{P} \subseteq \mathcal{R}$ and hence $(u, \mathcal{P})$ also is QS. Then, Lemma 3.11 establishes that $d^T L^{-1} d > 0$. □

Theorem 4.16 (Characterisation of CAQS Quantisers) *Let $d \in \mathbb{R}^n$, $d \neq 0$, consider β and γ as defined in (4.15), and define*

$$\rho \triangleq \frac{|\beta| - \gamma}{|\beta| + \gamma}. \tag{4.22}$$

Let q be a parallel-hyperplane quantiser with direction d, satisfying $q(x) = -q(-x)$ for all $x \in \mathbb{R}^n$. Then, q is CAQS if and only if $d^T L^{-1} d > 0$, $0 < \rho < 1$, $|\beta| \pm \gamma > 0$ and there exists $u_0 \in \mathbb{R}$, $\beta u_0 > 0$,

such that q satisfies

$$q(x) = \begin{cases} 0 & \text{if } x \in \bar{\mathcal{R}}, \quad (4.23a) \\ u_i & \text{if } x \in \mathcal{R}_i, \quad (4.23b) \\ -u_i & \text{if } x \in -\mathcal{R}_i, \quad (4.23c) \end{cases}$$

$$\bar{\mathcal{R}} = \{x \in \mathbb{R}^n : d^T x = 0\}, \quad (4.24)$$

$$u_i = \rho^{-i} u_0, \quad (4.25)$$

$$\mathcal{R}_i = \rho^{-i} \mathcal{R}_0, \quad (4.26)$$

$$\sigma_0 = \beta u_0 - \gamma |u_0|. \quad (4.27)$$

$$\mathcal{R}_0 = \{x \in \mathbb{R}^n : \sigma_0 < d^T x \leq \rho^{-1}\sigma_0\}, \text{ or} \quad (4.28a)$$

$$\mathcal{R}_0 = \{x \in \mathbb{R}^n : \sigma_0 \leq d^T x < \rho^{-1}\sigma_0\}, \quad (4.28b)$$

for all $i \in \mathbb{Z}$.

Proof. Sufficiency. Since $\beta u_0 > 0$ and $|\beta| \pm \gamma > 0$, using (4.27) we have $\sigma_0 > 0$. Since $d^T L^{-1} d > 0$, then $d \neq 0$. It is then straightforward to check that q, as defined above, is indeed a parallel-hyperplane quantiser with direction d. We next prove that q is CAQS. Since $d^T L^{-1} d > 0$, from (4.24), Lemma 4.13 shows that $(0, \bar{\mathcal{R}})$ is CAQS. Suppose that $\mathcal{R}_0$ has the form (4.28a). Using (4.27) and (4.22), and since $\beta u_0 > 0$, we can straightforwardly show that

$$\mathcal{R}_0 = \{x \in \mathbb{R}^n : \beta u_0 - \gamma |u_0| < d^T x \leq \beta u_0 + \gamma |u_0|\}. \quad (4.29)$$

By Lemma 4.13, we see that $(u_0, \mathcal{R}_0)$ is CAQS. From (4.22), (4.25), (4.26), (4.29) and the facts that $0 < \rho$ and $\beta u_0 > 0$, it follows that

$$\mathcal{R}_i = \{x \in \mathbb{R}^n : \beta u_i - \gamma |u_i| < d^T x \leq \beta u_i + \gamma |u_i|\}.$$

Then, Lemma 4.13 shows that $(u_i, \mathcal{R}_i)$ is CAQS for all $i \in \mathbb{Z}$. The same follows by considering $\mathcal{R}_0$ in (4.28b) and a similar procedure can be used to prove that $(-u_i, -\mathcal{R}_i)$ is CAQS for $i \in \mathbb{Z}$. Recalling Definition 4.14, the sufficiency proof is concluded.

Necessity. Since q is a CAQS parallel-hyperplane quantiser with direction d, then Proposition 4.15 establishes that $d^T L^{-1} d > 0$. Let $\bar{\mathcal{R}}$ denote the quantisation region of q that contains the origin and let $\bar{u}$ denote its corresponding value. Since $q(x) = -q(-x)$, then $q(0) = 0$ and hence $\bar{u} = 0$. Since q is CAQS, then $(\bar{u}, \bar{\mathcal{R}})$ is CAQS and Lemma 4.13 shows that $\bar{\mathcal{R}} = \{x \in \mathbb{R}^n : d^T x = 0\}$, establishing (4.23a) and (4.24). Using Lemma 4.6 and (4.15), we have $\gamma > 0$ and from Lemma 4.8, we prove that $|\beta| \pm \gamma > 0$ and $0 < \rho < 1$. Since q is CAQS and $q(x) = -q(-x)$, using Lemma 4.13 it is

straightforward to check that q must have at least one quantisation value u_0 satisfying $\beta u_0 > 0$ (note that $\beta \neq 0$ since $|\beta| > \gamma > 0$). Let $\mathcal{R}_0$ denote the quantisation region of q whose corresponding value is u_0. Since $(u_0, \mathcal{R}_0)$ is CAQS, using Lemma 4.13 we obtain

$$\mathcal{R}_0 = \{x \in \mathbb{R}^n : \beta u_0 - \gamma|u_0| < d^T x \leq \beta u_0 + \gamma|u_0|\}, \text{ or} \tag{4.30a}$$

$$\mathcal{R}_0 = \{x \in \mathbb{R}^n : \beta u_0 - \gamma|u_0| \leq d^T x < \beta u_0 + \gamma|u_0|\}. \tag{4.30b}$$

Since $\beta u_0 > 0$, then $\beta u_0 - \gamma|u_0| = (|\beta| - \gamma)|u_0|$ and $\beta u_0 + \gamma|u_0| = (|\beta| + \gamma)|u_0|$. Using (4.22), then $\beta u_0 + \gamma|u_0| = \rho^{-1}(\beta u_0 - \gamma|u_0|)$. Defining then σ_0 to satisfy (4.27), it follows that (4.30) establishes (4.28). We now proceed to prove (4.23b), (4.25) and (4.26) for all $i \in \mathbb{Z}$ by induction. Let $j \in \mathbb{Z}_{+,0}$, and assume that q contains a region of the form (4.26) for $i = j$. Note that this assumption is satisfied when $j = 0$.

Using (4.26) and (4.30), we have

$$\mathcal{R}_j = \{x \in \mathbb{R}^n : \rho^{-j}(\beta u_0 - \gamma|u_0|) < d^T x \leq \rho^{-j}(\beta u_0 + \gamma|u_0|)\} \text{ or} \tag{4.31a}$$

$$\mathcal{R}_j = \{x \in \mathbb{R}^n : \rho^{-j}(\beta u_0 - \gamma|u_0|) \leq d^T x < \rho^{-j}(\beta u_0 + \gamma|u_0|)\}, \tag{4.31b}$$

where (4.31a) corresponds to (4.30a) and (4.31b) to (4.30b). Let u_j denote the value of q corresponding to $\mathcal{R}_j$. Since $(u_j, \mathcal{R}_j)$ is CAQS, using Lemma 4.13 and (4.31), we obtain

$$\rho^{-j}(\beta u_0 - \gamma|u_0|) = \beta u_j - \gamma|u_j| \quad \text{and} \quad \rho^{-j}(\beta u_0 + \gamma|u_0|) = \beta u_j + \gamma|u_j|. \tag{4.32}$$

From (4.32) and the facts that $\rho > 0$, $\beta u_0 > 0$ and $|\beta| \pm \gamma > 0$, it follows that $\beta u_j > 0$ and $u_j = \rho^{-j} u_0$, which establishes (4.25) for $i = j$. We next show that (4.23b), (4.25) and (4.26) hold for $i = j + 1$.

If $\mathcal{R}_j$ has the form (4.31a), let $x^\star \in \mathbb{R}^n$ satisfy

$$d^T x^\star = \rho^{-j}(\beta u_0 + \gamma|u_0|) + \epsilon = \beta u_j + \gamma|u_j| + \epsilon, \tag{4.33}$$

for some $\epsilon > 0$ (arbitrarily small) such that $x^\star \notin \mathcal{R}_j$. Let $\mathcal{R}_1^\star$ be the quantisation region of q that contains $x^\star$ and let $u_1^\star$ be the value of q corresponding to $\mathcal{R}_1^\star$. Since $(u_1^\star, \mathcal{R}_1^\star)$ is CAQS, we have, using Lemma 4.13,

$$\mathcal{R}_1^\star = \{x \in \mathbb{R}^n : \beta u_1^\star - \gamma|u_1^\star| < d^T x \leq \beta u_1^\star + \gamma|u_1^\star|\}, \text{ or} \tag{4.34a}$$

$$\mathcal{R}_1^\star = \{x \in \mathbb{R}^n : \beta u_1^\star - \gamma|u_1^\star| \leq d^T x < \beta u_1^\star + \gamma|u_1^\star|\}. \tag{4.34b}$$

Since $x^\star \notin \mathcal{R}_j$, $x^\star \in \mathcal{R}_1^\star$ and $\mathcal{R}_1^\star \cap \mathcal{R}_j = \emptyset$, it follows from (4.31), (4.33) and (4.34) that $\rho^{-j}(\beta u_0 + \gamma|u_0|) \leq \beta u_1^\star - \gamma|u_1^\star|$. Note that $0 < \rho^{-j}(\beta u_0 + \gamma|u_0|)$ because $\rho > 0$, $\beta u_0 > 0$ and $|\beta| > \gamma > 0$. Then, $\beta u_1^\star - \gamma|u_1^\star| > 0$ and $\beta u_1^\star > 0$.

Recall that we have selected $\epsilon > 0$ arbitrarily small. For a contradiction, suppose that $\beta u_1^\star - \gamma|u_1^\star| \neq \rho^{-j}(\beta u_0 + \gamma|u_0|)$. Then, q must have a quantisation region, $\mathcal{R}_2^\star$, satisfying $\mathcal{R}_2^\star \subseteq \mathcal{R}_3^\star$, where

$$\mathcal{R}_3^\star = \{x \in \mathbb{R}^n : \rho^{-j}(\beta u_0 + \gamma|u_0|) \leq d^T x \leq (\beta u_1^\star - \gamma|u_1^\star|)\}.$$

Since $x^\star \in \mathcal{R}_1^\star$ and satisfies (4.33), using (4.34) we have

$$\beta u_1^\star - \gamma|u_1^\star| - \rho^{-j}(\beta u_0 + \gamma|u_0|) \leq \epsilon. \tag{4.35}$$

Let $u_2^\star$ denote the value of q corresponding to $\mathcal{R}_2^\star$. Using the same argument as for $u_1^\star$, we can prove that $\beta u_2^\star > 0$. Since $(u_2^\star, \mathcal{R}_2^\star)$ is CAQS, using Lemma 4.13, we obtain

$$\mathcal{R}_2^\star = \{x \in \mathbb{R}^n : \beta u_2^\star - \gamma|u_2^\star| < d^T x \leq \beta u_2^\star + \gamma|u_2^\star|\}, \text{ or} \tag{4.36a}$$

$$\mathcal{R}_2^\star = \{x \in \mathbb{R}^n : \beta u_2^\star - \gamma|u_2^\star| \leq d^T x < \beta u_2^\star + \gamma|u_2^\star|\}, \tag{4.36b}$$

where, since $\mathcal{R}_2^\star \subseteq \mathcal{R}_3^\star$,

$$\beta u_2^\star - \gamma|u_2^\star| \geq \rho^{-j}(\beta u_0 + \gamma|u_0|), \tag{4.37}$$

$$\beta u_2^\star + \gamma|u_2^\star| \leq \beta u_1^\star - \gamma|u_1^\star|. \tag{4.38}$$

From (4.35), (4.37) and (4.38), we have

$$\beta u_2^\star + \gamma|u_2^\star| - (\beta u_2^\star - \gamma|u_2^\star|) = 2\gamma|u_2^\star| \leq \epsilon. \tag{4.39}$$

Since $\gamma > 0$, then (4.39) can be written as $|u_2^\star| \leq \epsilon/2\gamma$. Since $u_2^\star \neq 0$ because $\beta u_2^\star > 0$, then we reach a contradiction by selecting $\epsilon > 0$ small enough. Then, $\beta u_1^\star - \gamma|u_1^\star| = \beta u_j + \gamma|u_j|$ and since $\mathcal{R}_1^\star$ and $\mathcal{R}_j$ must be disjoint, we have, using (4.22),

$$\mathcal{R}_1^\star = \{x \in \mathbb{R}^n : \beta u_j + \gamma|u_j| < d^T x \leq \rho^{-1}(\beta u_j + \gamma|u_j|)\}. \tag{4.40}$$

If $\mathcal{R}_j$ has the form (4.31b), then let $x^\star \in \mathbb{R}^n$ satisfy

$$0 < d^T x^\star = \rho^{-j}(\beta u_0 + \gamma|u_0|) = \beta u_j + \gamma|u_j|, \tag{4.41}$$

and note that $x^\star \notin \mathcal{R}_j$. Let $\mathcal{R}_1^\star$ be the quantisation region of q that contains $x^\star$ and let $u_1^\star$ be the corresponding value. Since $(u_1^\star, \mathcal{R}_1^\star)$ is CAQS, then Lemma 4.13 shows that $\mathcal{R}_1^\star$ satisfies (4.34b). Then, $\beta u_1^\star - \gamma|u_1^\star| = \beta u_j + \gamma|u_j|$ and

$$\mathcal{R}_1^\star = \{x \in \mathbb{R}^n : \beta u_j + \gamma|u_j| \leq d^T x < \rho^{-1}(\beta u_j + \gamma|u_j|)\}. \tag{4.42}$$

Substituting $u_j = \rho^{-j} u_0$ into (4.40) or (4.42), and defining $\mathcal{R}_{j+1} \triangleq \mathcal{R}_1^\star$ and $u_{j+1} \triangleq u_1^\star$, we obtain $\mathcal{R}_{j+1} = \rho^{-(j+1)}\mathcal{R}_0$ and $u_{j+1} = \rho^{-(j+1)}u_0$, which establishes (4.23b), (4.25) and (4.26) for $i = j+1$. Hence, we have established that if q contains a region of the form (4.26), for $i = j$, then it must also contain a region of the form (4.26), for $i = j+1$. Also, we have established that the values of q corresponding to these regions satisfy (4.25). Hence, we have proved by induction that (4.23b), (4.25) and (4.26) hold for all $i \in \mathbb{Z}_{+,0}$. In a similar manner, we can prove that if q contains a region of the form (4.26), for $i = j$, then it must also contain a region of the form (4.26), for $i = j-1$. This establishes

(4.23b), (4.25) and (4.26) for all $i \in \mathbb{Z}$. Then, (4.23c) follows since $q(x) = -q(-x)$ for all $x \in \mathbb{R}^n$ by assumption. This concludes the proof. □

Theorem 4.16 shows how we may build a CAQS quantiser for a given direction d. This theorem also shows that a CAQS quantiser is logarithmic, with a logarithmic base given by ρ in (4.22). We next show how to construct logarithmic QS quantisers having a logarithmic base that is arbitrarily close to that of any given CAQS quantiser. Before proving Theorem 4.18, we need the following result, whose proof is straightforward and is therefore omitted.

Lemma 4.17 *Let $\mathcal{R} \subset \mathbb{R}^n$, and $u, \rho \in \mathbb{R}$. If $(u, \mathcal{R})$ is QS, then $(\rho u, \rho\mathcal{R})$ is QS.*

Theorem 4.18 *Let $d \in \mathbb{R}^n$ satisfy $d^T L^{-1} d > 0$ and consider β and γ as defined in (4.15). Then $|\beta| \pm \gamma > 0$. Let $\epsilon > 0$ satisfy*

$$0 < \frac{|\beta| - \gamma}{|\beta| + \gamma} + \epsilon < 1, \tag{4.43}$$

and define $\rho \triangleq \frac{|\beta|-\gamma}{|\beta|+\gamma} + \epsilon$. Let $\sigma_0, u_0 \in \mathbb{R}$ satisfy $\beta u_0 > 0$ and

$$\beta u_0 - \gamma|u_0| < \sigma_0 < \rho(\beta u_0 + \gamma|u_0|). \tag{4.44}$$

Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quantiser defined as in (4.23)–(4.26) and (4.28)*. Then, q is QS.*

Proof. Since $d^T L^{-1} d > 0$ and using Lemma 4.6, then $\gamma > 0$. From Lemma 4.8 then $|\beta| \pm \gamma > 0$. We next establish that q is QS. Consider $\bar{\mathcal{R}}$ in (4.24). By Theorem 3.12 and the fact that $d^T L^{-1} d > 0$, we have that $(0, \bar{\mathcal{R}})$ is QS. Since $\beta u_0 > 0$ and $|\beta| \pm \gamma > 0$, then $\beta u_0 - \gamma|u_0| > 0$ and from (4.44) then $\sigma_0 > 0$. Also, since $0 < \rho < 1$, multiplying the second inequality in (4.44) by $\rho^{-1} > 1$, and combining with the first inequality in (4.44), yields $\beta u_0 - \gamma|u_0| < \sigma_0 < \rho^{-1}\sigma_0 < \beta u_0 + \gamma|u_0|$. Then, from (4.28) and the fact that $d^T L^{-1} d > 0$, Theorem 3.12 shows that $(u_0, \mathcal{R}_0)$ is QS. Then, using Lemma 4.17 it follows from (4.25) and (4.26) that $(u_i, \mathcal{R}_i)$ and $(-u_i, -\mathcal{R}_i)$ are QS, for all $i \in \mathbb{Z}$. By Lemma 3.7, the result follows. □

4.5 CAQS Quantisers and Quantisation Density

We next discuss the relationship between CAQS quantisers and the standard concept of quantisation density, which we have considered in previous chapters.

Recall from Chapter 2 that the density of a quantiser $q : \mathbb{R}^r \to \mathbb{R}^s$, denoted $\eta(q)$, is defined as

$$\eta(q) \triangleq \limsup_{\epsilon \to 0} \frac{\#[\mathcal{U}(q) \cap C^s(\epsilon)]}{-2 \ln \epsilon}, \tag{4.45}$$

where $\#[\cdot]$ denotes the number of elements of a set, $\mathcal{U}(q)$ denotes the range of q, and $C^s(\epsilon) \triangleq \{u \in \mathbb{R}^s : \epsilon \leq \|u\|_2 \leq 1/\epsilon\}$.

The first link between CAQS quantisers and quantisation density is given in the following theorem, which shows that the density of a CAQS quantiser with direction d is the infimum over all QS parallel-hyperplane quantisers with direction d [with respect to a given and fixed CLF $V(x) = x^T P x$].

Theorem 4.19 *Let $q_d : \mathbb{R}^n \to \mathbb{R}$ be a CAQS quantiser with direction $d \in \mathbb{R}^n$. Then, $\eta(q_d) = \eta^\star$, where*

$$\eta^\star = \inf \eta(q), \quad \textit{subject to} \tag{4.46}$$

$$q \textit{ is QS parallel-hyperplane with direction } d, \tag{4.47}$$

and $\eta(\cdot)$ denotes the density of a quantiser, as defined in (4.45).

Proof. Recall that the density of a quantiser depends only on its values and not on its regions. Then, considering (4.22), (4.23) and (4.25) in Theorem 4.16, we have to prove that $\eta^\star$, as defined in (4.46)–(4.47), coincides with the density of a quantiser having logarithmically spaced values with logarithmic base $\rho = \frac{|\beta|-\gamma}{|\beta|+\gamma}$, with β and γ as defined in (4.15).

Let q be a QS parallel-hyperplane quantiser with direction d. Then Theorem 3.14 shows that $d^T L^{-1} d > 0$. From Lemma 4.6, Lemma 4.8 and (4.15), then $|\beta| > \gamma > 0$. Let $u_0, u_1 \in \mathbb{R}$ be adjacent quantisation values of q, that is, $u_0 \neq u_1$ and q has no other quantisation values in the interval $(\min\{u_0, u_1\}, \max\{u_0, u_1\})$. Suppose, without loss of generality, that $\beta u_0 > 0$. Note that this implies that $u_0 \neq 0$. Let $\mathcal{R}_0$ and $\mathcal{R}_1$ denote the quantisation regions of q with corresponding values u_0 and u_1, respectively. Since q is QS, then Lemma 3.7 shows that $(u_0, \mathcal{R}_0)$ and $(u_1, \mathcal{R}_1)$ are QS. We next proceed to find the maximum separation between the adjacent values u_0 and u_1 of q, subject to the constraints that $(u_0, \mathcal{R}_0)$ and $(u_1, \mathcal{R}_1)$ are QS. Since q is parallel-hyperplane with direction d, we have

$$\mathcal{R}_i = \bigcup_{a \in \mathcal{A}_i} \{x \in \mathbb{R}^n : d^T x = a\}, \tag{4.48}$$

for $i = 0, 1$. Since $(u_0, \mathcal{R}_0)$ is QS, $d^T L^{-1} d > 0$ and $u_0 \neq 0$, Theorem 3.12 imposes that

$$\mathcal{A}_i \subseteq (\beta u_i - \gamma |u_i|, \beta u_i + \gamma |u_i|), \tag{4.49}$$

for $i = 0$.

Claim A): $\rho < u_1/u_0 < \rho^{-1}$, with ρ as defined in (4.22).

Proof of Claim A). Note that since $|\beta| > \gamma > 0$, then $\rho > 0$. Since u_0 and u_1 are adjacent, $u_0 \neq 0$ and q is QS, then $u_1/u_0 \geq 0$, since otherwise u_0 and u_1 would not be adjacent because 0 necessarily is a value of q. Suppose for a contradiction that $\rho \geq u_1/u_0$. Since $u_1/u_0 \geq 0$, $\beta u_0 > 0$ and $\rho \geq u_1/u_0$, then

$$\beta u_1 \leq \rho \beta u_0 \quad \text{and} \quad |u_1| \leq \rho |u_0|. \tag{4.50}$$

Using (4.50), (4.22), and since $\gamma > 0$ and $\beta u_0 > 0$, we obtain

$$\beta u_1 + \gamma |u_1| \leq \rho(\beta u_0 + \gamma |u_0|) = \beta u_0 - \gamma |u_0|. \tag{4.51}$$

Let $x \in \mathbb{R}^n$ satisfy $d^T x = \beta u_0 - \gamma|u_0|$. From (4.48) and (4.49), it follows that $x \notin \mathcal{R}_0$. If $u_1 \neq 0$, then Theorem 3.12 part 2) imposes (4.49) for $i = 1$ because $(u_1, \mathcal{R}_1)$ is QS. Then, from (4.49) and (4.51), it follows that $x \notin \mathcal{R}_1$. If $u_1 = 0$, then note that $x \notin \mathcal{R}_1$ from Theorem 3.12 part 1) and since $\beta u_0 - \gamma|u_0| > 0$. Therefore, $x \notin \mathcal{R}_0$ and $x \notin \mathcal{R}_1$. Thus, let $\mathcal{R}_2$ be the quantisation region of q that contains x and let u_2 be its corresponding value. Since q is parallel-hyperplane, then $\mathcal{R}_2$ has the form (4.48) with $i = 2$. Since q is QS, then Lemma 3.7 shows that $(u_2, \mathcal{R}_2)$ is QS. Since $\beta u_0 - \gamma|u_0| > 0$ and $\beta u_0 - \gamma|u_0| \in \mathcal{A}_2$, then from Theorem 3.12 part 1), it follows that $u_2 \neq 0$. Then, Theorem 3.12 part 2) establishes (4.49) with $i = 2$. Since $x \in \mathcal{R}_2$, it follows that

$$\beta u_2 - \gamma|u_2| < \beta u_0 - \gamma|u_0| < \beta u_2 + \gamma|u_2|. \tag{4.52}$$

Since $|\beta| > \gamma$ and $\beta u_0 - \gamma|u_0| > 0$, then the second inequality in (4.52) implies that $\beta u_2 > 0$. Operating on the first inequality in (4.52) yields

$$(|\beta| - \gamma)|u_2| < (|\beta| - \gamma)|u_0|.$$

Therefore, since $|\beta| > \gamma$, then $|u_2| < |u_0|$, whence $|\beta||u_2| < |\beta||u_0|$ and then $\beta u_2 < \beta u_0$. From (4.51) and (4.52), we have

$$\beta u_1 + \gamma|u_1| \leq \beta u_0 - \gamma|u_0| < \beta u_2 + \gamma|u_2|, \tag{4.53}$$

whence $\beta u_1 < \beta u_2$. Note also that $u_1/u_0 \geq 0$ and $\beta u_0 > 0$ imply that $\beta u_1 \geq 0$. Therefore, we have $0 \leq \beta u_1 < \beta u_2 < \beta u_0$ and hence $\min\{u_0, u_1\} < u_2 < \max\{u_0, u_1\}$, showing that u_0 and u_1 are not adjacent, and reaching a contradiction. We have thus proved that $\rho < u_1/u_0$. The proof that $u_1/u_0 < \rho^{-1}$ follows similar arguments. This concludes the proof of Claim A).

Claim A) proves that any two adjacent values u_0 and u_1 of a QS parallel-hyperplane quantiser necessarily satisfy

$$\frac{|\beta| - \gamma}{|\beta| + \gamma} < u_1/u_0 < \frac{|\beta| + \gamma}{|\beta| - \gamma}$$

Moreover, note that, for any $\epsilon > 0$ arbitrarily small, we can always build a logarithmic QS quantiser whose adjacent values satisfy

$$\frac{|\beta| - \gamma}{|\beta| + \gamma} + \epsilon < u_1/u_0 < \left(\frac{|\beta| - \gamma}{|\beta| + \gamma} + \epsilon\right)^{-1}$$

using Theorem 4.18. Hence, it follows that the infimum density $\eta^\star$ is equal to the density of a logarithmic quantiser with base $\rho = \frac{|\beta|-\gamma}{|\beta|+\gamma}$. This concludes the proof. □

The link between CAQS quantisers and quantisation density given by Theorem 4.19 can be directly exploited to design static output feedback laws that employ a quantiser with infimum density, as the following theorem shows.

Theorem 4.20 *Consider system (4.1) with a single output*

$$y(k) = Cx(k), \tag{4.54}$$

where $C \in \mathbb{R}^{1\times n}$. Suppose that a linear feedback $u = \alpha y$ exists that stabilises the single-input system (4.1) with the single-output (4.54), and that $V(x)$, defined in (4.4), is a Lyapunov function for the linear closed-loop system $x(k+1) = (A + \alpha BC)x(k)$. Let q_d be a CAQS quantiser with direction $d = C^T$. Then $\eta(q_d) = \eta^\star$, where

$$\eta^\star = \inf \eta(\mathring{q}), \quad \textit{subject to} \tag{4.55}$$

$$\mathring{q} \textit{ is such that } V \textit{ is a Lyapunov function for } x(k+1) = Ax(k) + B\mathring{q}(Cx(k)), \tag{4.56}$$

and $\eta(\cdot)$ denotes the density of a quantiser, as defined in (4.45)*.*

Proof. Define $q : \mathbb{R}^n \to \mathbb{R}$ by $q(x) = \mathring{q}(Cx)$. Since the range of q, namely $\mathcal{U}(q)$, satisfies $\mathcal{U}(q) = \{q(x) : x \in \mathbb{R}^n\} = \{\mathring{q}(a) : a \in \mathbb{R}\} = \mathcal{U}(\mathring{q})$, then $\eta(q) = \eta(\mathring{q})$. Note also that q is parallel-hyperplane with direction $d = C^T$. We can then rewrite (4.55)–(4.56) as

$$\eta^\star = \inf \eta(q), \quad \text{subject to} \tag{4.57}$$

$$q \text{ is QS and parallel-hyperplane with direction } d = C^T. \tag{4.58}$$

Using Theorem 4.19, the result follows. □

The link between CAQS quantisers and quantisation density given by Theorem 4.19 can be exploited further to recover the result of Elia and Mitter (2001, Theorem 2.1), which derives the optimum quantisation density over all quantisers that are QS for a single-input system with respect to a given CLF.

Theorem 4.21 *Consider the matrix M defined in* (4.5)*, and let q_M be a CAQS quantiser with direction M. Then,*

$$\eta(q_M) = \inf \eta(q), \quad \textit{subject to} \tag{4.59}$$

$$q \textit{ is QS.} \tag{4.60}$$

Proof. Note that we have to prove that $\eta(q_M)$ is the solution to Problem 2.16 in Chapter 2, when we consider the single-input system (4.1). By Theorem 2.17 and Remark 2.20, it follows that we need to optimise density only over parallel-hyperplane quantisers with direction αM, where $\alpha \in \mathbb{R}$. By Remark 4.3, then we need to optimise density only over parallel-hyperplane quantisers with direction M. Therefore, we have

$$\begin{pmatrix} \inf \eta(q), \quad \text{subject to} \\ q \text{ is QS} \end{pmatrix} = \begin{pmatrix} \inf \eta(q), \quad \text{subject to} \\ q \text{ is QS parallel-hyperplane with direction } M \end{pmatrix}. \tag{4.61}$$

By Theorem 4.19, the result then follows. □

Theorem 4.21 shows that a CAQS quantiser with direction $M = A^T PB$ optimises density over all QS quantisers. In Theorem 2.1 of Elia and Mitter (2001), it is shown that a logarithmic parallel-hyperplane quantiser with direction $K_{GD}^T = -A^T PB(B^T PB)^{-1}$ optimises density over all QS quantisers. Recalling Remark 4.3 and since $0 \neq B^T PB \in \mathbb{R}$, it follows that a parallel-hyperplane quantiser with direction M also has direction K_{GD}^T. Thus, Theorem 4.21 recovers the optimum density result in Theorem 2.1 of Elia and Mitter (2001).

We next verify that the expressions for the optimum quantisation density derived according to the results in this thesis and according to Elia and Mitter (2001) actually yield identical values. From (4.23) and (4.25) in Theorem 4.16, a CAQS quantiser with direction d has a range $\mathcal{U}$ satisfying

$$\mathcal{U} = \{\rho^k u_0 : k \in \mathbb{Z}\} \uplus \{-\rho^k u_0 : k \in \mathbb{Z}\} \cup \{0\}, \tag{4.62}$$

where ρ satisfies (4.22), with β and γ as in (4.15). By Theorem 2.12, then the density of a CAQS quantiser with direction d is $2/-\ln\rho$. From Theorem 2.1 of Elia and Mitter (2001), the optimum density is $2/-\ln\tilde{\rho}$, where

$$\tilde{\rho} \triangleq \frac{\sqrt{\frac{B^T PAQ^{-1}A^T PB}{B^T PB}} - 1}{\sqrt{\frac{B^T PAQ^{-1}A^T PB}{B^T PB}} + 1}, \tag{4.63}$$

with

$$Q = P - A^T PA + A^T PB(B^T PB)^{-1}B^T PA. \tag{4.64}$$

The following result then verifies that the density of a CAQS quantiser with direction M coincides with the optimum density given in Theorem 2.1 of Elia and Mitter (2001).

Theorem 4.22 *Consider β and γ, defined in* (4.15)*, with $d = M$, where M was defined in* (4.5)*. Let ρ be defined in* (4.22)*. Then, $\rho = \tilde{\rho}$, where $\tilde{\rho}$ is given in* (4.63) *with Q satisfying* (4.64)*.*

Proof. From (4.22), (4.15), and setting $d = M$, we have

$$\rho = \frac{|\beta| - \gamma}{|\beta| + \gamma} = \frac{|\beta|/\gamma - 1}{|\beta|/\gamma + 1} = \frac{\sqrt{\frac{\beta^2}{\gamma^2}} - 1}{\sqrt{\frac{\beta^2}{\gamma^2}} + 1} = \frac{\sqrt{\frac{M^T L^{-1} M}{-H}} - 1}{\sqrt{\frac{M^T L^{-1} M}{-H}} + 1}. \tag{4.65}$$

Using (4.5), we can rewrite (4.64) as:

$$Q = M(B^T PB)^{-1}M^T - L.$$

Using a matrix inversion formula yields

$$Q^{-1} = -\left[L^{-1} + L^{-1}M(B^T PB - M^T L^{-1}M)^{-1}M^T L^{-1}\right]$$

Then,

$$\begin{aligned}\frac{B^TPAQ^{-1}A^TPB}{B^TPB} &= \frac{M^TQ^{-1}M}{B^TPB} = -\frac{M^TL^{-1}M}{B^TPB}\left[1+\frac{M^TL^{-1}M}{B^TPB-M^TL^{-1}M}\right]\\ &= -\frac{M^TL^{-1}M}{B^TPB}\frac{B^TPB}{B^TPB-M^TL^{-1}M}\\ &= \frac{M^TL^{-1}M}{-H},\end{aligned} \tag{4.66}$$

where we have used (4.5). From (4.65) and (4.66), the result follows. □

Theorem 4.22 thus verifies that the optimum density as given in Elia and Mitter (2001, Theorem 2.1) and the one corresponding to a CAQS quantiser with direction $d = M$ are identical. When the given direction is M or a scalar multiple of it, then Theorem 4.18 becomes the state-space-based counterpart to the construction of quadratically stabilising logarithmic quantisers with density arbitrarily close to the infimum in Theorem 2.1 of Elia and Mitter (2001).

The fact that a CAQS quantiser with direction M optimises quantisation density can be derived directly by optimising density over QS parallel-hyperplane quantisers, as shown by the following result.

Theorem 4.23 *Consider the matrix M defined in* (4.5)*, and let q_M be a CAQS quantiser with direction M. Then,*

$$\eta(q_M) = \inf \eta(q), \quad \textit{subject to} \tag{4.67}$$

$$q \textit{ is QS and parallel-hyperplane.} \tag{4.68}$$

Proof. Note that $\inf \eta(q)$ subject to (4.68) is equivalent to

$$\inf_{d\in\mathbb{R}^n,\, d\neq 0} \begin{pmatrix} \inf \eta(q), \text{ subject to} \\ q \text{ is QS and parallel-hyperplane with direction } d \end{pmatrix}.$$

Using Theorem 4.19, then $\inf \eta(q)$ subject to (4.68) is equivalent to

$$\inf_{d\in\mathbb{R}^n,\, d\neq 0} \eta(q_d),$$

where q_d is CAQS with direction d. From Proposition 4.15, it follows that in order that q_d be CAQS, then $d^TL^{-1}d > 0$. Hence, $\inf \eta(q)$ subject to (4.68) is equivalent to

$$\inf_{d:d^TL^{-1}d>0} \eta(q_d),$$

where q_d is CAQS with direction d. By Theorem 4.16, the range of q_d, namely $\mathcal{U}(q_d)$, satisfies

$$\mathcal{U}(q_d) = \{\rho^j u_0 : j \in \mathbb{Z}\} \cup \{-\rho^j u_0 : j \in \mathbb{Z}\} \cup \{0\},$$

with $0 < \rho < 1$. Then, from Theorem 2.12 it follows that $\eta(q_d) = -2/\ln\rho$. Minimising $\eta(q_d)$ is then equivalent to minimising $\rho = \frac{|\beta|-\gamma}{|\beta|+\gamma}$.

Since $d^T L^{-1} d > 0$ and, by Lemma 4.6 $H < 0$, then γ, defined in (4.15), is real and positive. Therefore,

$$\frac{|\beta| - \gamma}{|\beta| + \gamma} = \frac{|\beta|/\gamma - 1}{|\beta|/\gamma + 1}. \tag{4.69}$$

Using Lemma 4.8, and since $\gamma > 0$, then $|\beta|/\gamma > 1$. The function $g(|\beta|/\gamma)$ defined by the right-hand side of (4.69) is strictly increasing for $|\beta|/\gamma \geq 1$ and hence minimising (4.69) is equivalent to minimising $|\beta|/\gamma$. In turn, this is also equivalent to minimising β^2/γ^2. From (4.15), it follows that

$$f(d) \triangleq \frac{\beta^2}{\gamma^2} = \frac{d^T L^{-1} M M^T L^{-1} d}{-H d^T L^{-1} d}. \tag{4.70}$$

From Lemma 4.8, we have $|\beta| > \gamma$, even in the case when $d^T L^{-1} d = 0$. Hence, the infimum of (4.70) over $d^T L^{-1} d > 0$ cannot occur at a point $d_\star$ satisfying $d_\star^T L^{-1} d_\star = 0$. Therefore, any $d_\star$ that minimises (4.70) necessarily satisfies $\nabla f(d_\star) = 0$. We have

$$\nabla f(d_\star) = 2 \frac{(d_\star^T L^{-1} d_\star) L^{-1} M M^T L^{-1} d_\star - (d_\star^T L^{-1} M M^T L^{-1} d_\star) L^{-1} d_\star}{-H (d_\star^T L^{-1} d_\star)^2} = 0 \tag{4.71}$$

From (4.71), then

$$\left[(d_\star^T L^{-1} d_\star) M M^T L^{-1} - (d_\star^T L^{-1} M M^T L^{-1} d_\star) \mathrm{I} \right] d_\star = 0, \tag{4.72}$$

and hence

$$\begin{aligned} &\det \left[(d_\star^T L^{-1} d_\star) M M^T L^{-1} - (d_\star^T L^{-1} M M^T L^{-1} d_\star) \mathrm{I} \right] = \\ &\quad (-d_\star^T L^{-1} M M^T L^{-1} d_\star)^n + (-d_\star^T L^{-1} M M^T L^{-1} d_\star)^{n-1} (d_\star^T L^{-1} d_\star) M^T L^{-1} M = 0, \end{aligned} \tag{4.73}$$

whence

$$d_\star^T L^{-1} M M^T L^{-1} d_\star = (d_\star^T L^{-1} d_\star) M^T L^{-1} M. \tag{4.74}$$

Substituting (4.74) into (4.72) and since $d_\star^T L^{-1} d_\star > 0$, yields

$$\left[M M^T L^{-1} - (M^T L^{-1} M) \mathrm{I} \right] d_\star = 0,$$

and it follows that any $d_\star$ that satisfies (4.71) necessarily has the form $d_\star = \alpha M$, for some nonzero $\alpha \in \mathbb{R}$. Note that $f(\alpha d) = f(d)$ for all nonzero $\alpha \in \mathbb{R}$ and all nonzero $d \in \mathbb{R}^n$. In particular, $f(\alpha M) = f(M)$, for all nonzero $\alpha \in \mathbb{R}$. Let S denote the bounded set $\{d \in \mathbb{R}^n : \|d\|_2 = \|M\|_2\}$ and let f_S denote the restriction of f to the set S. Note that given any nonzero $d \in \mathbb{R}^n$, $f(d) = f_S(\|M\|_2 \, d / \|d\|_2) = f_S(-\|M\|_2 \, d / \|d\|_2)$. Also, note then that M is a minimiser of f if and only if M is a minimiser of f_S. The function f_S is defined over the bounded set $D \triangleq \{d \in \mathbb{R}^n : \|d\|_2 = \|M\|_2, d^T L^{-1} d > 0\}$, is differentiable in D and tends to infinity whenever d approaches any point in the boundary of D from within D. Since f, and hence f_S, is bounded below by 1, it follows that the only points of f_S where its gradient vanishes, M and $-M$, must correspond to a minimum of f_S. This proves that $d_\star = M$ is a minimiser of (4.70) and the proof is concluded. □

As pointed out in Elia and Mitter (2001), there are an infinite number of ways to define a quantiser so that its density is optimal in the sense that it is the infimum over QS quantisers. The results in this chapter have shown that CAQS quantisers are among those quantisers whose density is optimal. In particular, CAQS quantisers are the ones whose quantisation regions are of the parallel-hyperplane form and contain at most one point at which the increment of the CLF $V(x) = x^T Px$ is nonnegative.

4.6 Example

Consider system (4.1) with the output $y = Cx$, where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & 3 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \quad C^T = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}. \tag{4.75}$$

Next, consider the quadratic function $V(x) = x^T Px$, where

$$P = \begin{bmatrix} 34.0053 & -3.8006 & -18.8372 \\ -3.8006 & 14.3579 & 8.6933 \\ -18.8372 & 8.6933 & 15.0464 \end{bmatrix}. \tag{4.76}$$

The aim is to design an output quantizer, $\mathring{q}$, such that the resulting closed-loop system, $x(k+1) = Ax(k) + B\mathring{q}(Cx(k))$, admits V as a Lyapunov function. In addition, we would like the quantisation density of $\mathring{q}$ to be as low as possible. We will achieve this aim by employing Theorems 4.16, 4.18 and 4.20.

First, we can readily verify that the matrix A is unstable and the pair (A, B) is stabilisable (otherwise the whole stabilisation problem would be trivial). Next, note that the matrix $L = A^T PA - P$ is invertible, which is required in our approach. In addition, the feedback $u = \alpha y$ with $\alpha = 0.49$ stabilises the system, and V is a Lyapunov function for the linear closed-loop system $x(k+1) = (A+\alpha BC)x(k)$. This enables application of Theorem 4.20. According to this theorem, the infimum quantisation density, $\eta^\star$, over all output quantizers $\mathring{q}$ that render V a Lyapunov function for the closed-loop system $x(k+1) = Ax(k)+B\mathring{q}(Cx(k))$ coincides with that of a CAQS quantiser with direction $d = C^T$. We next compute this infimum quantisation density, $\eta^\star$.

From Theorem 4.16, the quantisation values of a CAQS quantiser with direction $d = C^T$ satisfy (4.25), where ρ satisfies (4.22). Using (4.15) with $d = C^T$ and (4.5), we obtain

$$\gamma = 0.1698, \quad \beta = 2.0376, \quad \frac{|\beta| - \gamma}{|\beta| + \gamma} = 0.8462.$$

By Theorem 2.12, the quantisation density of such a quantiser is $-2/\ln 0.8462$. It then follows that the infimum quantisation density over all output quantisers that render the given V a Lyapunov function for

the resulting closed-loop system is given by

$$\eta^\star = \frac{-2}{\ln 0.8462} = 11.9752.$$

Having computed the infimum quantisation density $\eta^\star$, we will next construct the required output quantiser. Recall that the composition of the output equation $y = Cx$ and the output quantiser $u = \mathring{q}(y)$ yields $u = \mathring{q}(Cx) = q(x)$, where q is a parallel-hyperplane quantiser with direction $d = C^T$. As we have previously explained, a CAQS quantiser may not quadratically stabilise the system because each of its quantisation regions may contain a point where the increment of V is nonnegative. However, we may construct a QS quantiser having a density arbitrarily close to $\eta^\star$ by means of Theorem 4.18. We thus select $\epsilon = 0.0038$, yielding

$$\rho = \frac{|\beta| - \gamma}{|\beta| + \gamma} + 0.0038 = 0.85.$$

We will next construct a QS parallel-hyperplane quantiser q with direction $d = C^T$ and with density $\eta(q) = -2/\ln 0.85 = 12.3063$. From Theorem 4.18, the required quantiser, q, satisfies (4.23)–(4.26), and (4.28). We choose $u_0 = 1$, so that $\beta u_0 > 0$, and $\sigma_0 = 1.87$, so that (4.44) is satisfied. Finally, we choose the form (4.28b) for $\mathcal{R}_0$.

Having constructed q, note that the required output quantiser $\mathring{q}$ is then given by $\mathring{q}(Cx) = q(x)$. Figure 4.2 depicts the output quantiser $\mathring{q}$.

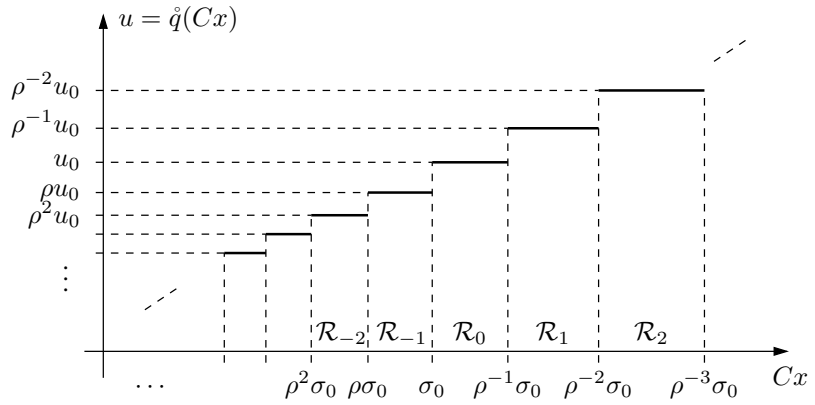
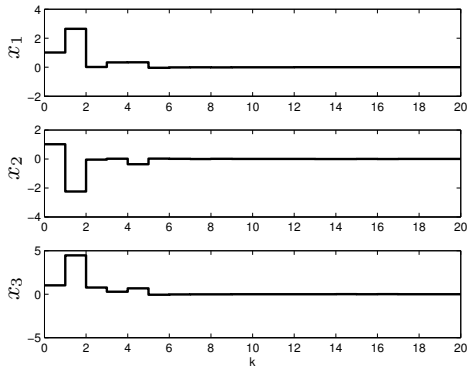
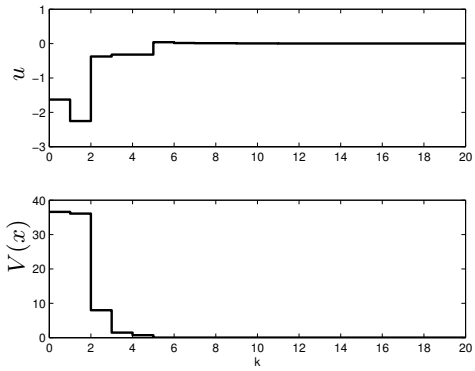


Figure 4.2: Input $u = \mathring{q}(Cx)$ as a function of Cx with $u_0 = 1$, $\sigma_0 = 1.87$ and $\rho = 0.85$.

We simulated the closed-loop system $x(k+1) = Ax(k) + B\mathring{q}(Cx(k))$ from the initial condition $x(0) = [1.015 \;\; 1.015 \;\; 1.015]^T$. Figure 4.3 (a) shows the evolution of the components of the state vector and Figure 4.3 (b) the evolution of the input and CLF. Note from Figure 4.3 (b) that the values $V(x(k))$ decrease to zero as k increases, a consequence of the fact that the output quantiser $\mathring{q}$ renders V a Lyapunov function for the closed-loop system.



(a) $x(k)$ vs k.



(b) $u(k)$ and $V(x(k))$ vs k.

Figure 4.3: Simulation results.

4.7 Chapter Summary

In this chapter, we have dealt with quadratic stabilisation of single-input systems by means of quantised static feedback. We have explored the concept of quantiser coarseness from a state-space standpoint. We followed the intuitive idea that the quantisation regions of a coarse quantiser must be as large as possible within the constraints of interest. These constraints refer to the fact that we seek quantisers that are able to quadratically stabilise a system with respect to a given CLF. We have defined CAQS quantisers and analysed connections with the standard concept of quantisation density. We have shown that a CAQS quantiser with direction d optimises quantisation density over all parallel-hyperplane quantisers with direction d that are quadratically stabilising with respect to the given CLF. We have also shown how CAQS quantisers can be employed to design quadratically stabilising static output feedback laws that utilise quantisers of infimum density. In addition, we have recovered a result of Elia and Mitter on infimum quantisation density through the use of CAQS quantisers.

In conclusion, this chapter has derived precise connections for single-input systems between quantisers whose quantisation *regions* are as large as possible and quantisers having quantisation *values* as separated as possible. Our derivations were made possible by imposing a specific structure on the quantisers considered, namely parallel-hyperplane quantisers.

Chapter 5

Quantisation Density and Multiple-input Systems: A Special Case

5.1 Overview

In the first part of Chapter 3, we have derived necessary and sufficient conditions for quantisers of a specific form to quadratically stabilise a given multiple-input system with respect to a given CLF. We have considered quantisers formed by linear operations and a "reduced-dimension" quantiser. The key feature of the quantisers considered is that the reduced-dimension quantisers operate between spaces having the minimum dimension possible.

In this chapter, we focus on the case when the reduced-dimension quantiser operates between one-dimensional spaces. In this setting, we will derive the infimum density over all quantisers that quadratically stabilise a system with respect to a given CLF. The derivation of this infimum density will require results from previous chapters, including Chapter 4, where we have dealt with single-input systems.

We will show that the infimum density problem considered in this chapter is a special case of a problem considered in Elia and Frazzoli (2002). We will also show, by means of a numerical example, that Theorem 1 of Elia and Frazzoli (2002) is incorrect. Our result will then provide a partial replacement for Theorem 1 of Elia and Frazzoli (2002).

5.2 Problem Statement

As in previous chapters, we consider a system of the following form

$$x(k+1) = Ax(k) + Bu(k), \tag{5.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $x(k) \in \mathbb{R}^n$ and $u(k) \in \mathbb{R}^m$. We assume that A is unstable, B has full column rank and (A, B) is stabilisable.

The problem that we address in this chapter is the derivation of the infimum quantisation density over all symmetric QS quantisers that have the form $q(x) = w\mathring{q}(d^T x)$, with $w \in \mathbb{R}^m$, $d \in \mathbb{R}^n$ and $\mathring{q} : \mathbb{R} \to \mathbb{R}$. We formulate this problem more precisely as follows.

Problem 5.1 *Given system* (5.1) *and a CLF*

$$V(x) = x^T Px, \quad \textit{where } P = P^T > 0, \tag{5.2}$$

such that the matrix

$$L \triangleq A^T PA - P \tag{5.3}$$

is invertible, solve

$$\eta^\star = \inf \eta(q), \quad \textit{subject to:} \tag{5.4}$$

C1) There exist $w \in \mathbb{R}^m$, $d \in \mathbb{R}^n$ and a quantiser $\mathring{q} : \mathbb{R} \to \mathbb{R}$ such that $q(x) = w\mathring{q}(d^T x)$, for all $x \in \mathbb{R}^n$,

C2) q is QS with respect to V,

C3) $q(x) = -q(-x)$ for all $x \in \mathbb{R}^n$,

and where $\eta(q)$ denotes the quantisation density of q.

We thus consider the setting of Figure 5.1. In Chapter 3, Theorem 3.2 states that the lowest possible

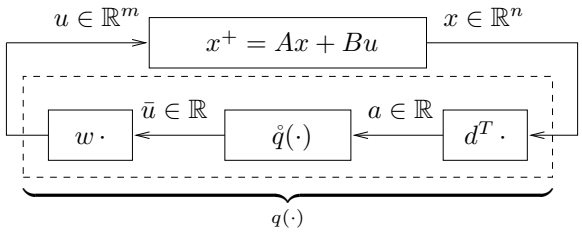


Figure 5.1: The quantised feedback considered: $u = q(x) = w\mathring{q}(d^T x)$.

value of the dimension p for a feedback of the form $q(x) = W\mathring{q}(D^T x)$ with $W \in \mathbb{R}^{m\times p}$ and $D \in \mathbb{R}^{n\times p}$ to be QS is the number of positive eigenvalues of the matrix L in (5.3). Moreover, recalling Remark 3.3, we know that L necessarily has one nonnegative eigenvalue because system (5.1) is open-loop unstable.

Therefore, solving Problem 5.1 is meaningful only when the given CLF (5.2) is such that L in (5.3) has only one positive eigenvalue, or else there would not exist any quantiser q satisfying constraints C1)–C3). This follows because, from C1), we seek QS quantisers of the form $q(x) = w\mathring{q}(d^T x)$, with $w \in \mathbb{R}^{m\times 1}$ and $d \in \mathbb{R}^{n\times 1}$.

5.3 Problem Solution

From Theorem 3.14, it follows that a quantiser $\mathring{q} : \mathbb{R} \to \mathbb{R}$ that causes $q(x) = w\mathring{q}(d^T x)$ to be QS exists if and only if $d^T L^{-1} d > 0$ and $-w^T H w > 0$, with L as in (5.3), and

$$M \triangleq A^T P B \quad \text{and} \quad H \triangleq B^T P B - M^T L^{-1} M. \tag{5.5}$$

In addition, the quantiser $\mathring{q}$ has to satisfy the conditions of Theorem 3.17. We may thus recast constraint C2) of Problem 5.1 as follows:

C2) The vectors $w \in \mathbb{R}^m$ and $d \in \mathbb{R}^n$ satisfy $d^T L^{-1} d > 0$ and $-w^T H w > 0$, and the quantiser $\mathring{q}$ satisfies the conditions of Theorem 3.17, with M and H as defined in (5.5).

Figure 5.1 shows that, conceptually, there exists a single-input system between the fictitious input $\bar{u}$ and the state x. This observation motivates us to divide Problem 5.1 into the following two subproblems. First, we assume that some "feasible" $w \in \mathbb{R}^m$ is given and consider the following single-input system:

$$x(k+1) = Ax(k) + \bar{B}\bar{u}(k), \quad \bar{B} = Bw. \tag{5.6}$$

The results of Chapter 4 can be used to obtain the infimum quantisation density for this single-input system over all quantisers of the form $\mathring{q}(d^T x)$ that are QS with respect to the given CLF. The resulting infimum density is a function of the input matrix $\bar{B} = Bw$ and hence a function of w. Second, we optimise density over all "feasible" vectors $w \in \mathbb{R}^m$. We next show that this procedure indeed leads to the solution to Problem 5.1.

Let $\bar{q} : \mathbb{R}^n \to \mathbb{R}$ be the quantiser defined by $\bar{q}(x) \triangleq \mathring{q}(d^T x)$ for all $x \in \mathbb{R}^n$. Then, we have $q(x) = w\mathring{q}(d^T x) = w\bar{q}(x)$. If $w \neq 0$, then Lemma 2.15 shows that $\eta(q) = \eta(\bar{q})$. We can then divide Problem 5.1 into the following two subproblems.

Subproblem 1 *For a given $w \in \mathbb{R}^m$ and CLF V as in* (5.2)*, where the number of positive eigenvalues of the matrix L in* (5.3) *is* 1*, solve*

$$\eta_w = \inf \eta(\bar{q}), \quad \textit{subject to:}$$

C1') There exists $d \in \mathbb{R}^n$ and a quantiser $\mathring{q} : \mathbb{R} \to \mathbb{R}$ such that $\bar{q}(x) = \mathring{q}(d^T x)$ for all $x \in \mathbb{R}^n$.

C2') The vector $d \in \mathbb{R}^n$ satisfies $d^T L^{-1} d > 0$, and the quantiser $\mathring{q}$ satisfies the conditions of Theorem 3.17.

C3') $\bar{q}(x) = -\bar{q}(-x)$ *for all* $x \in \mathbb{R}^n$.

Subproblem 2

$$\eta^\star = \inf \eta_w, \quad \textit{subject to:}$$

- $w \in \mathbb{R}^m$, *and* $-w^T H w > 0$,

and where η_w *is the solution to Subproblem 1.*

The solution to Subproblem 2 is also the solution to Problem 5.1.

Remark 5.2 *At this point, we can straightforwardly note that the solution to Problem 5.1 is the density of a single-input system of the form* (5.6)*, derived from the given multiple-input system* (5.1)*. This fact is* imposed *by the constraints of the problem and does not necessarily imply that the solution to Problem 5.1 is the infimum density over all quantisers that are QS with respect to the given CLF. In other words, even if the given CLF is such that* L *has only one positive eigenvalue, we are imposing the* additional *constraint that we optimise density only over quantisers* $q : \mathbb{R}^n \to \mathbb{R}^m$ *involving a one-dimensional reduced-dimension quantiser* $\mathring{q}$.

5.3.1 Solution to Subproblem 1

Note that w in Subproblem 1 is fixed. Therefore, η_w is the infimum density over all QS quantisers $\bar{q}$ for the single-input system (5.6). We have dealt with the optimisation of quantisation density for single-input systems in Chapter 4. Then, we may apply the results of Chapter 4 to find η_w. First, we must bear in mind that the results of Chapter 4 will be applied to the single-input system (5.6), and hence B must be replaced by $\bar{B}$, yielding $\bar{M} = A^T P \bar{B} = Mw$ and $\bar{H} = \bar{M}^T L^{-1} \bar{M} - \bar{B}^T P \bar{B} = w^T H w$.

Theorem 4.21 then shows that η_w is the density of a CAQS quantiser with direction $d = \bar{M} = Mw$, and from Theorem 4.16 and Theorem 2.12, we have

$$\eta_w = -2/\ln \frac{|\beta(w)| - \gamma(w)}{|\beta(w)| + \gamma(w)}, \tag{5.7}$$

where

$$\beta(w) \triangleq w^T M^T L^{-1} M w = \bar{M}^T L^{-1} \bar{M}, \quad \gamma(w) \triangleq \sqrt{-w^T H w\, \beta(w)} = \sqrt{-\bar{H} \beta(w)}, \tag{5.8}$$

L was defined in (5.3), and M and H in (5.5).

5.3.2 Solution to Subproblem 2

We are now ready to state the main result of the chapter:

Theorem 5.3 *The solution to Subproblem 2, and hence to Problem 5.1, is given by*

$$\eta^\star = -2/\ln \frac{\beta(w^\star) - \gamma(w^\star)}{\beta(w^\star) + \gamma(w^\star)}, \text{ where} \tag{5.9}$$

$$w^\star = (B^T P B)^{-1/2} v^\star, \tag{5.10}$$

β and γ are defined in (5.8), and $v^\star$ is an eigenvector corresponding to the greatest eigenvalue of the matrix

$$(B^T P B)^{-1/2} M^T L^{-1} M (B^T P B)^{-1/2}, \tag{5.11}$$

with L as defined in (5.3) *and M as in* (5.5).

Proof. Since the density of a quantiser is always nonnegative, we have $\eta_w \geq 0$ and thus any optimiser of Subproblem 2 is also an optimiser of [see (5.7)]

$$\inf \frac{|\beta(w)| - \gamma(w)}{|\beta(w)| + \gamma(w)}, \tag{5.12}$$

subject to $w^T H w < 0$. From (5.5), and since $P > 0$ and B has full column rank, then $w^T H w < 0$ implies that $w^T M^T L^{-1} M w > 0$ and hence $\beta(w) > 0$ [see (5.8)]. Then, $\gamma(w) \neq 0$ and (5.12) is equivalent to

$$\inf \frac{\beta(w)/\gamma(w) - 1}{\beta(w)/\gamma(w) + 1}. \tag{5.13}$$

Also, we have $\gamma^2(w) = -w^T H w\, \beta(w) = [\beta(w) - w^T B^T P B w]\beta(w)$ and then $\gamma^2(w) < \beta^2(w)$. Combining this last inequality with the fact that $\beta(w) > 0$ and $\gamma(w) > 0$ whenever $w^T H w < 0$, then $\beta(w) > \gamma(w) > 0$, and thus $\beta(w)/\gamma(w) > 1$ whenever $w^T H w < 0$. Since the expression to be optimised in (5.13), considered as a function of $\beta(w)/\gamma(w)$, is increasing for $\beta(w)/\gamma(w) > 1$, then any optimiser is also an optimiser of $\inf \beta(w)/\gamma(w)$, which in turn is an optimiser of

$$\inf \frac{\beta^2(w)}{\gamma^2(w)} = \inf \frac{w^T M^T L^{-1} M w}{w^T M^T L^{-1} M w - w^T B^T P B w}, \tag{5.14}$$

subject to $w^T H w < 0$. Since $w^T B^T P B w > 0$, (5.14) is equivalent to

$$\inf \frac{\dfrac{w^T M^T L^{-1} M w}{w^T B^T P B w}}{\dfrac{w^T M^T L^{-1} M w}{w^T B^T P B w} - 1}, \tag{5.15}$$

and since $w^T H w < 0$, using (5.5) then $\frac{w^T M^T L^{-1} M w}{w^T B^T P B w} > 1$ and any optimiser of (5.15) is also an optimiser of

$$\sup \frac{w^T M^T L^{-1} M w}{w^T B^T P B w}. \tag{5.16}$$

Let $v = (B^T P B)^{1/2} w$ and substitute into (5.16) to obtain

$$\sup \frac{v^T (B^T P B)^{-1/2} M^T L^{-1} M (B^T P B)^{-1/2} v}{v^T v}. \tag{5.17}$$

Note that any optimiser $v^\star$ of (5.17) is an eigenvector corresponding to the greatest eigenvalue of the matrix (5.11). Therefore, $w^\star = (B^T P B)^{-1/2} v^\star$ and the result follows. □

Remark 5.4 *Theorem 5.3 solves a quantisation density optimisation problem over all QS quantisers that involve a one-dimensional reduced-dimension quantiser ($\mathring{q}$), when the CLF is* given *and is such that the matrix L in* (5.3) *has only one positive eigenvalue. As anticipated in Remark 5.2, the solution to Problem 5.1 is the infimum density of a single-input system derived from the original multiple-input system. This single-input system is system* (5.6) *with $w = w^\star$. Optimising the solution to Problem 5.1 over all quadratic CLFs that are such that L has only one positive eigenvalue yields the result in Theorem 2.2 of Elia and Mitter (2001). It is important to interpret the correct meaning of the infimum density so derived for the multiple-input system considered. This density is the infimum over all CLFs such that L has only one positive eigenvalue, provided we impose the* additional *constraint that the QS quantisers (over which density is optimised) have values in a one-dimensional subspace of the input space. This resulting density only depends on the unstable eigenvalues of A, that is, it is independent of the input matrix $\bar{B} = Bw$, so long as (A, Bw) is stabilisable.*

5.4 Relationship to a Similar Claim in the Literature

In this section, we review Theorem 1 of Elia and Frazzoli (2002) in order to be able to compare that claim to the one above.

Elia and Frazzoli (2002) define a CLF of the form $V(x) = x^T P x$ to be of Type$_J$ if the number of strictly positive eigenvalues of the matrix $L = A^T P A - P$, defined in (5.3), is J. Then, the CLF considered in this chapter is of Type$_1$. For simplicity, we next restate Theorem 1 in p. 183 of Elia and Frazzoli (2002) using the current notation.

Theorem 5.5 (Theorem 1 in p. 183 of Elia and Frazzoli (2002)) *Let $V(x) = x^T P x$, $P > 0$, be a CLF of Type$_1$ for system* (5.1). *Then $V(x)$ is also a CLF for the single-input system (5.6) obtained by replacing B with $\bar{B} = Bw^\natural$ where $w^\natural = (B^T P B)^{-1} B^T P A v^\natural$, and $v^\natural$ denotes the eigenvector associated with the only positive eigenvalue of $L = A^T P A - P$. Moreover the coarsest (infimum density) quantiser for system (5.1) with respect to such a V is given by*

$$q(x) = w^\natural \bar{q}(x)$$

where $\bar{q}(x)$ is the coarsest quantiser for system (5.6) *with $w = w^\natural$.*

According to this theorem, the infimum density $\eta^\natural \triangleq \eta(\bar{q})$ can be obtained from Elia and Mitter (2001) as:

$$\eta^\natural = -2/\ln \rho, \tag{5.18}$$

where

$$\rho = \frac{\sqrt{\dfrac{\bar{B}^T P A Q^{-1} A^T P \bar{B}}{\bar{B}^T P \bar{B}}} - 1}{\sqrt{\dfrac{\bar{B}^T P A Q^{-1} A^T P \bar{B}}{\bar{B}^T P \bar{B}}} + 1}, \tag{5.19}$$

and

$$Q = P - A^T PA + \frac{A^T P\bar{B}\bar{B}^T PA}{\bar{B}^T P\bar{B}}, \tag{5.20}$$

with $\bar{B} = Bw^\natural$.

In the next section, we show that the result of Theorem 1 of Elia and Frazzoli (2002) is incorrect, and we show to what extent the solution to Problem 5.1 replaces that result.

5.5 Comparison of Results

In this section, we compare the solution to Problem 5.1 obtained in Theorem 5.3 with the claim in Theorem 1 of Elia and Frazzoli (2002) (Theorem 5.5 above) by means of a numerical example. Let system (5.1) be defined with matrices

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and consider the CLF $V(x) = x^T Px$, where

$$P = \begin{bmatrix} 1744 & 3901 & -4574 \\ 3901 & 8809 & -10356 \\ -4574 & -10356 & 12187 \end{bmatrix}.$$

Note that $P = P^T$, and $P > 0$ since the eigenvalues of P are approximately 1, 25 and 22714. The eigenvalues of the matrix L defined in (5.3) are approximately -253, -1 and 146756, showing that L has only one positive eigenvalue and is invertible. According to Elia and Frazzoli (2002), then $V(x)$ is a CLF of Type$_1$.

Evaluating $\eta^\star$ according to Theorem 5.3, that is, according to (5.9) and (5.10), gives

$$\eta^\star = -2/\ln 0.9318 \approx 28.3, \text{ with}$$
$$w^\star = \begin{bmatrix} 0.712 & 0.7022 \end{bmatrix}^T.$$

Also, the same result is obtained by means of (5.18)–(5.20), with $\bar{B} = Bw^\star$. If we evaluate the density using Theorem 5.5, that is, according to Theorem 1 of Elia and Frazzoli (2002), we obtain

$$w^\natural = \begin{bmatrix} -0.8509 & 0.5254 \end{bmatrix}^T,$$

and the argument of the square root in (5.19) is negative, approximately equal to -7.2773, which results in an inconsistent (complex) value of the density. Note also that in this case, $V(x) = x^T Px$ is *not* a CLF for the single-input system (5.6) with $\bar{B} = Bw^\natural$, since Q in (5.20) is not positive definite. The reason for this inconsistency seems to stem from the fact that, in the proof of Theorem 1 of Elia and Frazzoli

(2002), the authors study the increment of V only along the direction of the eigenvector corresponding to the positive eigenvalue of $L = A^T PA - P$.

We have thus shown that Theorem 1 of Elia and Frazzoli (2002) is incorrect. The solution to Problem 5.1, given by Theorem 5.3, replaces Theorem 1 of Elia and Frazzoli (2002) only partially. This is so because, as we have previously mentioned in Remark 5.2, we are imposing the additional constraint that quantisation density be optimised over QS quantisers that have levels in a one-dimensional subspace. On the other hand, Theorem 1 of Elia and Frazzoli (2002) claims that, given a CLF of Type$_1$ and imposing no additional constraints, the infimum density over all QS quantisers for the given multiple-input system corresponds to the infimum density for a single-input system derived from the original multiple-input system. Therefore, Theorem 1 of Elia and Frazzoli (2002) claims that, given a CLF of Type$_1$, the infimum density over all QS quantisers for the given multiple-input system corresponds to the density of a quantiser that has levels in a one-dimensional subspace. Although intuitively this may seem to be the case, at this point we have no proof that such a claim is true (recall also Remark 2.21 in Chapter 2).

Elia and Frazzoli (2002) interpret their Theorem 1 by stating that the results of Elia and Mitter (2001) for single-input systems cannot be improved by the presence of more than one input, when the given CLF is of Type$_1$. We emphasise that we still have no proof of this claim, though it intuitively seems to be true. Elia and Frazzoli (2002) also derive lower bounds on the infimum density with respect to CLFs of Type$_2$. These latter results are not affected by the fact that their results for CLFs of Type$_1$ are incorrect. However, one would have to stress, when referring to the lower bounds derived by Elia and Frazzoli (2002), that they are valid only over CLFs of Type$_2$. Similar considerations have to be born in mind for the results of Elia (2002), as well.

5.6 Chapter Summary

We have derived a new result on infimum quantisation density for linear time-invariant multiple-input systems that can be stabilised using a one-dimensional subspace of the input space. Specifically, we have derived the infimum density over all quantisers that have levels in a one-dimensional subspace and are QS with respect to a given CLF. The infimum density derived was shown to differ from a previously published claim (Elia and Frazzoli, 2002, Theorem 1). This discrepancy was explored by means of a numerical example that shows that the previously published claim is incorrect. This previously published claim also suggests that the infimum density over all quantisers that are QS with respect to a given CLF of a specific type corresponds to the infimum density for a single-input system derived from the original multiple-input system. Whether this latter claim is true or not still remains unanswered.

Part II

Componentwise Ultimate Bounds for Perturbed Systems

Chapter 6

Componentwise Ultimate Bounds for Quantised Systems

6.1 Introduction

6.1.1 Overview of Part II

Part I of the thesis dealt with quadratic stabilisation by means of static feedback employing quantisers. In order to quadratically stabilise a given system by means of such feedback, quantisers having an infinite number of levels were needed. In addition, the required quantisers needed increasingly greater precision towards the origin. For scalar quantisers, these requirements were met by, so-called, logarithmic quantisers.

When quantisers do not have increasingly greater precision towards the origin, asymptotic stabilisation of an open-loop unstable system is not possible, even if the quantisers still have an infinite number of levels (like, for example, uniform quantisers). In this case, however, global practical stabilisation of the system may be possible, that is, any arbitrary initial state may evolve into a bounded region containing the origin. The size of this bounded region is a measure —though not the only one— of the system performance. A typical setting where estimating the size of this bounded region is of interest is when estimating the effects of quantisation on steady-state error specifications in digital control systems (Miller et al., 1989).

When nonlinearities such as quantisers are present, the computation of the smallest bounded region to which the state of the system asymptotically converges is a difficult task. Therefore, different methods exist that obtain bounds on this region (Yakowitz and Parker, 1973; Green and Turner, 1988; Miller et al., 1989; Farrell and Michel, 1989). A feature that is common to all these works is that a quantised signal is regarded as a perturbed copy of the corresponding unquantised signal. This viewpoint contrasts with

that adopted in Part I of the thesis, and gives the impression that disregarding the precise information provided by a quantised signal must necessarily lead to conservativeness in the approach. However, Fu and Xie (2005) have shown that for single-input systems, regarding a logarithmic quantiser as a sector-bound nonlinearity (and hence disregarding precise information on the system state) is not a conservative approach in the context of quadratic stabilisation of discrete-time linear systems.

In this part of the thesis, we regard a quantised variable as a perturbed copy of the unquantised variable. This approach turns a system involving quantisers into a perturbed system. Our goal is then to derive ultimate bounds for perturbed systems. We deal with continuous-time, discrete-time and sampled-data systems and aim to obtain ultimate bounds that are as tight as possible. A key feature of our approach is that we seek *componentwise* ultimate bounds on the system state. Our approach is motivated by the work of Yakowitz and Parker (1973), who obtained componentwise ultimate bounds on the system state by analysing the Jordan canonical form of the system evolution matrix. This work dealt with discrete-time linear time-invariant systems affected by uniform quantisation.

In this chapter, we provide a number of extensions of the approach of Yakowitz and Parker (1973). Specifically, we derive componentwise global ultimate bounds for discrete-time and sampled-data systems involving different types of quantisers. A very important feature of our results is that they can directly accommodate feedback schemes where quantisers of different characteristics and/or types affect different signals in the same system. We further show that our results have application in a recently analysed setting aimed at efficiently utilising available data-rate in networked control systems.

In Chapter 7, we will consider discrete- and continuous-time perturbed systems with more general componentwise perturbation bounds. We will also derive componentwise ultimate bounds for these systems but the global feature of the bounds will be lost. This latter extension of the approach to more general perturbation bounds allows the application of the method to a class of nonlinear systems. We will demonstrate this via several examples.

6.1.2 Ultimate Bound Computation Tools

A standard tool for computing ultimate bounds is based on the use of Lyapunov functions (see, for example, Khalil, 2002, §9.2). This approach is very general and powerful although there is an inherent difficulty associated with the selection of a suitable Lyapunov function. For linear systems, however, quadratic Lyapunov functions can be easily computed and ultimate bounds can be obtained in the form of balls by using the system state 2-norm. This approach is based on bounding the norm of the perturbation and may lead to conservative bounds if information on the perturbation structure is lost when bounding its norm.

A closely related approach to estimate ultimate bounds is via the input-to-state stability (ISS) framework for systems with disturbances (Sontag, 1989; Sontag and Wang, 1995; Jiang and Wang, 2001).

Using this framework, ultimate bounds on the system state trajectories as a function of the disturbance bound can be computed using the system disturbance-to-state asymptotic gain. This gain can be obtained, for example, from a Lyapunov-like characterisation of ISS. As is the case with the Lyapunov-function-based approach, the ISS approach also requires a bound on the norm of the perturbation.

A different approach to estimate ultimate bounds for perturbed continuous-time linear time-invariant (LTI) systems was introduced in Kofman (2005). The latter work derived a closed-form ultimate bound formula based on componentwise analysis of the system in modal coordinates. The result of Kofman (2005) requires componentwise constant perturbation bounds and exploits the system geometry as well as the perturbation structure without utilising a bound on the norm of the perturbation. The examples in Kofman (2005) show that the bounds provided by this method may, in some cases, be much tighter than those obtained by means of the Lyapunov-function-based approach using quadratic functions. The results of Kofman (2005) apply to continuous-time LTI systems and cannot be directly applied to discrete-time or sampled-data systems.

The approach of Kofman (2005) can be seen as the continuous-time counterpart to that of Yakowitz and Parker (1973). The results derived in both works can be applied only to perturbed systems having constant perturbation bounds. However, Kofman (2005) suggests a possible extension of the approach to continuous-time systems with state-dependent perturbation bounds.

In the remainder of this chapter, we derive componentwise ultimate bounds for quantised discrete-time and sampled-data systems. Our approach is based on the analysis of the system dynamics in modal coordinates. In this sense, our results for discrete-time systems can be regarded as extensions of the results of Yakowitz and Parker (1973).

A strong motivation for considering sampled-data systems involving quantisation arises in the control of systems over communication networks (see Antsaklis and Baillieul, Guest Eds., 2004, and the references therein). Quantised sampled-data systems, where a continuous-time plant is controlled via some controller through the use of quantisers and sample and hold devices, have been considered explicitly in, for example, Elia and Mitter (2001); Ishii and Francis (2002b, 2003); Ishii et al. (2004); Ishii and Başar (2005); Fu and Hara (2005). Most of these works deal with the design of quantised control strategies to achieve different objectives in sampled-data systems. Our focus in this part of the thesis is on the analysis, rather than on the design, of a given quantised sampled-data scheme in terms of the computation of componentwise ultimate bounds on the state trajectories.

The remainder of the chapter is organised as follows. In §6.1.3, we introduce notation and some preliminary tools that will be used in the sequel. In §6.2, we present the quantised discrete-time and sampled-data schemes that we consider. In §6.3, we derive the perturbation bounds to utilise when different types of quantisers are employed. In §6.4, we derive componentwise ultimate bounds for the schemes considered. In §6.5, we apply the previous results to several examples. We conclude this chapter in §6.6.

6.1.3 Notation and Preliminary Tools

Notation

If M is a matrix with (real or complex) entries $M_{i,j}$, then $|M|$ and $\mathbb{R}\mathrm{e}(M)$ denote its elementwise magnitude and real part, respectively, that is, $|M|$ is the matrix with entries $|M_{i,j}|$ and $\mathbb{R}\mathrm{e}(M)$, the one with entries $\mathbb{R}\mathrm{e}(M_{i,j})$.

If $x, y \in \mathbb{R}^n$, then $x \preceq y$ and $x \prec y$ denote the sets of componentwise inequalities $x_i \leq y_i$ and $x_i < y_i$, respectively, for $i = 1, \ldots, n$, and similarly for $x \succeq y$ and $x \succ y$. The expression '$x \not\preceq y$' is used as equivalent to '$x \preceq y$ is not true'. Thus, $x \not\preceq y$ does not necessarily imply that $x \succ y$. If $M, N \in \mathbb{R}^{m\times n}$, then $M \preceq N$, $M \prec N$, $M \succeq N$ and $M \succ N$ also denote the corresponding componentwise inequalities.

According to these definitions, it is easy to show that[1]

$$|x + y| \preceq |x| + |y|, \quad |M\,x| \preceq |M| \cdot |x|, \tag{6.1}$$

$$|x| \preceq |y| \Rightarrow |M| \cdot |x| \preceq |M| \cdot |y|, \tag{6.2}$$

whenever $x, y \in \mathbb{C}^n$ and $M \in \mathbb{C}^{m\times n}$.

$\mathbb{R}^n_+$ and $\mathbb{R}^n_{+,0}$ denote the sets of vectors in $\mathbb{R}^n$ with positive and nonnegative components, respectively. Consequently, if $x \in \mathbb{R}^n$ then $x \in \mathbb{R}^n_+ \Leftrightarrow x \succ 0$ and $x \in \mathbb{R}^n_{+,0} \Leftrightarrow x \succeq 0$. Similarly for $\mathbb{R}^{m\times n}_+$ and $\mathbb{R}^{m\times n}_{+,0}$.

$\mathbf{1}_n$ denotes the vector in $\mathbb{R}^n$ all of whose components are equal to 1.

$\boldsymbol{\rho}(\cdot)$ denotes the spectral radius of a square matrix.

For $x, y \in \mathbb{R}^n$, we use the notation $\max\{x, y\} = z$ to denote the vector $z \in \mathbb{R}^n$ with components $z_i = \max\{x_i, y_i\}$, for $i = 1, \ldots, n$.

Ultimate Boundedness

Since our results aim at exploiting the system and perturbation structures, we employ a nonstandard definition of ultimate boundedness that is better suited to this goal. We next provide a standard definition of ultimate boundedness, adapted from Khalil (2002), and derive a preliminary result that links the results obtained later in the chapter to this definition.

Definition 6.1 *The solutions of $\dot{x} = f(t, x)$ are said to be uniformly ultimately bounded with ultimate bound d if there exist a vector norm $\|\cdot\|$ and positive constants d and c, independent of $t_0 \geq 0$, and for every $\alpha \in (0, c)$ there is $\mathcal{T} = \mathcal{T}(\alpha, d) \geq 0$, independent of t_0, such that*

$$\|x(t_0)\| \leq \alpha \Rightarrow \|x(t)\| \leq d, \quad \forall t \geq t_0 + \mathcal{T}. \tag{6.3}$$

[1] For an introduction to the properties of matrices with nonnegative entries, see, for example, Horn and Johnson (1985, §8).

In essence, the results that we will derive guarantee that if $|U^{-1}x(t_0)| \preceq \beta$, then the following implicit ultimate bound holds:

$$|U^{-1}x(t)| \preceq c, \quad \text{for all } t \geq t_0 + \mathcal{T}, \tag{6.4}$$

where $\beta, c \in \mathbb{R}^n_+$, $U \in \mathbb{C}^{n\times n}$ is a nonsingular matrix and $\mathcal{T} \in \mathbb{R}_{+,0}$. Since (6.4) implies the componentwise bound

$$|x(t)| \preceq |U| \cdot |U^{-1}x(t)| \preceq |U|c, \quad \text{for all } t \geq t_0 + \mathcal{T},$$

then the following Lemma shows that such results lead to ultimate bounds in the sense of Definition 6.1.

Lemma 6.2 *Consider the system $\dot{x} = f(t,x)$, where $x(t) \in \mathbb{R}^n$, and suppose that there exist $\beta, b \in \mathbb{R}^n_+$, $\mathcal{T} \in \mathbb{R}_{+,0}$ and a nonsingular matrix $U \in \mathbb{C}^{n\times n}$ such that, irrespective of t_0,*

$$|U^{-1}x(t_0)| \preceq \beta \Rightarrow |x(t)| \preceq b, \quad \forall t \geq t_0 + \mathcal{T}.$$

Then, the solutions of $\dot{x} = f(t,x)$ are uniformly ultimately bounded.

Proof. Let $\beta_{\min} \triangleq \min_i \beta_i$, $b_{\max} \triangleq \|b\|_\infty$, where $\|\cdot\|_\infty$ denotes the infinity norm of a vector and the corresponding induced norm for a matrix. Note that $\beta_{\min} > 0$. Then, for any $\alpha \in (0, \beta_{\min}/\left\|U^{-1}\right\|_\infty)$, we have

$$\begin{aligned}
\|x(t_0)\|_\infty < \alpha &\Rightarrow \left\|U^{-1}x(t_0)\right\|_\infty < \left\|U^{-1}\right\|_\infty \alpha \\
&\Rightarrow \left\|U^{-1}x(t_0)\right\|_\infty < \beta_{\min} \\
&\Rightarrow |U^{-1}x(t_0)| \prec \beta \\
&\Rightarrow |x(t)| \preceq b \Rightarrow \|x(t)\|_\infty \leq b_{\max},
\end{aligned}$$

for all $t \geq t_0 + \mathcal{T}$. This concludes the proof. □

The discrete-time counterparts to Definition 6.1 and Lemma 6.2 are straightforward.

6.2 Quantised System Description

In this section, we present the quantised discrete-time and sampled-data schemes that we consider. The main feature of these schemes is that each individual signal connecting plant and controller may be affected by an independent scalar quantiser. The scalar quantisers affecting the different signals can be of different types and characteristics. In §6.2.1 and §6.2.2, we present the quantised discrete-time and sampled-data schemes, respectively.

6.2.1 Quantised Discrete-time Scheme

We consider a discrete-time plant connected to a discrete-time controller, as shown in Figure 6.1. Each component of the plant and controller outputs has an independent scalar quantiser, which can be of any of the following three types: uniform, logarithmic, and semitruncated logarithmic (see §6.3). Our derivations do not require all quantisers to have the same features, nor to be of the same type. In addition, cases where quantisation affects some (but not necessarily all) connecting signals are also directly dealt with by our method. The plant and controller in Figure 6.1 can be described by the following equations:

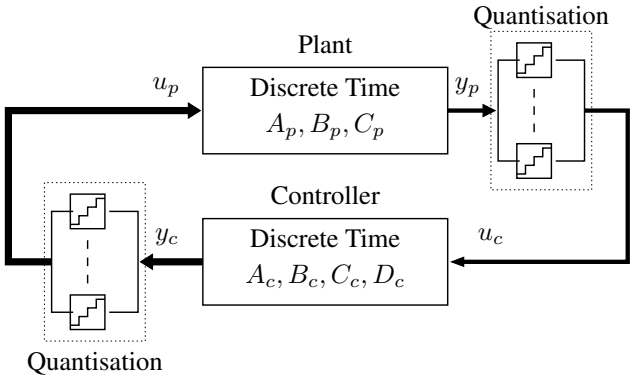


Figure 6.1: Quantised discrete-time control scheme.

$$\begin{aligned} x_p(k+1) &= A_p x_p(k) + B_p u_p(k), && \text{(6.5a)} \\ y_p(k) &= C_p x_p(k), && \text{(6.5b)} \\ x_c(k+1) &= A_c x_c(k) + B_c u_c(k), && \text{(6.5c)} \\ y_c(k) &= C_c x_c(k) + D_c u_c(k), && \text{(6.5d)} \end{aligned}$$

where $x_p(k) \in \mathbb{R}^{\mathrm{N}_p}$, $u_p(k) \in \mathbb{R}^{\mathrm{M}}$ and $y_p(k) \in \mathbb{R}^{\mathrm{P}}$ are the plant state, input and output, respectively, and $x_c(k) \in \mathbb{R}^{\mathrm{N}_c}$, $u_c(k) \in \mathbb{R}^{\mathrm{P}}$ and $y_c(k) \in \mathbb{R}^{\mathrm{M}}$ are the controller state, input and output, respectively. For future reference, we define n as the total number of system states (plant + controller), and S as the total maximum number of quantised signals:

$$n \triangleq \mathrm{N}_p + \mathrm{N}_c, \quad \mathrm{S} \triangleq \mathrm{P} + \mathrm{M}. \tag{6.6}$$

We may express the connections between controller and plant as follows:

$$\begin{aligned} u_c(k) &= y_p(k) + \Delta y_p(k), && \text{(6.7a)} \\ u_p(k) &= y_c(k) + \Delta y_c(k), && \text{(6.7b)} \end{aligned}$$

where Δy_p and Δy_c are the perturbations introduced by the quantisers at the plant and controller outputs, respectively. Defining

$$x_k \triangleq \begin{bmatrix} x_p(k) \\ x_c(k) \end{bmatrix} \in \mathbb{R}^n, \quad \Delta y_k \triangleq \begin{bmatrix} \Delta y_p(k) \\ \Delta y_c(k) \end{bmatrix} \in \mathbb{R}^{\mathrm{S}}, \tag{6.8}$$

and combining (6.5) and (6.7), we can write

$$x_{k+1} = A_d x_k + B_d \Delta y_k, \tag{6.9}$$

where

$$A_d = \begin{bmatrix} A_p + B_p D_c C_p & B_p C_c \\ B_c C_p & A_c \end{bmatrix}, \quad B_d = \begin{bmatrix} B_p D_c & B_p \\ B_c & 0 \end{bmatrix}. \tag{6.10}$$

6.2.2 Quantised Sampled-data Scheme

We consider the sampled-data system of Figure 6.2, where an LTI continuous-time plant is controlled via a discrete-time controller using quantisers, and sampling and zero-order hold devices. As in the discrete-

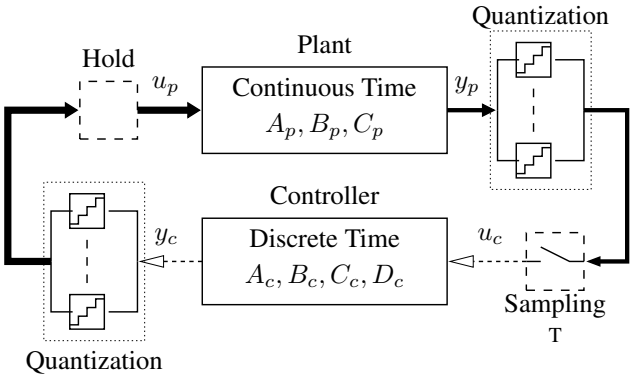


Figure 6.2: Quantised sampled-data control scheme.

time scheme in §6.2.1, each component of the plant and controller outputs has an independent scalar quantiser, which can be of any of the following three types: uniform, logarithmic, and semitruncated logarithmic (see §6.3).

The plant and controller in Figure 6.2 can be described by the following equations:

$$\dot{x}_p(t) = A_p x_p(t) + B_p u_p(t), \tag{6.11a}$$

$$y_p(t) = C_p x_p(t), \tag{6.11b}$$

$$x_c(k+1) = A_c x_c(k) + B_c u_c(k), \tag{6.11c}$$

$$y_c(k) = C_c x_c(k) + D_c u_c(k), \tag{6.11d}$$

where $x_p(t) \in \mathbb{R}^{\mathrm{N}_p}$, $u_p(t) \in \mathbb{R}^{\mathrm{M}}$ and $y_p(t) \in \mathbb{R}^{\mathrm{P}}$ are the continuous-time plant state, input and output, respectively, and $x_c(k) \in \mathbb{R}^{\mathrm{N}_c}$, $u_c(k) \in \mathbb{R}^{\mathrm{P}}$ and $y_c(k) \in \mathbb{R}^{\mathrm{M}}$ are the discrete-time controller state, input

and output, respectively. As in the purely discrete-time case, we define n and S as the total number of system states (plant + controller) and the total maximum number of quantised signals, respectively. Recall that n and S satisfy (6.6). We may express the connections between controller and plant as follows, where we also take account of sampling and hold:

$$u_c(k) = y_p(t_k) + \Delta y_p(k), \tag{6.12a}$$

$$u_p(t) = y_c(k) + \Delta y_c(k), \quad t_k \le t < t_k + \mathrm{T}, \tag{6.12b}$$

where T is the sampling period,

$$t_k = k\mathrm{T}, \quad k = 0, 1, \ldots,$$

and Δy_p and Δy_c are the perturbations introduced by the quantisers at the plant and controller outputs, respectively. Combining (6.11b), (6.11d) and (6.12) yields

$$u_c(k) = C_p x_p(t_k) + \Delta y_p(k),$$
$$u_p(t) = C_c x_c(k) + D_c C_p x_p(t_k) + D_c \Delta y_p(k) + \Delta y_c(k), \quad \text{for } t_k \le t < t_{k+1}.$$

The model of the system of Figure 6.2 at the sampling instants is

$$x_{k+1} = A_d x_k + B_d \Delta y_k, \tag{6.13}$$

where we have defined

$$x_k \triangleq \begin{bmatrix} x_p(t_k) \\ x_c(k) \end{bmatrix} \in \mathbb{R}^n, \quad \Delta y_k \triangleq \begin{bmatrix} \Delta y_p(k) \\ \Delta y_c(k) \end{bmatrix} \in \mathbb{R}^{\mathrm{S}}, \tag{6.14}$$

and A_d and B_d can be readily obtained from (6.11) and (6.12) as

$$A_d = \begin{bmatrix} A_{11} & A_{12} \\ B_c C_p & A_c \end{bmatrix}, \quad B_d = \begin{bmatrix} B_{11} & B_{12} \\ B_c & 0 \end{bmatrix}, \tag{6.15}$$

with

$$A_{11} \triangleq e^{A_p \mathrm{T}} + \Psi(\mathrm{T}) B_p D_c C_p, \tag{6.16}$$

$$A_{12} \triangleq \Psi(\mathrm{T}) B_p C_c, \tag{6.17}$$

$$B_{11} \triangleq \Psi(\mathrm{T}) B_p D_c, \quad B_{12} \triangleq \Psi(\mathrm{T}) B_p, \tag{6.18}$$

$$\Psi(t) \triangleq \int_0^t e^{A_p \tau} d\tau. \tag{6.19}$$

Remark 6.3 *The derivation of ultimate bounds for the sampled-data scheme of Figure 6.2 will be performed in two stages. In the first stage, we derive ultimate bounds that are valid only at the sampling instants by analysing system* (6.13)*. The second stage then derives bounds on the continuous-time plant states that are valid at all times greater than a finite time, that is, not just at the sampling instants. Since* (6.13) *is identical to* (6.9)*, it is clear that the first stage will directly employ the results derived for discrete-time systems.*

6.3 Quantiser Perturbations

In this section, we show how to bound the different perturbations introduced by the quantisers, according to the type of quantiser employed. Recall that, in the schemes that we consider, each component of the plant and controller outputs may have an independent scalar quantiser. In §6.3.1, we derive the different perturbation bounds for the three types of scalar quantisers considered, in terms of the corresponding unquantised signal. In §6.3.2, we show how to utilise these bounds to express componentwise bounds on the quantiser perturbation vector Δy_k [defined in (6.8) or (6.14)], in a form suitable to the subsequent derivations.

6.3.1 Single Scalar Quantiser

Given a scalar quantiser $q : \mathbb{R} \to \mathbb{R}$, we regard the quantised variable $q(s)$ as a perturbed copy of the unquantised variable s:

$$q(s) = s + \Delta s.$$

We next explain the different quantiser types considered (uniform, logarithmic, and semitruncated logarithmic) and derive the corresponding bounds on Δs. For simplicity, we introduce the different quantiser types using symmetric quantisers, that is, quantisers that satisfy $q(s) = -q(-s)$, for all $s \in \mathbb{R}$. Note however that our derivations hold for any quantiser whose corresponding perturbation Δs can be bounded according to the expression that we derive [see (6.24)].

Uniform quantiser

A uniform quantiser has uniformly spaced levels, as shown in Figure 6.3. In this case, the quantiser

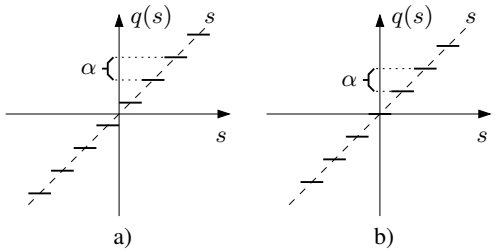


Figure 6.3: Uniform quantisers: a) Midrise. b) Midtread.

perturbation $\Delta s = q(s) - s$ can be bounded by

$$|\Delta s| \leq u^\circ \triangleq \alpha/2, \tag{6.20}$$

where α is the quantisation step, as shown in Figure 6.3.

Logarithmic quantiser

A symmetric logarithmic quantiser has levels in a set $W \subset \mathbb{R}$ satisfying

$$W = \{\pm\rho^{-i}u^\circ, i = 0, \pm1, \pm2, \dots\} \cup \{0\}, \tag{6.21}$$

where $0 < \rho < 1$ and $u^\circ > 0$. We will consider logarithmic quantisers as that depicted in Figure 6.4 a) for positive values of the unquantised variable. For this type of quantiser, the corresponding quantiser

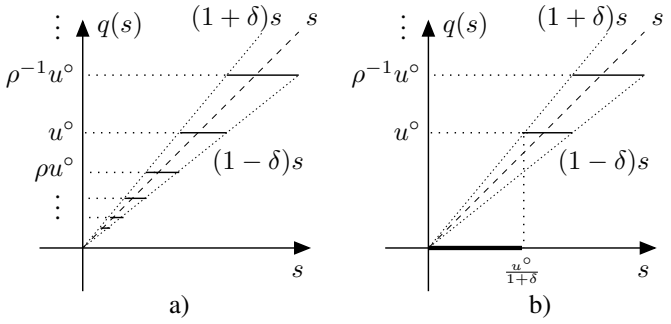


Figure 6.4: a) Logarithmic quantiser. b) Semitruncated logarithmic quantiser.

perturbation $\Delta s = q(s) - s$ satisfies

$$|\Delta s| \leq \delta|s|. \tag{6.22}$$

Thus, the quantity δ represents the maximum relative error of the logarithmically quantised variable. From Figure 6.4 a), it can easily be verified that

$$\delta = \frac{1-\rho}{1+\rho}. \tag{6.23}$$

Semitruncated logarithmic quantiser

Practical logarithmic quantisers arise from truncating a logarithmic quantiser so that the resulting quantiser has only a finite number of levels. We consider a semitruncated quantiser in the sense that it is truncated only towards the origin, that is, it has values in the following set [cf. (6.21)]:

$$W = \{\pm\rho^{-i}u^\circ, i = 0, 1, \dots\} \cup \{0\},$$

where $0 < \rho < 1$ and $u^\circ > 0$. This form of quantiser is illustrated in Figure 6.4 b) for positive values of the unquantised variable. As seen from this figure, the quantiser perturbation $\Delta s = q(s) - s$ satisfies

$$|\Delta s| \leq \max\left\{\delta|s|, \frac{u^\circ}{1+\delta}\right\}. \tag{6.24}$$

Note that (6.24) encompasses the three types of quantiser perturbations considered, that is, (6.20) can be obtained from (6.24) by setting $\delta = 0$ and $u^\circ = \alpha/2$, and (6.22) can be obtained by setting $u^\circ = 0$.

6.3.2 Quantiser Perturbations in Vector Form

We next utilise the bounds derived in §6.3.1 to express a bound on the quantiser perturbation vector Δy_k, defined in (6.8) or (6.14), in a form that is needed in §6.4 for the derivation of componentwise ultimate bounds. Let Δy_{p_i}, for $i = 1, \ldots, \mathrm{P}$, denote the i-th component of Δy_p, and Δy_{c_j}, for $j = 1, \ldots, \mathrm{M}$, denote the j-th component of Δy_c, where Δy_p and Δy_c are the perturbations introduced by the quantisers at the plant and controller outputs, respectively [recall (6.7) and (6.12)]. Since (6.24) encompasses (6.20) and (6.22) as special cases, irrespective of the type of quantiser affecting each signal, we can write

$$|\Delta y_{p_i}(k)| \leq \max\left\{\delta_i |y_{p_i}(k)|, \frac{u_i^\circ}{1+\delta_i}\right\}, \tag{6.25}$$

$$|\Delta y_{c_j}(k)| \leq \max\left\{\delta_{j+\mathrm{P}} |y_{c_j}(k)|, \frac{u_{j+\mathrm{P}}^\circ}{1+\delta_{j+\mathrm{P}}}\right\}, \tag{6.26}$$

for $i = 1, \ldots, \mathrm{P}$, $j = 1, \ldots, \mathrm{M}$ and all $k \geq 0$. In (6.25) and (6.26), δ_i and $\delta_{j+\mathrm{P}}$ correspond to the quantiser at the i-th plant output and j-th controller output, respectively, and are zero if that plant or controller output is uniformly quantised or if they are not quantised. Similarly, u_i° and $u_{j+\mathrm{P}}^\circ$ correspond to the quantiser at the i-th plant output and j-th controller output, respectively, and are zero if that plant or controller output is logarithmically quantised or if they are not quantised. Recalling (6.6) and defining

$$\Gamma_p \triangleq \mathrm{diag}(\delta_1, \ldots, \delta_{\mathrm{P}}), \tag{6.27}$$

$$\Gamma_c \triangleq \mathrm{diag}(\delta_{\mathrm{P}+1}, \ldots, \delta_{\mathrm{S}}), \tag{6.28}$$

$$\theta_p \triangleq \begin{bmatrix} \frac{u_1^\circ}{1+\delta_1} & \cdots & \frac{u_{\mathrm{P}}^\circ}{1+\delta_{\mathrm{P}}} \end{bmatrix}^T, \tag{6.29}$$

$$\theta_c \triangleq \begin{bmatrix} \frac{u_{\mathrm{P}+1}^\circ}{1+\delta_{\mathrm{P}+1}} & \cdots & \frac{u_{\mathrm{S}}^\circ}{1+\delta_{\mathrm{S}}} \end{bmatrix}^T, \tag{6.30}$$

we can express (6.25) and (6.26) in vector form as

$$|\Delta y_p(k)| \preceq \max\{\Gamma_p |y_p(k)|, \theta_p\}, \tag{6.31}$$

$$|\Delta y_c(k)| \preceq \max\{\Gamma_c |y_c(k)|, \theta_c\}, \tag{6.32}$$

where the maximum is taken componentwise. In (6.31)–(6.32), $y_p(k)$ and $y_c(k)$ satisfy, from (6.5) and (6.7),

$$y_p(k) = C_p x_p(k), \tag{6.33}$$

$$y_c(k) = C_c x_c(k) + D_c C_p x_p(k) + D_c \Delta y_p(k). \tag{6.34}$$

In §6.4, we will need bounds on $\Delta y_p(k)$ and $\Delta y_c(k)$ in terms of a linearly transformed version of the state. Therefore, let $U \in \mathbb{C}^{n \times n}$ denote an arbitrary invertible matrix and consider the transformation $x_k = U z_k$, where U is partitioned according to x_k in (6.8) as

$$U = \begin{bmatrix} U_p \\ U_c \end{bmatrix}, \quad \text{with } U_p \in \mathbb{C}^{\mathrm{N}_p \times n} \text{ and } U_c \in \mathbb{C}^{\mathrm{N}_c \times n}. \tag{6.35}$$

Operating on (6.31)–(6.34), and using (6.35) yields

$$|\Delta y_p(t_k)| \preceq \max\{\Theta_p|z_k|, \theta_p\}, \tag{6.36a}$$

$$|\Delta y_c(k)| \preceq \max\Big\{\Theta_c|z_k| + \Theta_s \max\{\Theta_p|z_k|, \theta_p\}, \theta_c\Big\}, \tag{6.36b}$$

where we have used properties (6.1) and (6.2), and defined

$$\Theta_p \triangleq \Gamma_p|C_pU_p|, \tag{6.37}$$

$$\Theta_c \triangleq \Gamma_c|C_cU_c + D_cC_pU_p|, \tag{6.38}$$

$$\Theta_s \triangleq \Gamma_c|D_c|. \tag{6.39}$$

Remark 6.4 *We emphasise that our approach allows for any combination of uniform, logarithmic, semitruncated logarithmic and no quantisation at the plant and controller outputs, that is, the quantisers need not be all of the same type, need not all have the same features, and not all signals need to be quantised. The corresponding perturbation bound of the form (6.36) can be obtained by adjusting the entries of Γ_p, Γ_c, θ_p and θ_c to match the type of quantisation used for each signal, as explained before.*

6.4 Componentwise Ultimate Bounds for Quantised Systems

In this section, we derive upper bounds on the individual components of the system state. Our derivations utilise the componentwise perturbation bound (6.36). In §6.4.1, we deal with the quantised discrete-time system depicted in Figure 6.1. In §6.4.2, we treat the sampled-data system of Figure 6.2.

6.4.1 Discrete-time Systems

We next derive componentwise global ultimate bounds on the state of the discrete-time perturbed system (6.8)–(6.9) when the perturbation Δy_k is bounded as in (6.36). In essence, our results stem from the application of a comparison principle to each component of the state vector, as follows. Given a discrete-time system $x(k+1) = Ax(k) + v(k)$, where $|v(k)| \preceq b$ for all k, then $|x(k)| \preceq b(k)$, where the sequence $b(k)$ satisfies $b(0) = |x(0)|$ and $b(k+1) = |A|b(k) + b$. To prove this by induction, note that $|x(0)| \preceq b(0)$ by definition. Then, assume that $|x(k)| \preceq b(k)$. We have:

$$\begin{aligned}|x(k+1)| &= |Ax(k) + v(k)| \preceq |Ax(k)| + |v(k)| \preceq |A||x(k)| + b\\ &\preceq |A|b(k) + b = b(k+1),\end{aligned}$$

which establishes by induction that $|x(k)| \preceq b(k)$ for all $k \geq 0$. Note that we cannot directly apply this procedure to derive ultimate bounds for the discrete-time perturbed system considered because (a) the perturbation bound (6.36) has a more complicated form and (b) the matrix $|A|$ may not be stable even if A is, causing $\lim_{k\to\infty} b(k)$ to not be finite. However, the foregoing explanation captures the essence of the following theorem.

Theorem 6.5 *Consider system (6.8)–(6.9) and express A_d in Jordan canonical form as $A_d = U\Lambda U^{-1}$. Let Δy_k be bounded as in (6.36) for all $k \geq 0$, where $z_k = U^{-1}x_k$ and $\Theta_p \in \mathbb{R}^{\mathrm{P}\times n}_{+,0}$, $\Theta_c \in \mathbb{R}^{\mathrm{M}\times n}_{+,0}$, $\theta_p \in \mathbb{R}^{\mathrm{P}}_{+,0}$, $\theta_c \in \mathbb{R}^{\mathrm{M}}_{+,0}$ and $\Theta_s \in \mathbb{R}^{\mathrm{M}\times \mathrm{P}}_{+,0}$. Define*

$$M \triangleq |\Lambda| + |U^{-1}B_d|\,\Theta, \tag{6.40}$$

$$\Theta \triangleq \begin{bmatrix} \Theta_p \\ \Theta_c + \Theta_s\Theta_p \end{bmatrix}, \tag{6.41}$$

$$\theta \triangleq \begin{bmatrix} \theta_p \\ \theta_c + \Theta_s\theta_p \end{bmatrix}, \tag{6.42}$$

and suppose that $\boldsymbol{\rho}(M) < 1$, where $\boldsymbol{\rho}(\cdot)$ denotes the spectral radius of a square matrix. Then, $\boldsymbol{\rho}(|\Lambda|) < 1$. Define

$$\beta \triangleq (\mathrm{I} - M)^{-1}\,|U^{-1}B_d|\,\theta, \tag{6.43}$$

$$\gamma \triangleq (\mathrm{I} - |\Lambda|)^{-1}\,|U^{-1}B_d|\,[\theta_p^T\ \theta_c^T]^T. \tag{6.44}$$

Then, $\beta \in \mathbb{R}^n_{+,0}$ and $\gamma \in \mathbb{R}^n_{+,0}$. Consider the map $T : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ defined by

$$T(w) = |\Lambda|w + |U^{-1}B_d| \begin{bmatrix} \max\{\Theta_p w, \theta_p\} \\ \max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} \end{bmatrix} \tag{6.45}$$

and the sequence $\{b_r\}_{r=0}^{\infty}$ defined by

$$b_0 \triangleq \beta, \quad b_r \triangleq T(b_{r-1}),\ \textit{for } r = 1, 2, \ldots. \tag{6.46}$$

Then,

1. *$0 \preceq b_r \preceq b_{r-1}$ for $r = 1, 2, \ldots,$ and $b_\infty \triangleq \lim_{r\to\infty} b_r$ exists and satisfies $\gamma \preceq b_\infty \preceq \beta$.*
2. *If $|U^{-1}x_0| \preceq b_r$ for some $0 \leq r \leq \infty$, then, for all $k \geq 0$,*
   a) *$|U^{-1}x_k| \preceq b_r$.*
   b) *$|x_k| \preceq |U|b_r$.*
3. *Given any $\epsilon \in \mathbb{R}^n_+$ and $x_0 \in \mathbb{R}^n$, there exists $\ell = \ell(\epsilon, x_0) \geq 0$ such that, for all $k \geq \ell$ and all $0 \leq r \leq \infty$,*
   a) *$|U^{-1}x_k| \preceq b_r + \epsilon$.*
   b) *$|x_k| \preceq |U|b_r + |U|\,\epsilon$.*

Proof. Note that $|\Lambda|$, $|U^{-1}B_d|$, Θ and M all have nonnegative entries. Then, from (6.40), it follows that $\boldsymbol{\rho}(M) \geq \boldsymbol{\rho}(|\Lambda|) + \boldsymbol{\rho}(|U^{-1}B_d|\Theta)$ (see, for example, Horn and Johnson, 1985, §8.1), and by assumption,

$\boldsymbol{\rho}(M) < 1$. Since the spectral radius is a nonnegative quantity, it follows that $\boldsymbol{\rho}(|\Lambda|) < 1$. Note then that $\mathrm{I} - M$ and $\mathrm{I} - |\Lambda|$ are invertible and β and γ in (6.43) and (6.44), respectively, are well defined.

Define the maps $T_M : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ and $T_\Lambda : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ as

$$T_M(w) = Mw + |U^{-1}B_d|\,\theta \quad \text{and} \quad T_\Lambda(w) = |\Lambda|w + |U^{-1}B_d| \begin{bmatrix} \theta_p \\ \theta_c \end{bmatrix}. \tag{6.47}$$

Note that, for any $w \in \mathbb{R}^n_{+,0}$, $T_M(w) \succeq 0$ and $T_\Lambda(w) \succeq 0$, and hence $T^r_M(w) \succeq 0$ and $T^r_\Lambda(w) \succeq 0$, for all $r \geq 0$ and all $w \in \mathbb{R}^n_{+,0}$. Since $\boldsymbol{\rho}(M) < 1$ and $\boldsymbol{\rho}(|\Lambda|) < 1$, then $\lim_{r\to\infty} T^r_M(w)$ and $\lim_{r\to\infty} T^r_\Lambda(w)$ exist and satisfy

$$\lim_{r\to\infty} T^r_M(w) = \beta \succeq 0 \quad \text{and} \quad \lim_{r\to\infty} T^r_\Lambda(w) = \gamma \succeq 0, \tag{6.48}$$

for all $w \in \mathbb{R}^n_{+,0}$. This establishes that $\beta \in \mathbb{R}^n_{+,0}$ and $\gamma \in \mathbb{R}^n_{+,0}$.

1. Since the matrices $|\Lambda|$, $|U^{-1}B_d|$, Θ_p, Θ_c, and Θ_s, and the vectors θ_p and θ_c all have nonnegative entries, it follows that the maps T in (6.45), and T_M and T_Λ in (6.47) have the following property:

$$w_1 \preceq w_2 \Longrightarrow \begin{cases} T(w_1) \preceq T(w_2), \\ T_M(w_1) \preceq T_M(w_2), \\ T_\Lambda(w_1) \preceq T_\Lambda(w_2), \end{cases} \tag{6.49}$$

for all $w_1, w_2 \in \mathbb{R}^n_{+,0}$. Since $\Theta_p \succeq 0$ and $\theta_p \succeq 0$, it follows that

$$0 \preceq \theta_p \preceq \max\{\Theta_p w, \theta_p\} \preceq \Theta_p w + \theta_p, \tag{6.50}$$

for all $w \in \mathbb{R}^n_{+,0}$. In addition, using (6.50) and since $\Theta_c \succeq 0$, $\Theta_s \succeq 0$ and $\theta_c \succeq 0$, then

$$0 \preceq \theta_c \preceq \max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} \preceq (\Theta_c + \Theta_s\Theta_p)w + (\theta_c + \Theta_s\theta_p), \tag{6.51}$$

for all $w \in \mathbb{R}^n_{+,0}$. Using (6.40)–(6.42), (6.45), (6.47), (6.50), and (6.51), we then have

$$T_\Lambda(w) = |\Lambda|w + |U^{-1}B_d| \left[\begin{smallmatrix} \theta_p \\ \theta_c \end{smallmatrix}\right] \preceq T(w) \preceq |\Lambda|w + |U^{-1}B_d|(\Theta w + \theta) = T_M(w), \tag{6.52}$$

for all $w \in \mathbb{R}^n_{+,0}$. Also, note that (6.49) and (6.52) imply that

$$T^r_\Lambda(w) \preceq T^r(w) \preceq T^r_M(w), \quad \text{for all } w \in \mathbb{R}^n_{+,0}, \quad \text{for } r = 1, 2, \ldots. \tag{6.53}$$

From (6.43) and (6.44), note that β and γ satisfy $M\beta + |U^{-1}B_d|\theta = \beta$ and $|\Lambda|\gamma + |U^{-1}B_d|[\theta_p^T\ \theta_c^T]^T = \gamma$, and using (6.47), then

$$\beta = T_M(\beta) \quad \text{and} \quad \gamma = T_\Lambda(\gamma). \tag{6.54}$$

Using (6.46), (6.52) and (6.54), it follows that $b_1 = T(\beta) \preceq T_M(\beta) = \beta = b_0$, whence $b_1 \preceq b_0$. Applying (6.49) iteratively to the latter inequality yields $b_r = T^r(\beta) \preceq T^{r-1}(\beta) = b_{r-1}$. The sequence $\{b_r\}_{r=0}^\infty$ is thus componentwise nonincreasing. Moreover, this sequence is lower bounded by the convergent sequence $T^r_\Lambda(\beta)$. Hence, $\{b_r\}_{r=0}^\infty$ converges to some point $b_\infty = \lim_{r\to\infty} b_r$ satisfying

$$\gamma = \lim_{r\to\infty} T^r_\Lambda(\beta) \preceq b_\infty \preceq \beta.$$

This establishes 1.

Let $x_k = Uz_k$ and substitute into (6.9) to obtain

$$z_{k+1} = \Lambda z_k + U^{-1}B_d\Delta y_k.$$

Taking magnitudes and using (6.8) and (6.36) yields

$$\begin{aligned}|z_{k+1}| &\preceq |\Lambda|\,|z_k| + |U^{-1}B_d||\Delta y_k| \\ &\preceq |\Lambda|\,|z_k| + |U^{-1}B_d| \begin{bmatrix} \max\{\Theta_p|z_k|, \theta_p\} \\ \max\Big\{\Theta_c|z_k| + \Theta_s \max\{\Theta_p|z_k|, \theta_p\}, \theta_c\Big\} \end{bmatrix} = T(|z_k|), \end{aligned} \tag{6.55}$$

where the equality above follows from (6.45).

2. We next proceed by induction on k. Note that 2 a) holds for $k = 0$ by assumption. Suppose now that 2 a) holds for some $k \geq 0$. Note that $z_k = U^{-1}x_k$ and hence $|z_k| = |U^{-1}x_k|$. From 2 a), we have $|z_k| \preceq b_r$ and by (6.49), applying T to this inequality yields $T(|z_k|) \preceq T(b_r) = b_{r+1} \preceq b_r$, where we have used part 1. Combining this inequality with (6.55), we obtain

$$|z_{k+1}| \preceq T(|z_k|) \preceq b_r.$$

Since $|z_{k+1}| = |U^{-1}x_{k+1}|$, then 2 a) holds for $k+1$ and we have proved by induction that 2 a) holds for all $k \geq 0$, concluding the proof of part 2 a). To prove 2 b), note that $|x_k| \preceq |U| \cdot |U^{-1}x_k|$ and use 2 a).

3. We first prove that 3 a) and 3 b) hold for $r = \infty$. Define

$$\bar{\beta} \triangleq \max\{\beta, |z_0|\}, \tag{6.56}$$

where $z_0 = U^{-1}x_0$, and note that $\bar{\beta} \succeq \beta$ and $|z_0| \preceq \bar{\beta}$. From (6.55), we have $|z_{k+1}| \preceq T(|z_k|)$, for all $k \geq 0$, and using (6.49) then

$$|z_k| \preceq T^k(|z_0|) \preceq T^k(\bar{\beta}), \text{ for all } k \geq 0. \tag{6.57}$$

Claim: Let $w, y \in \mathbb{R}^n_{+,0}$ and suppose that $w \succeq y \succeq 0$. Then, $0 \preceq T(w) - T(y) \preceq M(w-y)$.

Proof: Since $w \succeq y \succeq 0$, then $0 \preceq T(w) - T(y)$ follows directly from (6.49). From (6.45) we have

$$\begin{aligned}T(w) - T(y) &= |\Lambda|(w-y) + \\ &|U^{-1}B_d| \begin{bmatrix} \max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\} \\ \max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} - \max\Big\{\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\}, \theta_c\Big\} \end{bmatrix}. \end{aligned} \tag{6.58}$$

We next prove that the expression between square brackets above is less than or equal to $\Theta(w-y)$. Let $(\cdot)_i$ denote the i-th component of a vector.

- If $(\Theta_p y)_i \geq (\theta_p)_i$, then $(\Theta_p w)_i \geq (\theta_p)_i$, since $w \succeq y \succeq 0$ and Θ_p has nonnegative entries. Then, $(\max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\})_i = (\Theta_p(w-y))_i$.

- If $(\Theta_p y)_i < (\theta_p)_i$, then
  - if $(\Theta_p w)_i < (\theta_p)_i$, then
$$(\max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\})_i = 0 \leq (\Theta_p(w-y))_i.$$
  - if $(\Theta_p w)_i \geq (\theta_p)_i$, then
$$(\max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\})_i = (\Theta_p w)_i - (\theta_p)_i < (\Theta_p(w-y))_i.$$

Hence, we have proved that

$$\max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\} \preceq \Theta_p(w-y). \tag{6.59}$$

- If $(\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\})_i \geq (\theta_c)_i$, then $(\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\})_i \geq (\theta_c)_i$, since $w \succeq y \succeq 0$ and all the matrices and vectors involved have nonnegative entries. Then,
$$\begin{aligned}
&\Big(\max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} - \max\Big\{\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\}, \theta_c\Big\}\Big)_i \\
&= \Big(\Theta_c(w-y) + \Theta_s\Big[\max\{\Theta_p w, \theta_p\} - \max\{\Theta_p y, \theta_p\}\Big]\Big)_i \\
&\leq ((\Theta_c + \Theta_s\Theta_p)(w-y))_i,
\end{aligned}$$
  where in the last line above we have used (6.59).

- If $(\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\})_i < (\theta_c)_i$, then
  - if $(\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\})_i < (\theta_c)_i$, then
$$\begin{aligned}
&\Big(\max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} - \max\Big\{\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\}, \theta_c\Big\}\Big)_i \\
&= 0 \leq ((\Theta_c + \Theta_s\Theta_p)(w-y))_i.
\end{aligned}$$
  - if $(\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\})_i \geq (\theta_c)_i$, then
$$\begin{aligned}
&\Big(\max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} - \max\Big\{\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\}, \theta_c\Big\}\Big)_i \\
&= (\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\} - \theta_c)_i \\
&< (\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\} - \Theta_c y - \Theta_s \max\{\Theta_p y, \theta_p\})_i \\
&\leq ((\Theta_c + \Theta_s\Theta_p)(w-y))_i,
\end{aligned}$$
    where in the last line above we have used (6.59).

Hence, we have proved that

$$\begin{aligned}
&\max\Big\{\Theta_c w + \Theta_s \max\{\Theta_p w, \theta_p\}, \theta_c\Big\} - \max\Big\{\Theta_c y + \Theta_s \max\{\Theta_p y, \theta_p\}, \theta_c\Big\} \\
&\qquad\qquad \preceq (\Theta_c + \Theta_s\Theta_p)(w-y).
\end{aligned} \tag{6.60}$$

Using (6.40)–(6.41) and (6.58)–(6.60), it follows that $T(w) - T(y) \preceq M(w - y)$, concluding the proof of the claim.

Since $\bar{\beta} \succeq \beta \succeq 0$, by the claim then $0 \preceq T(\bar{\beta}) - T(\beta) \preceq M(\bar{\beta} - \beta)$. Then, $T(\bar{\beta}) \succeq T(\beta) \succeq 0$ and using the claim again yields $T^2(\bar{\beta}) - T^2(\beta) \preceq M(T(\bar{\beta}) - T(\beta)) \preceq M^2(\bar{\beta} - \beta)$. Iterating this procedure yields $T^k(\bar{\beta}) - T^k(\beta) \preceq M^k(\bar{\beta} - \beta)$ and since $\boldsymbol{\rho}(M) < 1$, then $\lim_{k\to\infty} \left(T^k(\bar{\beta}) - T^k(\beta)\right) = 0$. Since by part 1 $\lim_{k\to\infty} T^k(\beta) = b_\infty$, then $\lim_{k\to\infty} T^k(\bar{\beta}) = b_\infty$. Therefore, given $\epsilon \in \mathbb{R}^n_+$, there exists $\ell \geq 0$ such that $T^k(\bar{\beta}) \preceq b_\infty + \epsilon$, for all $k \geq \ell$. Recalling (6.57) yields $|z_k| \preceq b_\infty + \epsilon$, for all $k \geq \ell$. The proof of part 3 a) for $r = \infty$ then follows by recalling that $|z_k| = |U^{-1}x_k|$ and noting that ℓ depends on x_0 since ℓ depends on $\bar{\beta}$ and $\bar{\beta}$ depends on $|z_0| = |U^{-1}x_0|$. To prove 3 b) for $r = \infty$, note that $|x_k| \preceq |U| \cdot |U^{-1}x_k|$ and use 3 a). For $0 \leq r < \infty$, 3 a) and 3 b) follow straightforwardly using the fact that $0 \preceq b_\infty \preceq b_r \preceq b_{r-1}$ for all $0 \leq r < \infty$. □

Theorem 6.5 gives a systematic method to compute componentwise ultimate bounds for a discrete-time system of the form (6.8)–(6.9) where the perturbation Δy_k is bounded as in (6.36). In particular, if the matrices in (6.36) have the form (6.37)–(6.39), then Theorem 6.5 provides componentwise ultimate bounds for the quantised discrete-time system of Figure 6.1. Note, however, that the result of Theorem 6.5 is valid irrespective of the form of the matrices Θ_p, Θ_c and Θ_s, provided they have nonnegative entries. The tightest ultimate bound given by this theorem is obtained using b_∞. In addition, the ultimate bounds corresponding to b_r, for $r < \infty$, may be more conservative but require only r iterations of the map T.

Remark 6.6 *Under the assumption that M in (6.40) has all its eigenvalues in the open unit disc ($\boldsymbol{\rho}(M) < 1$), part 2 of Theorem 6.5 characterises a bounded invariant region in the state space of the discrete-time system (6.8) and part 3 shows that its state trajectories asymptotically converge to this region from any initial condition. Thus, the assumption that $\boldsymbol{\rho}(M) < 1$ is a sufficient condition for global practical stability of the discrete-time system (6.9), that is, for its trajectories to be ultimately bounded from any initial condition. Since $\boldsymbol{\rho}(M) < 1$ implies that $\boldsymbol{\rho}(|\Lambda|) < 1$, and since $\boldsymbol{\rho}(|\Lambda|) = \boldsymbol{\rho}(\Lambda) = \boldsymbol{\rho}(A_d)$, it follows that $\boldsymbol{\rho}(M) < 1$ implies that $\boldsymbol{\rho}(A_d) < 1$. Hence, a necessary condition for application of Theorem 6.5 is that the unperturbed (closed-loop) discrete-time system $x_{k+1} = A_d x_k$ be stable.*

Remark 6.7 *If the perturbation Δy_k arises from the use of uniform quantisers in all signals, then $\Theta = 0$ in (6.41), since $\Gamma_p = 0$ and $\Gamma_c = 0$ in (6.37)–(6.39) from (6.27)–(6.28). Then, the map T in (6.45) reduces to*

$$T(w) = |\Lambda| w + |U^{-1}B_d|[\theta_p^T \; \theta_c^T]^T,$$

and M in (6.40) to $M = |\Lambda|$. Then, $\beta = \gamma$ and there is no need to iterate the map T since $T(\beta) = \beta = \gamma$. In this case, $\boldsymbol{\rho}(A_d) < 1$ is a necessary and sufficient condition for the discrete-time model

(6.9) to have ultimately bounded trajectories. If, on the contrary, Δy_k arises from the use of logarithmic quantisers in all signals, then $\theta = 0$ in (6.42), since $\theta_p = 0$ and $\theta_c = 0$ from (6.29)–(6.30). In this case, provided $\boldsymbol{\rho}(M) < 1$, then the ultimate bound is zero, since from (6.43) $\beta = 0$, implying that system (6.9) is asymptotically stable. If Δy_k arises from the use of any combination of quantisers, the fact that $\gamma \preceq b_r$ for all $0 \leq r \leq \infty$ can be interpreted as saying that the ultimate bound provided by Theorem 6.5 can never be tighter than the one that would be obtained if all signals were uniformly quantised, where each uniform quantiser had a half-step $\alpha_i/2$ equal to $u_i^\circ/(1+\delta_i)$, for $i = 1, \ldots, \mathrm{s}$.

6.4.2 Sampled-data Systems

We have seen that the analysis of the quantised sampled-data system of Figure 6.2 at the sampling instants reduces to the analysis of a discrete-time system as that shown in Figure 6.1. Componentwise ultimate bounds for this discrete-time system are provided by Theorem 6.5. To derive ultimate bounds for the sampled-data system of Figure 6.2 that are valid at all time instants greater than a finite time, we need to combine the bounds derived in Theorem 6.5 with bounds on the variation of the plant states between sampling instants.

Theorem 6.8 *Consider the perturbed sampled-data system of equations (6.11) and (6.12), and its discrete-time description (6.13)–(6.19). Express A_d in Jordan canonical form as $A_d = U\Lambda U^{-1}$ and consider U partitioned as in* (6.35)*. Let the perturbation Δy_k be bounded as in (6.36), where $z_k = U^{-1}x_k$ and $\Theta_p \in \mathbb{R}^{\mathrm{P}\times n}_{+,0}$, $\Theta_c \in \mathbb{R}^{\mathrm{M}\times n}_{+,0}$, $\theta_p \in \mathbb{R}^{\mathrm{P}}_{+,0}$, $\theta_c \in \mathbb{R}^{\mathrm{M}}_{+,0}$ and $\Theta_s \in \mathbb{R}^{\mathrm{M}\times \mathrm{P}}_{+,0}$. Consider the matrix M defined in (6.40)–(6.41) and suppose that $\boldsymbol{\rho}(M) < 1$. Then, given any $\epsilon \in \mathbb{R}^n_+$ and $x_0 \in \mathbb{R}^n$, there exists $\ell = \ell(\epsilon, x_0) \geq 0$ such that for all $t \geq t_\ell = \ell\mathrm{T}$,*

$$|x_p(t)| \preceq \sup_{0\leq\sigma<\mathrm{T}} \left[\bar{x}_p^1(\sigma) + \bar{x}_p^2(\sigma) + \bar{x}_p^3(\sigma)\right], \tag{6.61}$$

where the supremum is taken componentwise,

$$\bar{x}_p^1(\sigma) \triangleq \left|U_p + \Psi(\sigma)\Big[A_pU_p + B_p\left(D_cC_pU_p + C_cU_c\right)\Big]\right| \bar{z}, \tag{6.62}$$

$$\bar{x}_p^2(\sigma) \triangleq |\Psi(\sigma)B_pD_c|\, \bar{y}_p, \tag{6.63}$$

$$\bar{x}_p^3(\sigma) \triangleq |\Psi(\sigma)B_p|\, \bar{y}_c, \tag{6.64}$$

$$\bar{z} \triangleq (b_r + \epsilon), \tag{6.65}$$

$$\bar{y}_p \triangleq \max\{\Theta_p\bar{z}, \theta_p\}, \tag{6.66}$$

$$\bar{y}_c \triangleq \max\{\Theta_c\bar{z} + \Theta_s\bar{y}_p, \theta_c\}, \tag{6.67}$$

and b_r is given by Theorem 6.5, for any $0 \leq r \leq \infty$.

Proof. The evolution of the plant state between sampling instants is given by

$$x_p(t) = e^{A_p(t-t_k)}x_p(t_k) + \int_0^{t-t_k} e^{A_p\tau}B_p d\tau u_p(t_k), \quad \text{for } t_k \leq t < t_{k+1}, \tag{6.68}$$

where

$$u_p(t_k) = C_c x_c(k) + D_c C_p x_p(t_k) + D_c \Delta y_p(k) + \Delta y_c(k). \tag{6.69}$$

Recalling that $x_p(t_k) = U_p z_k$ and $x_c(k) = U_c z_k$, (6.69) can be rewritten as

$$u_p(t_k) = (C_c U_c + D_c C_p U_p) z_k + D_c \Delta y_p(k) + \Delta y_c(k). \tag{6.70}$$

Operating on (6.68), (6.70), and using (6.19) and the identity

$$e^{A_p t} = \mathrm{I} + \Psi(t) A_p,$$

we can obtain

$$\begin{aligned} |x_p(t)| \preceq & \left| U_p + \Psi(t - t_k) \Big[A_p U_p + B_p \left(D_c C_p U_p + C_c U_c \right) \Big] \right| |z_k| \\ & + |\Psi(t - t_k) B_p D_c| \, |\Delta y_p(k)| \\ & + |\Psi(t - t_k) B_p| \, |\Delta y_c(k)|, \end{aligned} \tag{6.71}$$

for $t_k \leq t < t_{k+1}$. From Theorem 6.5 part 3, we know that for any $\epsilon \in \mathbb{R}^n_+$ and $x_0 \in \mathbb{R}^n$, there exists $\ell \geq 0$ such that

$$|z_k| = |U^{-1} x_k| \leq b_r + \epsilon = \bar{z},$$

for all $k \geq \ell$, for any $0 \leq r \leq \infty$. From (6.36) and the bound above, we have, for all $k \geq \ell$,

$$\begin{aligned} |\Delta y_p(k)| &\preceq \max\{\Theta_p \bar{z}, \theta_p\} = \bar{y}_p, \text{ and} \\ |\Delta y_c(k)| &\preceq \max\{\Theta_c \bar{z} + \Theta_s \bar{y}_p, \theta_c\} = \bar{y}_c. \end{aligned}$$

Using the above bounds in (6.71) yields

$$|x_p(t)| \preceq \bar{x}_p^1(t - t_k) + \bar{x}_p^2(t - t_k) + \bar{x}_p^3(t - t_k), \tag{6.72}$$

for all $k \geq \ell$, for all $t_k \leq t < t_{k+1}$. Eq. (6.61) then follows by taking componentwise suprema in (6.72) and noting that $0 \leq t - t_k < \mathrm{T}$ for $t_k \leq t < t_{k+1}$. □

The bound (6.61)–(6.67) given by Theorem 6.8 requires one to compute the componentwise suprema of a function. This calculation has to be performed numerically, since in general the bound (6.61)–(6.67) will not admit an explicit expression.

We next provide some alternative bounds that, though being more conservative than the bound (6.61)–(6.67), allow explicit expressions.

Lemma 6.9 *Consider the perturbed sampled-data system of equations (6.11) and (6.12), and its discrete-time description (6.13)–(6.19). Express A_p and A_d in Jordan canonical form as $A_d = U \Lambda U^{-1}$ and $A_p = \tilde{U} \tilde{\Lambda} \tilde{U}^{-1}$, and consider U partitioned as in (6.35). Let the perturbation Δy_k be bounded as in*

(6.36), where $z_k = U^{-1}x_k$ *and* $\Theta_p \in \mathbb{R}^{\mathrm{P}\times n}_{+,0}$, $\Theta_c \in \mathbb{R}^{\mathrm{M}\times n}_{+,0}$, $\theta_p \in \mathbb{R}^{\mathrm{P}}_{+,0}$, $\theta_c \in \mathbb{R}^{\mathrm{M}}_{+,0}$ *and* $\Theta_s \in \mathbb{R}^{\mathrm{M}\times \mathrm{P}}_{+,0}$. *Consider the matrix* M *defined in (6.40)–(6.41) and suppose that* $\boldsymbol{\rho}(M) < 1$. *Define*

$$G \triangleq \int_0^{\mathrm{T}} e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau} d\tau. \tag{6.73}$$

Then, given any $\epsilon \in \mathbb{R}^n_+$ *and* $x_0 \in \mathbb{R}^n$, *there exists* $\ell = \ell(\epsilon, x_0) \geq 0$ *such that for all* $t \geq t_\ell = \ell\mathrm{T}$,

$$\begin{aligned} |x_p(t)| \preceq \left(|U_p| + |\tilde{U}|G\left|\tilde{U}^{-1}\left[(A_p + B_pD_cC_p)U_p + B_pC_cU_c\right]\right|\right)\bar{z} \\ + |\tilde{U}|G|\tilde{U}^{-1}B_pD_c|\bar{y}_p + |\tilde{U}|G|\tilde{U}^{-1}B_p|\bar{y}_c, \end{aligned} \tag{6.74}$$

with $\bar{z}$, $\bar{y}_p$ *and* $\bar{y}_c$ *as defined in (6.65)–(6.67), and* b_r *given by Theorem 6.5, for any* $0 \leq r \leq \infty$.

Proof. From Theorem 6.8, it follows that

$$|x_p(t)| \preceq \sup_{0\leq\sigma<\mathrm{T}} \left[\sum_{i=1}^{3} \bar{x}_p^i(\sigma)\right] \preceq \sum_{i=1}^{3} \sup_{0\leq\sigma<\mathrm{T}} \bar{x}_p^i(\sigma). \tag{6.75}$$

Consider $\bar{x}_p^3(\sigma)$. Using (6.64) and (6.19), we have

$$\begin{aligned} \bar{x}_p^3(\sigma) &= \left|\int_0^\sigma e^{A_p\tau} d\tau B_p\right| \bar{y}_c \\ &\preceq \int_0^\sigma \left|e^{A_p\tau}B_p\right| d\tau \bar{y}_c \\ &\preceq |\tilde{U}| \int_0^\sigma \left|e^{\tilde{\Lambda}\tau}\right| d\tau |\tilde{U}^{-1}B_p|\bar{y}_c. \end{aligned} \tag{6.76}$$

Substituting $\left|e^{\tilde{\Lambda}\tau}\right| = e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau}$ (which holds since $\tilde{\Lambda}$ is in Jordan form) into (6.76) and finding the supremum yields

$$\begin{aligned} \sup_{0\leq\sigma<\mathrm{T}} \bar{x}_p^3(\sigma) &\preceq \sup_{0\leq\sigma<\mathrm{T}} |\tilde{U}| \int_0^\sigma e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau} d\tau |\tilde{U}^{-1}B_p|\bar{y}_c \\ &\preceq |\tilde{U}| \left[\sup_{0\leq\sigma<\mathrm{T}} \int_0^\sigma e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau} d\tau |\tilde{U}^{-1}B_p|\bar{y}_c\right] \\ &\preceq |\tilde{U}| \left[\sup_{0\leq\sigma<\mathrm{T}} \int_0^\sigma e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau} d\tau\right] |\tilde{U}^{-1}B_p|\bar{y}_c \\ &\preceq |\tilde{U}|G|\tilde{U}^{-1}B_p|\bar{y}_c. \end{aligned}$$

The last line above follows since $e^{\mathbb{R}\mathrm{e}(\tilde{\Lambda})\tau} \succeq 0$ for all $0 \leq \tau \leq \mathrm{T}$. The bounds for $\sup_{0\leq\sigma<\mathrm{T}} \bar{x}_p^i(\sigma)$, for $i = 1, 2$, are proved in a similar manner, and the result follows by substituting these bounds into (6.75).

□

Remark 6.10 *The matrix* G *in (6.73) can be explicitly calculated and even admits a simple expression if* $\mathbb{R}\mathrm{e}(\tilde{\Lambda})$ *is invertible. In addition, the result of Lemma 6.9 is still valid if we replace* G *by a componentwise upper bound* $\tilde{G}$, *such that* $G \preceq \tilde{G}$, *though this again leads to more conservative ultimate bounds.*

6.5 Examples

6.5.1 Static Controller with a Single Quantiser

To illustrate the application of the method developed, we consider the magnetic ball levitation system used as an application example in Ishii et al. (2004) and Ishii and Francis (2002a, §4.7). This system consists of a steel ball that is suspended in the air by means of an electromagnet. The control objective is to keep the position of the ball, y, at an equilibrium by controlling the voltage applied to the electromagnet, v. The current through the electromagnet's coil is denoted by i. The plant state is taken as $x_p = [y \;\; \dot{y} \;\; i]^T$ and the input is $\bar{u}_p = v$. We directly consider the same linearisation of the model as in Ishii and Francis (2002a, §4.7) and Ishii et al. (2004), with the same numerical values for the system parameters. This yields the quantised sampled-data system depicted in Figure 6.5, where the state of the continuous-time plant with matrices

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 2798 & 0 & -19.6 \\ 0 & 0 & -24.39 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 2.439 \end{bmatrix}, \tag{6.77}$$

is regularly sampled every $\mathrm{T} = 4.605 \cdot 10^{-3}$ seconds. [This sampling period was used in Ishii et al. (2004).] The state samples are multiplied by the feedback gain

$$K = \begin{bmatrix} 10315.67 & 195.02 & -49.47 \end{bmatrix}, \tag{6.78}$$

and then passed through a scalar quantiser q to generate the control inputs at times $t_k = k\mathrm{T}$, for all $k \geq 0$. At times $t_k \leq t < t_k + \mathrm{T}$, the plant input is held at its value by means of a zero-order hold device.

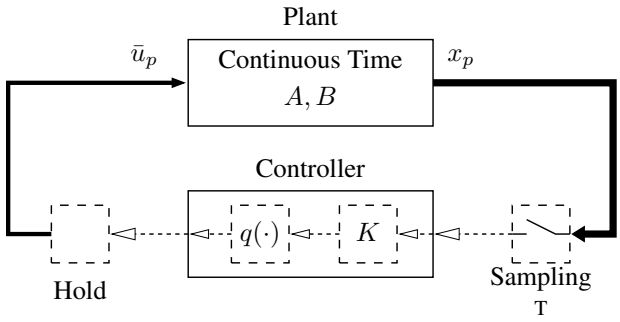


Figure 6.5: Quantised sampled-data system in Ishii et al. (2004).

The quantiser q is defined as:

$$q(\sigma) = u_j \text{ if and only if } \sigma \in \mathcal{I}_j, \tag{6.79}$$

for all $j \in \mathbb{Z}$, where

$$u_j = \begin{cases} 0 & \text{if } j = 0, \\ \operatorname{sgn}(j)\beta_0\rho^{1-|j|} & \text{if } j \neq 0, \end{cases} \tag{6.80}$$

$$\mathcal{I}_j = \begin{cases} (-\alpha_0, \alpha_0) & \text{if } j = 0, \\ [\operatorname{sgn}(j)\alpha_0\rho^{1-|j|}, \operatorname{sgn}(j)\alpha_0\rho^{-|j|}) & \text{if } j \neq 0, \end{cases} \tag{6.81}$$

$$\beta_0 = 0.652, \quad \alpha_0 = 0.451, \quad \rho = 1/1.78 = 0.5618. \tag{6.82}$$

Our aim is to apply the proposed ultimate bound estimation method to this magnetic levitation system. To do this, we need to regard this system as a perturbed sampled-data system described by equations (6.11) and (6.12), where the perturbations are introduced by the quantiser and are bounded as was described in §6.3. The scalar quantiser q in (6.79)–(6.82), depicted in Figure 6.6 a), can be expressed as $q(s) = r\tilde{q}(s)$, where $\tilde{q}$ is a semitruncated logarithmic quantiser of the form considered in §6.3. Straightforward calculations yield $r = \beta_0(1+\rho)/(2\alpha_0) = 1.1289$. The quantiser $\tilde{q}$ is shown in Figure 6.6 b), where $\tilde{\beta}_0 = \beta_0/r = 0.5775$.

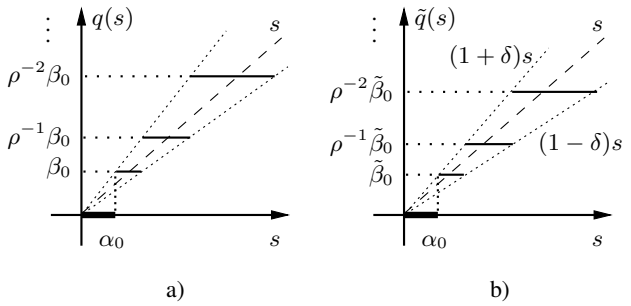


Figure 6.6: Quantiser rescaling.

We can now straightforwardly put the system into the form of equations (6.11) and (6.12), with the perturbation bounded as in (6.36) by interchanging the scalar gain r with the hold device in Figure 6.5. We then obtain a continuous-time plant of equations (6.11a) and (6.11b) with $A_p = A$, $B_p = Br$ and $C_p = \mathrm{I}_3$, and a static discrete-time controller of equation (6.11d) with $D_c = K$ and zero C_c. We also have (6.12) and note that $\mathrm{N}_p = 3$, $\mathrm{N}_c = 0$, $\mathrm{M} = 1$, $\mathrm{P} = 3$, $n = \mathrm{N}_p + \mathrm{N}_c = 3$ and $\mathrm{S} = \mathrm{P} + \mathrm{M} = 4$. The discrete-time model (6.13)–(6.19) is then given by

$$x_k = x_p(t_k), \qquad A_d = A_{11}, \qquad B_d = [B_{11}\ B_{12}]. \tag{6.83}$$

We express A_d in Jordan canonical form as $A_d = U\Lambda U^{-1}$ and note that the partition of U in (6.35) is just $U = U_p \in \mathbb{C}^{3\times 3}$, since $\mathrm{N}_c = 0$. From (6.27) and (6.29), and since there is no quantisation at the plant outputs, we have $\Gamma_p = \mathbf{0}_{3\times 3}$ and $\theta_p = \mathbf{0}_{3\times 1}$. Then, from (6.37), $\Theta_p = \mathbf{0}_{3\times 3}$. From (6.28), (6.23)

and Figure 6.6 b), we have $\Gamma_c = (1-\rho)/(1+\rho)$ and $\theta_c = \alpha_0$. Also, from (6.38), $\Theta_c = \Gamma_c|KU_p|$. Then, Δy_k can be bounded as

$$|\Delta y_p(k)| = 0, \tag{6.84}$$

$$|\Delta y_c(k)| \leq \max\{\Theta_c|z_k|, \theta_c\} = \max\left\{\frac{1-\rho}{1+\rho}|KU_p||z_k|, \alpha_0\right\}. \tag{6.85}$$

We can now readily compute the matrix M in (6.40) and verify that $\boldsymbol{\rho}(M) < 1$. Then, using Theorem 6.5, we have

$$\beta = [0.1121 \;\; 0.0336 \;\; 0.0507]^T \quad \text{and} \quad \gamma = [6.067 \;\; 1.817 \;\; 2.745]^T \cdot 10^{-2}.$$

Iteration of the map T defined in (6.45) from the initial condition β yields

$$b_\infty = \gamma = [6.067 \;\; 1.817 \;\; 2.745]^T \cdot 10^{-2}.$$

Application of Theorem 6.5 using $\epsilon = [1 \;\; 1 \;\; 1]^T \cdot 10^{-10}$ and b_∞ shows that for any $x_0 = x_p(0) \in \mathbb{R}^3$, there exists $\ell = \ell(\epsilon, x_0) \geq 0$ such that

$$|x_p(t_k)| \preceq [9.2 \;\; 363.5 \;\; 862.2]^T \cdot 10^{-4}, \tag{6.86}$$

for all $k \geq \ell$. Moreover, application of Theorem 6.8 using b_∞ and the same value of ϵ yields

$$|x_p(t)| \preceq [9.2 \;\; 363.5 \;\; 862.2]^T \cdot 10^{-4}, \tag{6.87}$$

and application of Lemma 6.9 yields

$$|x_p(t)| \preceq [13.7 \;\; 544.2 \;\; 1096.3]^T \cdot 10^{-4}, \tag{6.88}$$

both for all $t \geq t_\ell = \ell\mathrm{T}$.

We observe some interesting features of this example. First, note that $b_\infty = \gamma$ and hence the lower bound on the discrete-time componentwise ultimate bound provided by Theorem 6.5 is achieved. In §6.5.2, we will see another quantised sampled-data scheme where this feature is not present. Second, note that the bound (6.87), that takes account of intersample behaviour, is identical to the bound (6.86), which is only valid at the sampling instants. The equality of these bounds shows that, in this example, there is no conservativeness in the bounding procedure of Theorem 6.8.

In Ishii et al. (2004), a randomized algorithm is developed that reduces conservatism in the analysis of sampled-data systems with quantisers. The approach in Ishii et al. (2004) can reduce conservatism not only in the ultimate bounds for a system, but also in the required sampling period. In addition, the approach considers a guaranteed decay rate to the ultimate bound. Here, we are only interested in comparing the ultimate bound obtained in Ishii et al. (2004) with the componentwise bounds (6.87) and (6.88). The ultimate bound obtained in Ishii et al. (2004) is, using our notation,

$$\|x_p(t)\|_2 \leq 0.053, \quad \text{for } t \geq t_{\ell'}. \tag{6.89}$$

From (6.87), it follows that

$$\|x_p(t)\|_2 \leq 0.0936, \tag{6.90}$$

and from (6.88), that

$$\|x_p(t)\|_2 \leq 0.1224, \tag{6.91}$$

both for all $t \geq t_\ell$.

It is not surprising, perhaps, that the bound (6.89), obtained in Ishii et al. (2004), is better than (6.90), since the algorithm in Ishii et al. (2004) involves the analysis of individual state trajectories. On the other hand, note that the componentwise bound (6.87) gives a tighter bound on the first two components of the state, which represent the position and velocity of the ball in the magnetic levitation system. In particular, the ultimate bound on the ball position, $9.2 \cdot 10^{-4}$, is more than 50 times lower than 0.053. Also, the ultimate bound on the ball position given by (6.88), $13.7 \cdot 10^{-4}$, is more than 35 times lower than 0.053.

It is worth emphasising that the method we propose is completely systematic, and does not require adjustment of any parameters or selection of, for example, appropriate probability density functions in order to provide an ultimate bound for a system.

6.5.2 Static Controller with Mixed Quantisers

Consider the same example of §6.5.1, but assume now that, in addition to the quantiser of Figure 6.5, the plant states are individually uniformly quantised before multiplication by the feedback gain K. For this new scheme, the results of Ishii et al. (2004) are not directly applicable since we have quantisation both at the plant and controller outputs.

We consider quantisation steps $\alpha_1 = 2 \cdot 10^{-4}$, $\alpha_2 = 0.01$ and $\alpha_3 = 0.02$ for the quantisers corresponding to x_{p_1}, x_{p_2} and x_{p_3}, respectively. For this new quantised sampled-data scheme, we have $\delta_i = 0$ and $u_i^\circ = \alpha_i/2$ for $i = 1, 2, 3$ (see §6.3). Hence, from (6.27) and (6.29) we have $\Gamma_p = 0$ and $\theta_p = [1\ 50\ 100]^T \cdot 10^{-4}$. Note from (6.40) that the matrix M is the same as in §6.5.1, since we consider the same plant and feedback gain, and we did not add or modify any logarithmic or semitruncated logarithmic quantisers. Then, using Theorem 6.5, we have

$$\beta = [0.9084\ \ 0.2721\ \ 0.4111]^T \quad \text{and} \quad \gamma = [0.4915\ \ 0.1473\ \ 0.2224]^T.$$

Iteration of the map T defined in (6.45) from the initial condition β yields

$$b_\infty = [0.7963\ \ 0.2385\ \ 0.3603]^T.$$

Application of Theorem 6.5 using $\epsilon = [1\ \ 1\ \ 1]^T \cdot 10^{-10}$ and b_∞ shows that for any $x_0 = x_p(0) \in \mathbb{R}^3$, there exists $\ell = \ell(\epsilon, x_0) \geq 0$ such that

$$|x_p(t_k)| \preceq [0.0120\ \ 0.4770\ \ 1.1317]^T, \tag{6.92}$$

for all $k \geq \ell$. Moreover, application of Theorem 6.8 using b_∞ and the same value of ϵ yields

$$|x_p(t)| \preceq [0.0120 \;\; 0.4770 \;\; 1.1317]^T, \tag{6.93}$$

while application of Lemma 6.9 yields

$$|x_p(t)| \preceq [0.0180 \;\; 0.7142 \;\; 1.4389]^T,$$

both for all $t \geq t_\ell$. Note that, as with the example in §6.5.1, the bound that is valid only at the sampling instants coincides with the one that takes intersample behaviour into account provided by Theorem 6.8. However, as opposed to that example, in this case $b_\infty \succ \gamma$.

6.5.3 Dynamic Controller with a Single Quantiser

Consider again the example of §6.5.1, but now assume that only x_{p_1}, that is, only the ball position, can be measured. To control the plant, we first designed, ignoring quantisation, a discrete-time partial state observer to estimate the remaining plant states, x_{p_2} and x_{p_3}, at the sampling instants. The measured state and the estimated states are multiplied by the feedback gain K given in (6.78) prior to being quantised by the quantiser q defined in (6.79)–(6.82). For this setting, we could not find any location of the observer poles that rendered $\boldsymbol{\rho}(M) < 1$, with M as in (6.40). Since, in this case, our method cannot be applied, we reduced the sampling period to $T = 3.07 \cdot 10^{-3}$, which is one of the values employed in Ishii and Francis (2002a, §4.7).

After recalculating the discrete-time model of the plant for the new sampling period, placing the discrete-time reduced state observer poles at -0.26 and -0.38, and taking account of the state feedback gain K, the resulting controller has matrices

$$A_c = \begin{bmatrix} -0.7715 & -0.001495 \\ 29.011 & -0.17693 \end{bmatrix}, \;\; B_c = \begin{bmatrix} -1000.4 \\ 27423 \end{bmatrix}, \tag{6.94}$$

$$C_c = \begin{bmatrix} 195.02 & -49.465 \end{bmatrix}, \;\; D_c = 5.691 \cdot 10^5. \tag{6.95}$$

The parameters Γ_p, Γ_c, θ_c and θ_p are the same as in §6.5.1. The matrix M in (6.40) now satisfies $\boldsymbol{\rho}(M) < 1$. Then, using Theorem 6.5, we have

$$\beta = \begin{bmatrix} 2.997 \cdot 10^{-2} \\ 6.393 \cdot 10^{-4} \\ 3.504 \\ 20.43 \\ 28.75 \end{bmatrix}, \qquad \gamma = \begin{bmatrix} 6.470 \cdot 10^{-3} \\ 1.381 \cdot 10^{-4} \\ 0.7564 \\ 4.409 \\ 6.206 \end{bmatrix} = b_\infty.$$

Application of Theorem 6.5 using $\epsilon = [1 \;\; 1 \;\; 1 \;\; 1 \;\; 1]^T \cdot 10^{-10}$ and b_∞ shows that for any $x_0 =$

$[x_p(0)^T \;\; x_c(0)^T]^T \in \mathbb{R}^5$, there exists $\ell = \ell(\epsilon, x_0) \geq 0$ such that

$$|x_p(t_k)| \preceq \begin{bmatrix} 1.25 \\ 51.52 \\ 115.5 \end{bmatrix} \cdot 10^{-3}, \qquad |x_c(k)| \preceq \begin{bmatrix} 0.7673 \\ 11.35 \end{bmatrix},$$

for all $k \geq \ell$. Moreover, application of Theorem 6.8 using b_∞ and the same value of ϵ yields

$$|x_p(t)| \preceq [1.25 \;\; 51.52 \;\; 115.5]^T \cdot 10^{-3}, \tag{6.96}$$

while application of Lemma 6.9 yields

$$|x_p(t)| \preceq [1.695 \;\; 69.38 \;\; 137.71]^T \cdot 10^{-3}, \tag{6.97}$$

both for all $t \geq t_\ell$.

As with the example in §6.5.1, note that $b_\infty = \gamma$ and the bound on the plant states that is valid only at the sampling instants coincides with the bound that takes account of intersample behaviour given by Theorem 6.8.

6.6 Chapter Summary

In this chapter, we have developed a novel systematic method to obtain componentwise ultimate bounds for perturbed discrete-time and sampled-data systems, especially when the perturbations arise due to the use of quantisers. The main features of the method are its systematic nature, whereby the method can be readily computer coded without requiring adjustment of parameters, and its flexibility in dealing with highly structured perturbation schemes, whereby the information on the perturbation structure is directly incorporated into the method. This last feature distinguishes the method from other methods that require a bound on the norm of the perturbation and thus may disregard information on the perturbation structure. We have illustrated the simplicity and potential of the method with numerical examples. Future research directions include the derivation of conditions under which the ultimate bounds obtained admit even simpler forms, as observed in the examples when $b_\infty = \gamma$ or when the bound that is valid only at the sampling instants coincides with the one that takes intersample behaviour into account, and the extension of the framework to nonlinear systems.

Chapter 7

General Perturbation Bounds

7.1 Overview

In Chapter 6, we have derived componentwise global ultimate bounds for discrete-time and sampled-data systems involving quantisers. Our derivations were based on regarding a quantised variable as a perturbed copy of the corresponding unquantised variable. The perturbation introduced by a quantiser was then bounded according to the type of quantiser. We have seen that the corresponding perturbation bound may depend on the system state. In summary, Chapter 6 derived componentwise global ultimate bounds for perturbed discrete-time and sampled-data systems having componentwise perturbation bounds of a specific form.

In this chapter, we extend the approach of Chapter 6 to more general componentwise perturbation bounds. We seek componentwise ultimate bounds for continuous- and discrete-time perturbed systems having componentwise perturbation bounds that may depend on the system state. The perturbation bounds considered in this chapter do not necessarily arise from quantisation, though the latter can be seen as a special case of perturbation bounds that fit into the development of this chapter.

As in Chapter 6, our derivations involve the analysis of the system in modal coordinates. This approach has been employed to derive componentwise ultimate bounds for systems having *constant* perturbation bounds in Yakowitz and Parker (1973) for discrete-time systems and in Kofman (2005) for continuous-time systems. The current work extends the approach of Yakowitz and Parker (1973) and Kofman (2005) by allowing the perturbation bound to be a state-dependent function. Specifically, we consider a system defined by

$$\dot{x}(t) = Ax(t) + u(t), \tag{7.1}$$

where $x(t) \in \mathbb{R}^n$ denotes the system state, $u(t) \in \mathbb{R}^n$ a perturbation input and $A \in \mathbb{R}^{n\times n}$ is Hurwitz (has all its eigenvalues in the open left half-plane). The result of Kofman (2005), which applies only when A is also diagonalisable, essentially consists in obtaining a componentwise ultimate bound on the

state x, when the perturbation term $u(t)$ is componentwise bounded as

$$|u_i(t)| \leq u_{m_i}, \quad \text{for } i = 1, \ldots, n.$$

Here, we derive ultimate bounds when A is Hurwitz (not necessarily diagonalisable) and the perturbation term is componentwise bounded by a (possibly) nonlinear function of the state, as follows

$$|u_i(t)| \leq \delta_i(x(t)), \quad \text{for } i = 1, \ldots, n. \tag{7.2}$$

We also derive a discrete-time counterpart of the method, considering a discrete-time system of the form

$$x(k+1) = Ax(k) + u(k),$$

where $A \in \mathbb{R}^{n \times n}$ has all its eigenvalues inside the unit circle and $u(k)$ is componentwise bounded by a state-dependent function.

These results are then utilised to derive ultimate bounds in perturbed nonlinear systems by regarding such a system as a linear system having a (nonlinear) state-dependent perturbation term. In all cases, we provide a systematic method for computing an ultimate bound and a set of initial states from which the ultimate bound obtained is guaranteed. Thus, extending the results of Chapter 6 to more general perturbation bounds comes at the expense of the global feature of the ultimate bounds previously derived. The method is based on the iteration of a map constructed from the modal decomposition of the matrix A and from the perturbation bounds (7.2).

The selected structure, (7.1)–(7.2), permits one to represent most problems where estimation of an ultimate bound is of practical importance. These problems include the presence of noise, the effect of uniform or logarithmic quantisation, systems with parametric uncertainty (where the product of an unknown matrix and the state can be modelled as a perturbation), etc. Notice that in these cases the perturbation does not necessarily affect each component of the right-hand side of (7.1) in the same way and hence it may be useful to bound the perturbation componentwise as in (7.2). Our method can also be easily extended to systems of the form $\dot{x}(t) = Ax(t) + Bu(t)$ or $x(k+1) = Ax(k) + Bu(k)$, where $u \in \mathbb{R}^m$, with straightforward modifications of the derived expressions.

The remainder of the chapter is organised as follows. In §7.2 and §7.3, we derive componentwise ultimate bounds for continuous- and discrete-time systems, respectively. Illustrative examples are provided in §7.4 and a summary is given in §7.5.

7.2 Ultimate Bounds for Continuous-time Systems

In this section, we develop a systematic method to obtain ultimate bounds for perturbed continuous-time systems. In §7.2.1, we derive ultimate bound expressions when the perturbation input is componentwise bounded by constants. This result is used as an intermediate tool to derive ultimate bound expressions

at the beginning of §7.2.2, where the perturbation input is bounded by a state-dependent function. In §7.2.2, we then proceed to develop the aforementioned systematic method. In §7.2.3, we show how the results of §7.2.2 may be applied to nonlinear systems.

7.2.1 Constant Perturbation Bounds

In this section, we present ultimate bounds for a linear system when the perturbation bound is constant. This result is presented in Theorem 7.3, which builds on the following two preliminary lemmas. Lemma 7.1 derives a result for a perturbed scalar system and Lemma 7.2 a similar result for a system whose evolution matrix consists of a single Jordan block.

Lemma 7.1 *Consider the complex scalar system*

$$\dot{z}(t) = \lambda z(t) + v(t) \tag{7.3}$$

where $\lambda, z(t), v(t) \in \mathbb{C}$*, and* $\mathbb{R}\mathrm{e}(\lambda) < 0$*. Let* $v_m \in \mathbb{R}_{+,0}$ *and suppose that* $|z(0)| \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_m$*.*

a) If $|v(t)| \leq v_m$ *for all* $0 \leq t \leq \tau$*, then* $|z(t)| \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_m$ *for all* $0 \leq t \leq \tau$*.*

b) If $|v(t)| \leq v_m$ *for all* $t \geq 0$*, then* $|z(t)| \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_m$ *for all* $t \geq 0$*.*

Proof. Express $z(t)$ in polar form as $z(t) = \rho(t)\, e^{j\theta(t)}$, where $\rho(t) \in \mathbb{R}_{+,0}$ and $\theta(t) \in \mathbb{R}$. Substituting into (7.3) and multiplying by $e^{-j\theta(t)}$ yields

$$\dot{\rho}(t) + j\rho(t)\,\dot{\theta}(t) = \lambda\rho(t) + v(t)\, e^{-j\theta(t)}.$$

Taking real part and using the bound on $v(t)$, we have

$$\dot{\rho}(t) = \mathbb{R}\mathrm{e}(\lambda)\rho(t) + \mathbb{R}\mathrm{e}\left(v(t)\, e^{-j\theta(t)}\right) \leq \mathbb{R}\mathrm{e}(\lambda)\rho(t) + v_m, \tag{7.4}$$

where the inequality is valid for $0 \leq t \leq \tau$ in a) or for $t \geq 0$ in b). Define the auxiliary system

$$\dot{y}(t) = \mathbb{R}\mathrm{e}(\lambda)y(t) + v_m, \tag{7.5}$$

with initial condition $y(0) \triangleq \rho(0) = |z(0)|$. This linear differential equation can be solved as

$$y(t) = |z(0)|e^{\mathbb{R}\mathrm{e}(\lambda)t} + \frac{v_m}{|\mathbb{R}\mathrm{e}(\lambda)|}\left(1 - e^{\mathbb{R}\mathrm{e}(\lambda)t}\right), \tag{7.6}$$

where we have used the fact that $\mathbb{R}\mathrm{e}(\lambda) < 0$. Using the assumption $|z(0)| \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_m$ in (7.6), it follows that $y(t) \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_m$ for all $t \geq 0$. Applying the Comparison Lemma to (7.4) and (7.5) (see, for example, Khalil, 2002, p.102), we conclude that $|z(t)| \leq y(t)$, for all $0 \leq t \leq \tau$ in a) or for all $t \geq 0$ in b). The result follows. □

Lemma 7.2 *Consider the (possibly complex) system*

$$\dot{z}(t) = \Lambda z(t) + v(t) \tag{7.7}$$

where $z(t), v(t) \in \mathbb{C}^r$ and $\Lambda \in \mathbb{C}^{r\times r}$ is a Jordan block with eigenvalue λ satisfying $\mathbb{R}\mathrm{e}(\lambda) < 0$. Let $v_m \in \mathbb{R}^r_{+,0}$ and suppose that $|z(0)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$.

a) If $|v(t)| \preceq v_m$ for all $0 \le t \le \tau$, then $|z(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$ for all $0 \le t \le \tau$.

b) If $|v(t)| \preceq v_m$ for all $t \ge 0$, then $|z(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$ for all $t \ge 0$.

Proof. Since Λ is a Jordan block with eigenvalue λ, note that the matrix $|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}|$ satisfies

$$|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}| = \begin{bmatrix} |\mathbb{R}\mathrm{e}(\lambda)^{-1}| & |\mathbb{R}\mathrm{e}(\lambda)^{-2}| & \dots & |\mathbb{R}\mathrm{e}(\lambda)^{-r}| \\ 0 & |\mathbb{R}\mathrm{e}(\lambda)^{-1}| & \dots & |\mathbb{R}\mathrm{e}(\lambda)^{-(r-1)}| \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & |\mathbb{R}\mathrm{e}(\lambda)^{-1}| \end{bmatrix}. \tag{7.8}$$

Define

$$a \triangleq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m, \tag{7.9}$$

and let a_i and v_{m_j} denote the i-th and j-th components of a and v_m, respectively. Using (7.8) and (7.9), we can write

$$a_i = \sum_{j=i}^{r} \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-(j-i+1)}\right| v_{m_j}, \tag{7.10}$$

for $i = 1, \dots, r$. Let $z_i(t)$ denote the i-th component of $z(t)$. We will prove by induction that

$$|z_i(t)| \le a_i, \quad \text{for } i = 1, \dots, r, \tag{7.11}$$

for $0 \le t \le \tau$ in a) or for $t \ge 0$ in b). By assumption, $|z(0)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$, and using a as defined above, we have

$$|z_i(0)| \le a_i, \quad \text{for } i = 1, \dots, r. \tag{7.12}$$

In particular, $|z_r(0)| \le a_r = \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right| v_{m_r}$. From (7.7) and the Jordan form of Λ, it follows that $\dot{z}_r(t) = \lambda z_r(t) + v_r(t)$, with $|v_r(t)| \le v_{m_r}$ for $0 \le t \le \tau$ in a) or for $t \ge 0$ in b). Applying Lemma 7.1 yields $|z_r(t)| \le |\mathbb{R}\mathrm{e}(\lambda)^{-1}| v_{m_r} = a_r$ for $0 \le t \le \tau$ in a) or for $t \ge 0$ in b), proving (7.11) for $i = r$.

We next prove that if z_{i+1} satisfies (7.11), then z_i also does. Thus, suppose that z_{i+1} satisfies (7.11). This implies

$$|z_{i+1}(t) + v_i(t)| \le a_{i+1} + v_{m_i} \tag{7.13}$$

for $0 \leq t \leq \tau$ in a) or for $t \geq 0$ in b). Using (7.10), the right-hand side of (7.13) satisfies

$$\begin{aligned}\left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right|(a_{i+1}+v_{m_i}) &= \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right|\left(\sum_{j=i+1}^{r}\left|[\mathbb{R}\mathrm{e}(\lambda)]^{-(j-i)}\right|v_{m_j}+v_{m_i}\right)\\ &= \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right|\sum_{j=i}^{r}\left|[\mathbb{R}\mathrm{e}(\lambda)]^{-(j-i)}\right|v_{m_j}\\ &= \sum_{j=i}^{r}\left|[\mathbb{R}\mathrm{e}(\lambda)]^{-(j-i+1)}\right|v_{m_j} = a_i. \end{aligned} \tag{7.14}$$

From (7.7) and the Jordan form of Λ, we have $\dot{z}_i(t) = \lambda z_i(t) + z_{i+1}(t) + v_i(t)$, for $i = 1, \ldots, r-1$, where the last two terms satisfy (7.13). From (7.12) and (7.14), we have

$$|z_i(0)| \leq a_i = \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right|(a_{i+1}+v_{m_i}).$$

Applying Lemma 7.1 then yields

$$|z_i(t)| \leq \left|[\mathbb{R}\mathrm{e}(\lambda)]^{-1}\right|(a_{i+1}+v_{m_i}) = a_i$$

valid for $0 \leq t \leq \tau$ in a) or for $t \geq 0$ in b). This shows that z_i also satisfies (7.11).

Since we have already shown that z_r satisfies (7.11), it follows that (7.11) is satisfied for $i = 1, \ldots, r$ and the proof is complete. □

The following theorem provides ultimate bounds for linear systems having constant perturbation bounds. This theorem extends the result of Kofman (2005) to the case where the system's evolution matrix is required to be only Hurwitz (not necessarily diagonalisable). The main feature of this result is that it does not require the calculation of a Lyapunov function for the system and may yield tighter bounds than those obtained via standard Lyapunov analysis using quadratic functions, as we will show in §7.4.1 by means of a numerical example.

Theorem 7.3 *Consider the system*

$$\dot{x}(t) = Ax(t) + u(t) \tag{7.15}$$

where $x(t), u(t) \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$ is a Hurwitz matrix with Jordan canonical form $\Lambda = U^{-1}AU$. Let $u_m \in \mathbb{R}^n_{+,0}$ and define

$$S \triangleq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| \cdot |U^{-1}|. \tag{7.16}$$

i) Invariance. Suppose that $|U^{-1}x(0)| \preceq Su_m$. If $|u(t)| \preceq u_m$ for $0 \leq t \leq \tau$, then the following hold for $0 \leq t \leq \tau$ and if $|u(t)| \preceq u_m$ for all $t \geq 0$, then the following hold for all $t \geq 0$:

a) $|U^{-1}x(t)| \preceq Su_m$.

b) $|x(t)| \preceq |U|Su_m$.

ii) Convergence. Suppose that $|u(t)| \preceq u_m$ *for all* $t \geq 0$*. Then, for each positive vector* $\epsilon \in \mathbb{R}^n_+$*, there exists a continuous function* $t_f(\epsilon, \cdot) : \mathbb{R}^n \to \mathbb{R}_{+,0}$ *so that for each initial condition* $x(0) \in \mathbb{R}^n$*, the following hold for all* $t \geq t_f(\epsilon, x(0))$*:*

a) $|U^{-1}x(t)| \preceq Su_m + \epsilon$*.*

b) $|x(t)| \preceq |U|Su_m + |U|\,\epsilon$*.*

Proof. Let $x(t) = Uz(t)$ and $v(t) \triangleq U^{-1}u(t)$. Then, using (7.15) we have

$$\dot{z}(t) = \Lambda z(t) + v(t), \tag{7.17}$$

where $v(t)$ satisfies

$$|v(t)| \preceq v_m \triangleq |U^{-1}|\,u_m, \quad \text{either for } 0 \leq t \leq \tau \text{ or for all } t \geq 0. \tag{7.18}$$

Note that (7.17) constitutes k ($k \leq n$) uncoupled sets of differential equations of the form

$$\dot{z}_i(t) = \Lambda_i z_i(t) + v_i(t), \quad \text{for } 1 \leq i \leq k, \tag{7.19}$$

where $z_i, v_i \in \mathbb{C}^{r_i}$, $\Lambda_i \in \mathbb{C}^{r_i \times r_i}$ is a Jordan block, and r_i is the multiplicity of the eigenvalue of the i–th block. From (7.18), $|v_i(t)| \preceq v_{m_i}$ either for $0 \leq t \leq \tau$ or for all $t \geq 0$.

i) By assumption, $|z(0)| = |U^{-1}x(0)| \preceq Su_m$. Using (7.16) and (7.18), then

$$|z(0)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$$

and hence $|z_i(0)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda_i)]^{-1}\right| v_{m_i}$, for $i = 1, \ldots, k$. Applying Lemma 7.2 to (7.19), we obtain

$$|z_i(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda_i)]^{-1}\right| v_{m_i}, \quad \text{for } i = 1, \ldots, k, \tag{7.20}$$

for all $0 \leq t \leq \tau$. A compact expression for (7.20) is

$$|z(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m, \text{ for all } 0 \leq t \leq \tau, \tag{7.21}$$

and the proof of i) a) for $0 \leq t \leq \tau$ follows by recalling that $z(t) = U^{-1}x(t)$, $v_m = |U^{-1}|\,u_m$ and (7.16). The proof for all $t \geq 0$ follows identical steps. To prove i) b) note that $|x(t)| = |Uz(t)| \preceq |U| \cdot |z(t)|$ and use (7.21). This completes the proof of i).

ii) Consider again system (7.17) with initial condition $z(0)$ and with the perturbation term bounded by (7.18) for all $t \geq 0$. Let $\tilde{z}(t)$ satisfy

$$\dot{\tilde{z}}(t) = \Lambda\tilde{z}(t), \quad \text{with } \tilde{z}(0) = z(0). \tag{7.22}$$

Since Λ is the Jordan form of A, which is Hurwitz, then the equilibrium point $\tilde{z} = 0$ of (7.22) is globally exponentially stable. Hence, there exist positive constants k and λ such that (see, for example, Khalil, 2002, §4)

$$\|\tilde{z}(t)\|_\infty \leq k\,\|\tilde{z}(0)\|_\infty\, e^{-\lambda t}, \quad \text{for all } t \geq 0,$$

for all $\tilde{z}(0)$. It then follows that for any $\xi \in \mathbb{R}_+$, we have

$$\|\tilde{z}(t)\|_\infty \leq \xi, \quad \text{for all } t \geq \max\left\{0, \frac{1}{\lambda} \ln \frac{k \|\tilde{z}(0)\|_\infty}{\xi}\right\}. \tag{7.23}$$

Therefore, given $\epsilon \in \mathbb{R}^n_+$ and selecting $\xi = \min_{i=1,\ldots,n} \epsilon_i$, it follows from (7.23) that

$$|\tilde{z}(t)| \preceq \epsilon, \text{for all } t \geq t_f^z(\epsilon, \tilde{z}(0)), \tag{7.24}$$

where we have defined

$$t_f^z(\epsilon, \tilde{z}) \triangleq \max\left\{0, \frac{1}{\lambda} \ln \frac{k \|\tilde{z}\|_\infty}{\min_{i=1,\ldots,n} \epsilon_i}\right\}. \tag{7.25}$$

Define $\hat{z}(t) \triangleq z(t) - \tilde{z}(t)$. Then, $\hat{z}(t)$ verifies (7.17) and (7.18). Note also that $|\hat{z}(0)| = 0 \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$. Thus, applying the result of part i), we conclude that $|\hat{z}(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m$ for all $t \geq 0$. Then, using the definition of $\hat{z}$ and (7.24), we obtain

$$|z(t)| \preceq |\hat{z}(t)| + |\tilde{z}(t)| \preceq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| v_m + \epsilon \quad \text{for all } t \geq t_f^z(\epsilon, \tilde{z}(0)). \tag{7.26}$$

Recalling that $z(t) = U^{-1}x(t)$, $v_m = |U^{-1}|\, u_m$, and (7.16), it follows from (7.26) that

$$|U^{-1}x(t)| \preceq S u_m + \epsilon \quad \text{for all } t \geq t_f^z(\epsilon, \tilde{z}(0)). \tag{7.27}$$

Recalling that $\tilde{z}(0) = z(0) = U^{-1}x(0)$, and defining $t_f(\epsilon, x) \triangleq t_f^z(\epsilon, U^{-1}x)$, it follows from (7.27) that

$$|U^{-1}x(t)| \preceq S u_m + \epsilon \quad \text{for all } t \geq t_f(\epsilon, x(0)). \tag{7.28}$$

This establishes ii) a). Part ii) b) follows from (7.28) and since $|x(t)| \preceq |U||U^{-1}x(t)|$. From the definition of t_f and (7.25), note that $t_f(\epsilon, \cdot)$ is continuous. This completes the proof of the theorem. □

Theorem 7.3 i) characterises a bounded invariant region in the state space, that is, a region with the property that trajectories originating in that region remain in the region while the perturbation remains bounded. Theorem 7.3 ii) shows that, if the perturbation is bounded for all $t \geq 0$, then the trajectories converge to the bounded invariant region from any initial condition.

Theorem 7.3 gives both implicit and componentwise ultimate bound estimations of LTI systems when the perturbation bound is constant. The regions of the state space defined by the implicit bounds given in Theorem 7.3 ii) a) are contained in the axis-aligned sets corresponding to Theorem 7.3 ii) b). The latter provides componentwise ultimate bounds on the state.

7.2.2 State-dependent Perturbation Bounds

In this section, we present the main contribution of the chapter for continuous-time systems. We provide ultimate bound expressions for linear systems with state-dependent perturbation bounds that satisfy a monotonicity condition [see (7.30) and (7.31) below]. The ultimate bounds are derived in Theorem 7.4,

which requires the existence of a point (x_m) satisfying a certain condition. We subsequently provide an algorithm to test whether this condition is satisfied, and a proof of the algorithm's convergence. All these results provide a systematic method to obtain ultimate bounds for continuous-time systems. As we will see in the examples, the bounds provided by this systematic method may be tighter than those obtained via standard Lyapunov analysis using quadratic functions, and can also be combined with the latter methodology to improve on the bounds provided by either approach.

Theorem 7.4 *Consider the system*

$$\dot{x}(t) = Ax(t) + u(t), \tag{7.29}$$

where $x(t), u(t) \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$ is Hurwitz with Jordan canonical form $\Lambda = U^{-1}AU$. Suppose that

$$|u(t)| \preceq \delta(x(t)) \quad \textit{for all } t \geq 0, \tag{7.30}$$

where $\delta : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ is a continuous map satisfying

$$|x| \preceq |y| \Rightarrow \delta(x) \preceq \delta(y) \quad \textit{for all } x, y \in \mathbb{R}^n. \tag{7.31}$$

Consider the map $T : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ defined by

$$T(x) \triangleq |U|S\delta(x), \tag{7.32}$$

with S as defined in (7.16). Suppose that there exists $x_m \in \mathbb{R}^n$ satisfying $T(x_m) \prec x_m$. Then,

i) $b \triangleq \lim_{k\to\infty} T^k(x_m)$ exists and satisfies $0 \preceq b \prec x_m$.

ii) If $|U^{-1}x(0)| \preceq S\delta(x_m)$ then, given any positive vector $\epsilon \in \mathbb{R}^n_+$, a finite time $t_f = t_f(\epsilon, x_m) \geq 0$ exists so that for all $t \geq t_f$,

a) $|U^{-1}x(t)| \preceq S\delta(b) + \epsilon$.

b) $|x(t)| \preceq b + |U|\,\epsilon$.

Proof. By (7.16), (7.31), (7.32), and a property of matrices with nonnegative entries [see (6.2) in Chapter 6], then T satisfies

$$|x| \preceq |y| \Rightarrow T(x) \preceq T(y) \quad \text{for all } x, y \in \mathbb{R}^n. \tag{7.33}$$

i) Note that $0 \preceq T(x_m) \prec x_m$ and hence $|T(x_m)| \prec |x_m|$. By (7.33), then $T(T(x_m)) \preceq T(x_m)$ and applying T repeatedly we obtain

$$0 \preceq T^k(x_m) \preceq T^{k-1}(x_m) \prec x_m \quad \text{for all } k \geq 2. \tag{7.34}$$

The sequence $T^k(x_m)$ is thus componentwise nonincreasing and lower bounded by 0, and hence it must converge to some point $b = \lim_{k\to\infty} T^k(x_m)$ that satisfies $0 \preceq b \prec x_m$.

ii) For any $\gamma \in \mathbb{R}^n_{+,0}$, define the map $T_\gamma : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ by

$$T_\gamma(x) \triangleq T(x) + |U|\,\gamma. \tag{7.35}$$

To proceed with the proof of Theorem 7.4, we require the following four claims. The proofs of these claims will be given later.

Claim 1 *For any $\gamma \in \mathbb{R}^n_{+,0}$ and $k \in \mathbb{Z}_{+,0}$, there exists $\bar{t}_f = \bar{t}_f(k, \gamma, x_m) \geq 0$ such that*

$$|U^{-1}x(t)| \preceq S\delta(T^k_\gamma(x_m)) + \gamma \quad \textit{for all } t \geq \bar{t}_f(k, \gamma, x_m). \tag{7.36}$$

Claim 2 *For any $\xi \in \mathbb{R}_+$, there exists $\eta_1 = \eta_1(\xi) > 0$ such that*

$$\|\delta(b + \Delta b) - \delta(b)\|_\infty < \xi \quad \textit{whenever } \|\Delta b\|_\infty < \eta_1(\xi).$$

Claim 3 *For any $\xi \in \mathbb{R}_+$, there exists $N = N(\xi) \in \mathbb{Z}_{+,0}$ such that*

$$\left\|T^k(x_m) - b\right\|_\infty < \xi \quad \textit{whenever } k \geq N(\xi).$$

Claim 4 *For any $\xi \in \mathbb{R}_+$ and $k \in \mathbb{Z}_{+,0}$, there exists $\eta_2 = \eta_2(\xi, k) > 0$ such that*

$$\left\|T^k_\gamma(x_m) - T^k(x_m)\right\|_\infty < \xi \quad \textit{whenever } \gamma \in \mathbb{R}^n_{+,0} \textit{ and } \|\gamma\|_\infty < \eta_2(\xi, k). \tag{7.37}$$

We next show that for any $\epsilon \in \mathbb{R}^n_+$, we may select $\gamma = \gamma(\epsilon, x_m) \in \mathbb{R}^n_+$ and $k = k(\epsilon, x_m) \in \mathbb{Z}_{+,0}$ so that

$$S\delta(T^k_\gamma(x_m)) + \gamma \preceq S\delta(b) + \epsilon. \tag{7.38}$$

For the given x_m, define

$$\Delta b(k, \gamma) \triangleq T^k_\gamma(x_m) - b = T^k_\gamma(x_m) - T^k(x_m) + T^k(x_m) - b \tag{7.39}$$

and write

$$\begin{aligned} S\delta(T^k_\gamma(x_m)) + \gamma &= S\delta(b) + S\left[\delta(b + \Delta b(k, \gamma)) - \delta(b)\right] + \gamma \\ &\preceq S\delta(b) + \left|S\left[\delta(b + \Delta b(k, \gamma)) - \delta(b)\right] + \gamma\right|. \end{aligned} \tag{7.40}$$

For any $\epsilon \in \mathbb{R}^n_+$, define $\xi \triangleq \min_{i=1,\ldots,n} \epsilon_i$ and select $k = k(\epsilon, x_m) \in \mathbb{Z}_{+,0}$ and $\gamma = \gamma(\epsilon, x_m) \in \mathbb{R}^n_+$ to satisfy

$$k = N\left(\frac{1}{2}\eta_1\left(\frac{\xi}{2\|S\|_\infty}\right)\right) \quad \text{and} \tag{7.41}$$

$$\|\gamma\|_\infty < \min\left\{\frac{\xi}{2}, \eta_2\left(\frac{1}{2}\eta_1\left(\frac{\xi}{2\|S\|_\infty}\right), k\right)\right\}, \tag{7.42}$$

where $\eta_1(\cdot)$, $N(\cdot)$, and $\eta_2(\cdot, \cdot)$ are the functions given by Claims 2 to 4. We next show that these selections yield (7.38).

For $k \in \mathbb{Z}_{+,0}$, selected according to (7.41), note that, by (7.42) and Claim 4, we have

$$\left\|T_\gamma^k(x_m) - T^k(x_m)\right\|_\infty < \frac{1}{2}\eta_1\left(\frac{\xi}{2\left\|S\right\|_\infty}\right). \tag{7.43}$$

By (7.41) and Claim 3, we have

$$\left\|T^k(x_m) - b\right\|_\infty < \frac{1}{2}\eta_1\left(\frac{\xi}{2\left\|S\right\|_\infty}\right). \tag{7.44}$$

From (7.39), it follows that

$$\left\|\Delta b(k,\gamma)\right\|_\infty \leq \left\|T_\gamma^k(x_m) - T^k(x_m)\right\|_\infty + \left\|T^k(x_m) - b\right\|_\infty. \tag{7.45}$$

Combining (7.43), (7.44), and (7.45), yields

$$\left\|\Delta b(k,\gamma)\right\|_\infty < \eta_1\left(\frac{\xi}{2\left\|S\right\|_\infty}\right). \tag{7.46}$$

By (7.46) and Claim 2, it follows that

$$\left\|\delta(b + \Delta b(k,\gamma)) - \delta(b)\right\|_\infty < \frac{\xi}{2\left\|S\right\|_\infty}. \tag{7.47}$$

By (7.47) and since $\left\|\gamma\right\|_\infty < \xi/2$ by (7.42), it follows that

$$\left\|S\right\|_\infty \left\|\delta(b + \Delta b(k,\gamma)) - \delta(b)\right\|_\infty + \left\|\gamma\right\|_\infty < \xi, \tag{7.48}$$

which implies that

$$\left\|S\left[\delta(b + \Delta b(k,\gamma)) - \delta(b)\right] + \gamma\right\|_\infty < \xi. \tag{7.49}$$

From (7.49) and since $\xi = \min_{i=1,\ldots,n} \epsilon_i$, then

$$\left|S\left[\delta(b + \Delta b(k,\gamma)) - \delta(b)\right] + \gamma\right| \prec \epsilon. \tag{7.50}$$

Therefore, using Claim 1 and (7.40) it follows that for any $\epsilon \in \mathbb{R}_+^n$ there exists $t_f = t_f(\epsilon, x_m) \triangleq \bar{t}_f(k(\epsilon, x_m), \gamma(\epsilon, x_m), x_m) \geq 0$ such that

$$|U^{-1}x(t)| \preceq S\delta(b) + \epsilon \quad \text{for all } t \geq t_f. \tag{7.51}$$

This establishes ii a). Part ii) b) follows from (7.51) and since $|x(t)| \preceq |U||U^{-1}x(t)|$. To complete the proof of Theorem 7.4, we finally prove Claims 1 to 4.

Proof of Claim 1. We begin by establishing that $|x(t)| \preceq x_m$ for all $t \geq 0$. For a contradiction, suppose that $|x(t_d)| \not\preceq x_m$, where $0 \leq t_d < \infty$. Define

$$t_c \triangleq \inf t, \quad \text{subject to } t \geq 0 \text{ and } |x(t)| \not\preceq x_m. \tag{7.52}$$

Note that $|x(0)| \preceq |U| \cdot |U^{-1}x(0)|$ and by assumption and (7.32), then

$$|x(0)| \preceq |U|S\delta(x_m) = T(x_m) \prec x_m.$$

Hence, $0 < t_c \leq t_d$ and since $x(t)$ is continuous we have $|x(t)| \preceq x_m$, for all $0 \leq t \leq t_c$. By (7.30) and (7.31) then $|u(t)| \preceq \delta(x_m)$, for all $0 \leq t \leq t_c$. Applying Theorem 7.3 i) b) and using (7.32), then $|x(t)| \preceq T(x_m) \prec x_m$, for all $0 \leq t \leq t_c$. Since $x(t)$ is continuous, then there exists a positive real constant $a > 0$ such that $|x(t)| \preceq x_m$ for all $0 \leq t \leq t_c + a$, contradicting (7.52) and proving that

$$|x(t)| \preceq x_m \quad \text{for all } t \geq 0. \tag{7.53}$$

We next proceed by induction on k. From (7.30), (7.31), and (7.53), it follows that $|u(t)| \preceq \delta(x_m)$ for all $t \geq 0$. Using Theorem 7.3 ii) a), then given $\gamma \in \mathbb{R}^n_+$ a finite time $\tilde{t}_1 = \tilde{t}_1(\gamma, x(0)) \geq 0$ exists so that

$$|U^{-1}x(t)| \preceq S\delta(x_m) + \gamma \tag{7.54}$$

for all $t \geq \tilde{t}_1$. By Theorem 7.3 ii), the function $\tilde{t}_1(\gamma, \cdot)$ is continuous. By assumption, $x(0)$ satisfies $|U^{-1}x(0)| \preceq S\delta(x_m)$, and for any $\gamma \in \mathbb{R}^n_{+,0}$, then $x(0)$ is contained in the compact set

$$\mathcal{C}_0(\gamma, x_m) \triangleq \{x \in \mathbb{R}^n : |U^{-1}x| \preceq S\delta(x_m) + \gamma\}.$$

Hence, the function $\tilde{t}_1(\gamma, \cdot)$ achieves a maximum over $\mathcal{C}_0(\gamma, x_m)$, and we can define

$$t_1(\gamma, x_m) \triangleq \max_{x \in \mathcal{C}_0(\gamma, x_m)} \tilde{t}_1(\gamma, x),$$

which is finite and nonnegative. Then, it follows that (7.54) holds for all $t \geq t_1$. This establishes (7.36) for $k = 0$ by defining $\bar{t}_f(0, \gamma, x_m) \triangleq t_1(\gamma, x_m)$.

Next, suppose that (7.36) holds for some $k \in \mathbb{Z}_{+,0}$. Since $|x(t)| \preceq |U||U^{-1}x(t)|$, using (7.36) we obtain

$$|x(t)| \preceq |U|S\delta(T^k_\gamma(x_m)) + |U|\,\gamma = T^{k+1}_\gamma(x_m),$$

where we have used (7.32) and (7.35). Therefore, using (7.30) and (7.31), it follows that

$$|u(t)| \preceq \delta(T^{k+1}_\gamma(x_m)) \quad \text{for all } t \geq \bar{t}_f(k, \gamma, x_m).$$

Applying Theorem 7.3 ii) a), there exists $\tilde{t}_{k+1} = \tilde{t}_{k+1}(\gamma, x(\bar{t}_f(k, \gamma, x_m))) \geq 0$ so that

$$|U^{-1}x(t)| \preceq S\delta(T^{k+1}_\gamma(x_m)) + \gamma, \quad \text{for all } t \geq \bar{t}_f(k, \gamma, x_m) + \tilde{t}_{k+1}(\gamma, x(\bar{t}_f(k, \gamma, x_m))). \tag{7.55}$$

By Theorem 7.3 ii), the function $\tilde{t}_{k+1}(\gamma, \cdot)$ is continuous. By our induction assumption (7.36), we have that $x(\bar{t}_f(k, \gamma, x_m))$ is contained in the compact set

$$\mathcal{C}_k(\gamma, x_m) \triangleq \{x \in \mathbb{R}^n : |U^{-1}x| \preceq S\delta(T^k_\gamma(x_m)) + \gamma\}. \tag{7.56}$$

Hence, $\tilde{t}_{k+1}(\gamma, \cdot)$ achieves a maximum over this compact set, and we can define

$$t_{k+1}(\gamma, x_m) \triangleq \max_{x \in \mathcal{C}_k(\gamma, x_m)} \tilde{t}_{k+1}(\gamma, x),$$

which is finite and nonnegative. Defining $\bar{t}_f(k+1,\gamma,x_m) \triangleq \bar{t}_f(k,\gamma,x_m) + t_{k+1}(\gamma,x_m)$, it follows from (7.55) that (7.36) holds for $k+1$. This concludes the proof of Claim 1. $\diamond$

Proof of Claim 2. Straightforward since the map δ is continuous by assumption. $\diamond$

Proof of Claim 3. Straightforward since, by part i), $b = \lim_{k\to\infty} T^k(x_m)$. $\diamond$

Proof of Claim 4. By induction on k. For $k=0$, for any $\xi \in \mathbb{R}_+$, any function $\bar{\eta}(\xi) > 0$ causes (7.37) to be satisfied with $\eta_2(\xi,0) \triangleq \bar{\eta}(\xi)$, since we adopt the convention that $T^0_\gamma(x_m) = T^0(x_m) = x_m$. Then, the result of Claim 4 holds trivially for $k=0$. Next, suppose that Claim 4 holds for some $k \in \mathbb{Z}_{+,0}$. Using (7.35), we have

$$\begin{aligned}\left\|T^{k+1}_\gamma(x_m) - T^{k+1}(x_m)\right\|_\infty &= \left\|T(T^k_\gamma(x_m)) - T(T^k(x_m)) + |U|\,\gamma\right\|_\infty \\ &\le \left\|T(T^k_\gamma(x_m)) - T(T^k(x_m))\right\|_\infty + \||U|\,\gamma\|_\infty \end{aligned} \tag{7.57}$$

Since, by assumption, $T(x_m) \prec x_m$, note from (7.35) that there exists $\bar{\gamma} \in \mathbb{R}^n_+$ such that

$$T_\gamma(x_m) \prec x_m \quad \text{for all } 0 \preceq \gamma \preceq \bar{\gamma}. \tag{7.58}$$

Then, from (7.33) and (7.58), it follows that

$$T(T_\gamma(x_m)) \preceq T(x_m), \quad \text{whence} \quad T(T_\gamma(x_m)) + |U|\,\gamma \preceq T(x_m) + |U|\,\gamma.$$

It then follows, from (7.35), that

$$T^2_\gamma(x_m) \preceq T_\gamma(x_m),$$

and repeating this recursively yields

$$T^k_\gamma(x_m) \preceq T^{k-1}_\gamma(x_m)$$

for all $k \in \mathbb{Z}_+$. Note then that $T^k_\gamma(x_m)$ is componentwise nonincreasing and satisfies

$$0 \preceq T^k_\gamma(x_m) \prec x_m, \quad \text{for all } k \in \mathbb{Z}_+ \text{ and for all } 0 \preceq \gamma \preceq \bar{\gamma}. \tag{7.59}$$

By (7.32) and since the map δ is continuous, then $T : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ also is continuous. Therefore, T is uniformly continuous in the compact set (see, for example, Corwin and Szczarba, 1995)

$$\mathcal{C} \triangleq \{x \in \mathbb{R}^n : 0 \preceq x \preceq x_m\}.$$

This implies that for any $\xi \in \mathbb{R}_+$, there exists $\eta_3 = \eta_3(\xi)$ (independent of x or y) such that

$$\|T(x) - T(y)\|_\infty < \xi \quad \text{whenever } x, y \in \mathcal{C} \text{ and } \|x - y\|_\infty < \eta_3(\xi).$$

Using (7.34) and (7.59), it follows that for any $\xi \in \mathbb{R}_+$, we have

$$\left\|T(T^k_\gamma(x_m)) - T(T^k(x_m))\right\|_\infty < \xi \quad \text{whenever } \left\|T^k_\gamma(x_m) - T^k(x_m)\right\|_\infty < \eta_3(\xi), \tag{7.60}$$

provided $0 \preceq \gamma \preceq \bar{\gamma}$.

We next prove that Claim 4 holds for $k+1$ with

$$\eta_2(\xi, k+1) \triangleq \min\left\{\eta_2\left(\eta_3(\xi/2), k\right), \min_{i=1,\ldots,n} \bar{\gamma}_i, \frac{\xi}{2\left\|\left|U\right|\right\|_\infty}\right\}. \tag{7.61}$$

By (7.61), note that

$$\gamma \in \mathbb{R}^n_{+,0} \quad \text{and} \quad \|\gamma\|_\infty < \eta_2(\xi, k+1) \quad \Longrightarrow \begin{cases} \|\gamma\|_\infty < \eta_2\left(\eta_3(\xi/2), k\right), & \text{(7.62a)} \\ 0 \preceq \gamma \preceq \bar{\gamma}, & \text{(7.62b)} \\ \|\gamma\|_\infty < \dfrac{\xi}{2\left\|\left|U\right|\right\|_\infty}. & \text{(7.62c)} \end{cases}$$

By (7.62a) and our induction assumption, it follows that

$$\gamma \in \mathbb{R}^n_{+,0} \text{ and } \|\gamma\|_\infty < \eta_2(\xi, k+1) \quad \Longrightarrow \quad \left\|T^k_\gamma(x_m) - T^k(x_m)\right\|_\infty < \eta_3(\xi/2),$$

and by (7.60) and (7.62b), it follows that

$$\left\|T(T^k_\gamma(x_m)) - T(T^k(x_m))\right\|_\infty < \xi/2. \tag{7.63}$$

Also, note that

$$\|\gamma\|_\infty < \frac{\xi}{2\left\|\left|U\right|\right\|_\infty} \quad \Longrightarrow \quad \left\|\left|U\right|\right\|_\infty \|\gamma\|_\infty < \xi/2 \quad \Longrightarrow \quad \left\|\left|U\right|\gamma\right\|_\infty < \xi/2. \tag{7.64}$$

From (7.57), (7.62), (7.63), and (7.64), it follows that

$$\gamma \in \mathbb{R}^n_{+,0} \text{ and } \|\gamma\|_\infty < \eta_2(\xi, k+1) \quad \Longrightarrow \quad \left\|T^{k+1}_\gamma(x_m) - T^{k+1}(x_m)\right\|_\infty < \xi.$$

This establishes Claim 4 for $k+1$ and concludes its proof. ◇

The proof of Theorem 7.4 is now complete. □

Theorem 7.4 provides a simple ultimate bound expression and shows that the set $\{x \in \mathbb{R}^n : |U^{-1}x| \preceq S\delta(x_m)\}$ is an estimate of the region of attraction of the ultimate bound. The theorem relies on finding a point x_m such that $T(x_m) \prec x_m$. Although checking this condition analytically might be possible, this cannot be ensured in all cases. Therefore, we provide the following numerical algorithm, and then analyse its convergence.

Algorithm 1 (Numerical Computation of x_m) *Consider a map $T : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$.*

1. *Choose a scalar $c > 0$.*
2. *Define the map $T_c(x) \triangleq T(x) + c\mathbf{1}_n$ and iterate it from $x = 0$, generating the sequence $T^k_c(0)$, for $k = 1, 2, \ldots$*

Theorem 7.5 *Suppose that a map $T : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ satisfies (7.33).*

a) If, choosing $c = \psi > 0$, Algorithm 1 converges to a point $\bar{x}_m^\psi \triangleq \lim_{k\to\infty} T_\psi^k(0)$, then $T(\bar{x}_m^\psi) \prec \bar{x}_m^\psi$. Also, if $0 < \phi < \psi$, then choosing $c = \phi$, Algorithm 1 converges to $\bar{x}_m^\phi \triangleq \lim_{k\to\infty} T_\phi^k(0)$, where $\bar{x}_m^\phi \prec \bar{x}_m^\psi$.

b) If x_m exists such that $T(x_m) \prec x_m$ and if $c > 0$ is chosen small enough in Step 1 of Algorithm 1, then the algorithm converges to a point $\bar{x}_m \triangleq \lim_{k\to\infty} T_c^k(0)$ satisfying $\lim_{k\to\infty} T^k(\bar{x}_m) \preceq \lim_{k\to\infty} T^k(x_m)$.

Proof. a). Convergence of Algorithm 1 to a point $\bar{x}_m^\psi$ implies that $\bar{x}_m^\psi = T_\psi(\bar{x}_m^\psi)$ and by definition of T_ψ and the fact that $\psi > 0$, then $T(\bar{x}_m^\psi) \prec T(\bar{x}_m^\psi) + \psi\mathbf{1}_n = \bar{x}_m^\psi$. Since $\phi < \psi$, we have

$$T_\phi(0) = T(0) + \phi\mathbf{1}_n \prec T(0) + \psi\mathbf{1}_n = T_\psi(0).$$

Since T satisfies (7.33), applying T to the inequality above yields $T(T_\phi(0)) \preceq T(T_\psi(0))$. Then, $T(T_\phi(0)) + \phi\mathbf{1}_n \prec T(T_\psi(0)) + \psi\mathbf{1}_n$ whence $T_\phi^2(0) \prec T_\psi^2(0)$. Repeating this procedure yields

$$T_\phi^k(0) \prec T_\psi^k(0), \text{ for all } k > 0. \tag{7.65}$$

Also, $0 \preceq T_\phi(0)$, whence $T(0) \preceq T(T_\phi(0))$ and $T(0) + \phi\mathbf{1}_n = T_\phi(0) \preceq T(T_\phi(0)) + \phi\mathbf{1}_n = T_\phi^2(0)$. Repeating this procedure yields

$$T_\phi^k(0) \preceq T_\phi^{k+1}(0), \text{ for all } k > 0. \tag{7.66}$$

From (7.66), the sequence $T_\phi^k(0)$ is nondecreasing and from (7.65) it is bounded above by the converging sequence $T_\psi^k(0)$. Therefore, $T_\phi^k(0)$ must converge to some point $\bar{x}_m^\phi$. From (7.65) it follows that $\bar{x}_m^\phi \preceq \bar{x}_m^\psi$. Using (7.33) then $T(\bar{x}_m^\phi) \preceq T(\bar{x}_m^\psi)$, whence $T_\phi(\bar{x}_m^\phi) = T(\bar{x}_m^\phi) + \phi\mathbf{1}_n \prec T(\bar{x}_m^\psi) + \psi\mathbf{1}_n = T_\psi(\bar{x}_m^\psi)$. Hence, $\bar{x}_m^\phi = T_\phi(\bar{x}_m^\phi) \prec T_\psi(\bar{x}_m^\psi) = \bar{x}_m^\psi$. This concludes the proof of a).

b). Since $T(x_m) \prec x_m$, then by Theorem 7.4 i) the limit $b \triangleq \lim_{k\to\infty} T^k(x_m)$ exists and satisfies $b \prec x_m$. This implies that $b = T(b) \prec T(b) + c\mathbf{1}_n = T_c(b) \preceq T_c(x_m)$, where the first inequality follows from $c > 0$ and the second one from the facts that T satisfies (7.33) and $b \prec x_m$. Also, by choosing $c > 0$ small enough, we can guarantee that $T_c(x_m) \prec x_m$. Applying T_c iteratively we arrive to

$$b \prec T_c^k(x_m) \preceq T_c^{k-1}(x_m) \prec x_m.$$

Then, the sequence $T_c^k(x_m)$ is nonincreasing and lower bounded, which implies that it converges to some point b_c satisfying

$$b \preceq b_c \prec x_m. \tag{7.67}$$

Consider next the sequence $T_c^k(0)$. Notice that since $0 \preceq b_c$ and T_c satisfies (7.33), then $0 \preceq T_c(0) \preceq b_c$, and applying T_c iteratively yields

$$T_c^{k-1}(0) \preceq T_c^k(0) \preceq b_c.$$

This implies that $T_c^k(0)$ is nondecreasing and upper bounded by b_c, which shows that Algorithm 1 must converge to some point $\bar{x}_m$ satisfying

$$\bar{x}_m \preceq b_c. \tag{7.68}$$

By a), then $T(\bar{x}_m) \prec \bar{x}_m$ and by assumption, $T(x_m) \prec x_m$. Thus, Theorem 7.4 i) proves that $\lim_{k\to\infty} T^k(\bar{x}_m)$ and $\lim_{k\to\infty} T^k(x_m)$ both exist. Also, (7.67) and (7.68) imply that $\bar{x}_m \prec x_m$. Since T satisfies (7.33), applying T iteratively yields $T^k(\bar{x}_m) \preceq T^k(x_m)$ for all $k \in \mathbb{Z}_+$, whence the result follows straightforwardly. □

Remark 7.6 *If Algorithm 1 converges then, by Theorem 7.5 a), the resulting $\bar{x}_m$ satisfies $T(\bar{x}_m) \prec \bar{x}_m$ and thus the hypotheses of Theorem 7.4 are satisfied. We emphasise that this holds irrespective of how large or small the chosen scalar c is (provided Algorithm 1 converges). On the other hand, the scalar c may need to be small enough to ensure the convergence of Algorithm 1. The use of different values of c for which Algorithm 1 converges yields different points $\bar{x}_m$. Larger values of c for which Algorithm 1 converges are more desirable since they provide larger $\bar{x}_m$, hence resulting in a larger estimate of the region of attraction of the ultimate bound. In some cases, iteration of the map T from different $\bar{x}_m$ provided by Algorithm 1 may converge to different points, corresponding to different ultimate bounds. In addition, if c is small enough, then iteration of the map T from the point $\bar{x}_m$ provided by Algorithm 1 leads to the smallest ultimate bound that can be obtained via application of Theorem 7.4.*

Remark 7.7 *Theorem 7.4, Algorithm 1 and Theorem 7.5 provide a* systematic *method to obtain ultimate bounds for continuous-time linear systems with perturbations bounded componentwise by state-dependent functions.*

7.2.3 Application to Nonlinear Systems

We next show how the method developed above can be applied to a nonlinear system of the form

$$\dot{x}(t) = f(x(t), u(t)), \tag{7.69}$$

where $f(0,0) = 0$ and $A \triangleq \left.\frac{\partial f}{\partial x}\right|_{(0,0)}$ is Hurwitz. Rewriting system (7.69) as

$$\dot{x}(t) = Ax(t) + [f(x(t), u(t)) - Ax(t)],$$

we see that if we can find a continuous function $\delta : \mathbb{R}^n \to \mathbb{R}^n_{+,0}$ so that

$$|f(x(t), u(t)) - Ax(t)| \preceq \delta(x(t)), \quad \text{for all } t \geq 0,$$

and (7.31) is satisfied, then we can analyse the map given by (7.32) and expect to be able to use Theorem 7.4 to estimate an ultimate bound and a region of attraction. In §7.4.2, we illustrate this procedure with an example.

7.3 Ultimate Bounds for Discrete-time Systems

In this section, we develop a systematic method to obtain ultimate bounds for perturbed discrete-time systems. This result is based on componentwise analysis of the system in modal coordinates. In this case, the ultimate bound expressions can be obtained in a more straightforward manner, via a procedure that is different from the one developed in the continuous-time case. In particular, the result for constant perturbation bounds is not needed as an intermediate tool to obtain ultimate bounds for state-dependent perturbation bounds. We therefore directly obtain ultimate bounds for this latter case in §7.3.1 where we also develop the systematic method discussed above. We then show how to apply this result to nonlinear systems in §7.3.2.

7.3.1 State-dependent Perturbation Bounds

As we did for continuous-time systems in §7.2.2, we next provide ultimate bound expressions for linear systems with state-dependent perturbation bounds that satisfy a monotonicity condition [see (7.72) and (7.73) below] and then develop a corresponding systematic method for the discrete-time case.

Theorem 7.8 *Consider the system*

$$x(k+1) = Ax(k) + u(k), \tag{7.70}$$

where $x(k), u(k) \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$ has all its eigenvalues strictly inside the unit circle and Jordan canonical form

$$\Lambda = U^{-1}AU. \tag{7.71}$$

Suppose that

$$|u(k)| \preceq \delta(|x(k)|), \quad \textit{for all } k \geq 0, \tag{7.72}$$

where $\delta : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ is a continuous map satisfying

$$x \preceq y \Rightarrow \delta(x) \preceq \delta(y) \quad \textit{for all } x, y \in \mathbb{R}^n_{+,0}. \tag{7.73}$$

Consider the map $T : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ defined by

$$T(y) \triangleq |\Lambda|\, y + |U^{-1}|\, \delta(|U|y). \tag{7.74}$$

Suppose that a point b satisfying $b = T(b)$ exists. Let $x_m \in \mathbb{R}^n$ denote any point satisfying

$$\lim_{k\to\infty} T^k(|U^{-1}x_m|) = b$$

(note that $x_m = Ub$ is one such point). If the initial condition $x(0)$ satisfies $|U^{-1}x(0)| \preceq |U^{-1}x_m|$, then for any $\epsilon \in \mathbb{R}^n_+$ there exists $\ell = \ell(\epsilon, x_m) \geq 0$, such that for all $k \geq \ell$

a) $|U^{-1}x(k)| \preceq b + \epsilon$.

b) $|x(k)| \preceq |U|\, b + |U|\, \epsilon$.

Proof. Let $x(k) = Uz(k)$ and substitute into (7.70) to obtain

$$z(k+1) = \Lambda z(k) + U^{-1}u(k).$$

Taking magnitudes and using (7.72) yields

$$\begin{aligned} |z(k+1)| &\preceq |\Lambda| \cdot |z(k)| + |U^{-1}|\delta(|Uz(k)|) \\ &\preceq |\Lambda| \cdot |z(k)| + |U^{-1}|\delta(|U| \cdot |z(k)|), \end{aligned}$$

where in the last line we have used (7.73). Define the auxiliary system

$$y(k+1) = |\Lambda| y(k) + |U^{-1}|\delta(|U|\, y(k)) = T(y(k)). \tag{7.75}$$

Note that by (7.73), (7.74), and properties (6.1) and (6.2) in Chapter 6, T satisfies

$$x \preceq y \Rightarrow T(x) \preceq T(y) \quad \text{for all } x, y \in \mathbb{R}^n_{+,0} \tag{7.76}$$

and also $|z(k)| \preceq y(k)$ for all $k \geq 0$ whenever the initial condition $y(0)$ satisfies $|z(0)| \preceq y(0)$. By assumption, $|z(0)| = |U^{-1}x(0)| \preceq |U^{-1}x_m|$ and hence set the initial condition $y(0) = |U^{-1}x_m|$. Then, by assumption, $\lim_{k\to\infty} T^k(y(0)) = b$. Thus, iteration of (7.75) converges to the point $b = T(b) = \lim_{k\to\infty} y(k)$. Therefore, given any $\epsilon \in \mathbb{R}^n_+$, there exists $\ell = \ell(\epsilon, x_m) \geq 0$ such that $y(k) \preceq b + \epsilon$, for all $k \geq \ell$. The proof of a) then follows by recalling that $|U^{-1}x(k)| = |z(k)| \preceq y(k)$. To prove b) note that $|x(k)| \preceq |U| \cdot |U^{-1}x(k)|$ and use a). □

The hypotheses of Theorem 7.8 are weaker than those of its continuous-time counterpart (Theorem 7.4). In Theorem 7.4, it is required that a point x_m exist such that $T(x_m) \prec x_m$. As shown in Theorem 7.4 i), this assumption is sufficient for T to have a fixed point. However, in the discrete-time case, we need only assume the latter, that is, that T has a fixed point. To find such a point, we may iterate T from the origin, as established in the following theorem.

Theorem 7.9 *Let $T : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ be a continuous map satisfying (7.76) and suppose that there exists b satisfying $b = T(b)$. Then, $\lim_{k\to\infty} T^k(0) = \bar{b}$, $\bar{b} = T(\bar{b})$ and $\bar{b} \preceq b$.*

Proof. Since $b = T(b)$, then $b \succeq 0$. Therefore, using (7.76), we have $b = T(b) \succeq T(0) \succeq 0$ and applying T iteratively yields $b \succeq T^k(0) \succeq T^{k-1}(0)$. Thus, the sequence $T^k(0)$ is nondecreasing and upper bounded by b and hence it converges to some point $\bar{b}$ satisfying $T(\bar{b}) = \bar{b}$ and $\bar{b} \preceq b$. □

Remark 7.10 *Theorems 7.8 and 7.9 provide a* systematic *method to obtain ultimate bounds for discrete-time linear systems with perturbations bounded componentwise by a state-dependent function.*

7.3.2 Application to Nonlinear Systems

We next show how the method developed above can be applied to a nonlinear system of the form

$$x(k+1) = f(x(k), u(k)), \tag{7.77}$$

where $f(0,0) = 0$ and $A \triangleq \left.\frac{\partial f}{\partial x}\right|_{(0,0)}$ has all its eigenvalues inside the unit circle. Rewriting system (7.77) as

$$x(k+1) = Ax(k) + \left[f(x(k), u(k)) - Ax(k)\right],$$

we see that if we can find a continuous function $\delta : \mathbb{R}^n_{+,0} \to \mathbb{R}^n_{+,0}$ so that

$$|f(x(k), u(k)) - Ax(k)| \preceq \delta(|x(k)|), \quad \text{for all } k \geq 0,$$

and (7.73) is satisfied, then we can analyse the map given by (7.74) and expect to be able to use Theorem 7.8 to estimate an ultimate bound and a region of attraction. In §7.4.3, we illustrate this procedure with an example.

7.4 Examples

In this section, we apply the results of this chapter to different numerical examples. We also compare the bounds resulting from the application of our method with those obtained via Lyapunov analysis by means of quadratic functions.

7.4.1 Continuous-time System with Constant Perturbation Bounds

Consider the system

$$\dot{x}(t) = \underbrace{\begin{bmatrix} 0 & 1 \\ -1 & -10 \end{bmatrix}}_{A} x(t) + u(t), \tag{7.78}$$

where $u_1(t) = 0$ and $|u_2(t)| \leq 0.1$ for all $t \geq 0$. A classical Lyapunov approach employs a quadratic function $V(x) = x^T P x$, where $P = P^T > 0$ is the solution of $A^T P + PA = -Q$, with $Q > 0$, and analyses its time derivative using the perturbation bound $\|u(t)\|_2 \leq 0.1$, for all $t \geq 0$. For example, this approach is used in Lemma 9.2 of Khalil (2002) to derive an ultimate bound for the 2-norm of x of the form

$$\mu = 2\frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}0.1 + \epsilon, \tag{7.79}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues, respectively, of a real symmetric matrix, and $\epsilon > 0$ can be made arbitrarily small. Numerical minimisation of (7.79) with respect to Q yields $\mu = 1.3837 + \epsilon$, whence $\|x(t)\|_2 \leq 1.3837 + \epsilon$ for all $t \geq t_f$, for some $t_f \geq 0$. We can also

obtain the componentwise bounds $|x_1(t)| \leq 1.3837 + \epsilon$ and $|x_2(t)| \leq 1.3837 + \epsilon$. Note that, since A is Hurwitz, the system (7.78) is ISS with respect to the input u and an ISS analysis then leads to the same bounds (see, for example, Khalil, 2002, Theorems 4.18 and 4.19).

Application of the formula derived in Kofman (2005, Theorem 4) (and extended here to the general —not necessarily diagonalisable— Hurwitz case in Theorem 7.3) results in the tighter bounds $|x_1(t)| \leq 0.1021$ and $|x_2(t)| \leq 0.0204$, and $\|x(t)\|_2 \leq 0.1041$, for all $t \geq t_f$, for some $t_f \geq 0$. In this case, we can identify two reasons why our method yields tighter bounds. First, the information on $u_1(t)$, namely $|u_1(t)| = 0$ is lost in the standard Lyapunov analysis, which requires a bound on $\|u(t)\|_2$. Second, obtaining ultimate bounds in the form of balls by means of quadratic Lyapunov functions seems to not be particularly well-suited to the system (7.78). We can verify this statement by supposing that the initial information on $u(t)$ is that $\|u(t)\|_2 \leq 0.1$ for all $t \geq 0$. Then, using the componentwise bounds $|u_1(t)| \leq 0.1$ and $|u_2(t)| \leq 0.1$ and applying again the formula derived in Kofman (2005, Theorem 4) (or here in Theorem 7.3) yields the bounds $|x_1(t)| \leq 1.1023$ and $|x_2(t)| \leq 0.1225$, whence $\|x(t)\|_2 \leq 1.1091$, for all $t \geq t_f$, for some $t_f \geq 0$. This bound is still tighter than the one obtained above via standard Lyapunov analysis.

7.4.2 Continuous-time System with State-dependent Perturbation Bounds

The system

$$\dot{x} = \underbrace{\begin{bmatrix} 0 & 1 \\ -1 & -10 \end{bmatrix}}_{A} x + \underbrace{\begin{bmatrix} 0 \\ x_1 - \sin(x_1) + \tau(t) \end{bmatrix}}_{u(t)}, \tag{7.80}$$

represents the dynamics of a pendulum with friction, where $x = [x_1\ x_2]^T$ and $\tau(t)$ represents a perturbation torque that is bounded by $|\tau(t)| \leq 0.1$. This system has been expressed in the form suggested in §7.2.3. The matrix A is Hurwitz with Jordan canonical form $\Lambda = U^{-1}AU$, where

$$U = \begin{bmatrix} 0.9949 & -0.1005 \\ -0.1005 & 0.9949 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -0.1010 & 0 \\ 0 & -9.8990 \end{bmatrix}.$$

The term $u(t)$ in (7.80) can be bounded by

$$|u(t)| \preceq \delta(x) \triangleq \begin{bmatrix} 0 \\ \frac{|x_1|^3}{6} + 0.1 \end{bmatrix}.$$

Note that δ satisfies (7.31). The map T, from (7.32), is $T(x) = |U|S\delta(x)$, where $S \triangleq \left|[\mathbb{R}\mathrm{e}(\Lambda)]^{-1}\right| \cdot |U^{-1}|$.

Choosing $c = 0.5$, Algorithm 1 converges to the point $[0.6484\, 0.5297]^T$, from which iteration of the map T converges to $b = [0.1023\ 0.0205]^T$. Hence, Theorem 7.4 concludes that if

$$|U^{-1}x(0)| \preceq S\delta([0.6484\ \ 0.5297]^T) = [0.1477\ \ 0.0149]^T,$$

then given any $\epsilon \in \mathbb{R}^n_+$, a finite time t_f exists so that for all $t \geq t_f$,

$$|U^{-1}x(t)| \preceq \begin{bmatrix} 0.1017 \\ 0.0103 \end{bmatrix} + \epsilon, \quad \text{and} \tag{7.81}$$

$$|x(t)| \preceq \begin{bmatrix} 0.1023 \\ 0.0205 \end{bmatrix} + |U|\,\epsilon. \tag{7.82}$$

These bounds yield the parallelogram and the axis-aligned rectangle shown in Figure 7.1. We have also checked that using $c = 10^{-6}$ in Algorithm 1 yields a point from which iteration of the map T also converges to $b = [0.1023\ 0.0205]^T$. The use of a higher value of c in Algorithm 1 provides a larger estimate of the region of attraction of the ultimate bound.

We next compare the ultimate bounds obtained above with the results obtained from Lyapunov analysis. Extension of a systematic Lyapunov analysis, such as that described in §7.4.1, to this nonlinear perturbation case is not straightforward. We tried analysing this system along the lines in Khalil (2002, Examples 9.2 and 9.5), using a quadratic function $V(x) = x^T P x$, with P to be determined. We performed this analysis bounding $u(t)$ by $\|u(t)\|_2 \leq \frac{|x_1|^3}{6} + 0.1$. Note at this point that the perturbation structure is already lost, since the fact that the first component of $u(t)$ is zero is not taken into account. On the other hand, taking this structure into account makes the analysis case-dependent and difficult to systematically generalise. We next proceed similarly to Khalil (2002, Examples 9.2 and 9.5), and bound the term $|x_1|^3$ by $\alpha d\|x\|$, where αd is the maximum value of $|x_1|^2$ on the level surface $V(x) = d$. Pursuing the analysis in this way, we conclude that, for the values of the parameters in this example, such a method does not yield useful information since the different constraints involved cannot be satisfied.

Having found this procedure uninformative for this example, we proceed in a non-systematic way by employing the function $V(x)$ used in §7.4.1 and analysing the exact possible values of $\dot{V}(x)$ [for all values of $\tau(t)$] on the level surfaces of $V(x)$. Note that the matrix A in (7.80) is the same as that in (7.78) and that the function $V(x)$ used in §7.4.1 minimises the Lyapunov-based formula (7.79) for the 2-norm ultimate bound on the state in the case of a constant perturbation bound. After performing this tedious numerical evaluation, we find that convergence of the system's trajectories to the set enclosed by the level surface $V(x) = 0.0378$ (see Figure 7.1) is guaranteed. For any x in this set, we have $|x_1| \leq 0.1076$ and $|x_2| \leq 0.1181$. These bounds are more conservative than the ones given in (7.82).

In an attempt to obtain a tighter bound on x_2, we find, via trial and error, the Lyapunov function $V_1(x) = x_1^2 + 5x_2^2 + x_1x_2$, which ensures convergence to the set enclosed by the level surface $V_1(x) = 0.0205$, also shown in Figure 7.1. For any x in this set, we have $|x_2| \leq 0.0657$, which is still larger than the bound given in (7.82).

This example illustrates that the Lyapunov approach using quadratic functions may be more conservative, and that finding an appropriate Lyapunov function may be a difficult task. In addition, if the systematic approach of minimising (7.79) is overly conservative, one is obliged to resort to the tedious

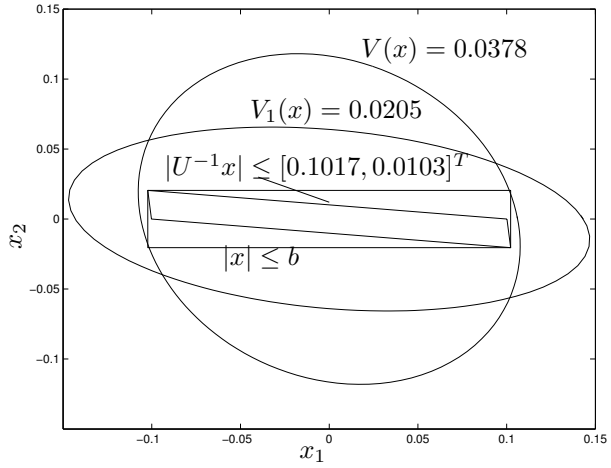


Figure 7.1: Different ultimate bounds in the pendulum system

and complicated procedure of evaluating the time derivative of the Lyapunov function along its level surfaces for all possible values of the perturbation. On the other hand, the approach that we propose provides a systematic method to obtain ultimate bounds that can be computer coded in a simpler way. Moreover, we have shown, in this example, that our approach can lead to tighter bounds.

7.4.3 Discrete-time System

Eq. (7.83) represents the Euler discretisation of a controlled inverted pendulum with a perturbation $w(k)$ that satisfies $|w(k)| \leq 0.01$ and where $x = [x_1\ x_2]^T$.

$$x(k+1) = \begin{bmatrix} 1 & 0.1 \\ -0.9 & 0 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 0.1[\sin(x_1(k)) - x_1(k)] - w(k) \end{bmatrix}. \tag{7.83}$$

The system has been written as suggested in §7.3.2 and has the form $x(k+1) = Ax(k) + u(k)$. The term $u(k)$ can be bounded by $|u(k)| \preceq \delta(|x(k)|)$, where the function $\delta : \mathbb{R}^2_{+,0} \to \mathbb{R}^2_{+,0}$ is given by

$$\delta(z) \triangleq \begin{bmatrix} 0 \\ \frac{z_1^3}{60} + 0.01 \end{bmatrix}$$

and can be easily shown to satisfy (7.73). The matrices U and Λ in the Jordan canonical form of the matrix A, $\Lambda = U^{-1}AU$, are

$$U = \begin{bmatrix} 0.7071 & -0.1104 \\ -0.7071 & 0.9939 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

The map $T : \mathbb{R}^2_{+,0} \to \mathbb{R}^2_{+,0}$, defined as

$$T(y) \triangleq |\Lambda|\, y + |U^{-1}|\delta(|U|y)$$

has a fixed point at $b = [0.0177 \;\; 0.0126]^T$. Also, $x_m = [5 \;\; 5]^T$ satisfies $\lim_{k\to\infty} T^k(|U^{-1}x_m|) = b$. Then, Theorem 7.8 states that if $|U^{-1}x(0)| \preceq |U^{-1}x_m|$, then for any $\epsilon \in \mathbb{R}^n_+$ there exists $\ell \geq 0$ such that for all $k \geq \ell$, $|U^{-1}x(k)| \preceq b + \epsilon$ and $|x(k)| \preceq |U|b + |U|\epsilon$. These bounds yield the parallelogram and the axis-aligned rectangle, respectively, shown in Figure 7.2.

To compare with a Lyapunov approach, the perturbed system was analysed using the quadratic function $V(x) \triangleq x^T P x$, where P is the solution to $A^T P A - P = -I$. After analysing the increment $\Delta V(x(k)) \triangleq V(x(k+1)) - V(x(k))$ on the level surfaces of V, ΔV satisfied $\Delta V(x) > 0$ at some point x for which $V(x) = 0.00147$ and then, using the function V we cannot insure an ultimate bound smaller than this level surface. This surface is shown in Figure 7.2. We stress that the exact value of the nonlinear function ΔV was numerically analysed, without bounding any term. If the Lyapunov analysis had been performed —as is usually done— by bounding some expressions (like $\sin(x)$ for instance), the resulting ultimate bound would have been significantly more conservative. Note that in the case of higher order systems, the numerical analysis of the exact value of the Lyapunov function increment is computationally intractable and thus the usual approach of bounding terms seems to be the only resort.

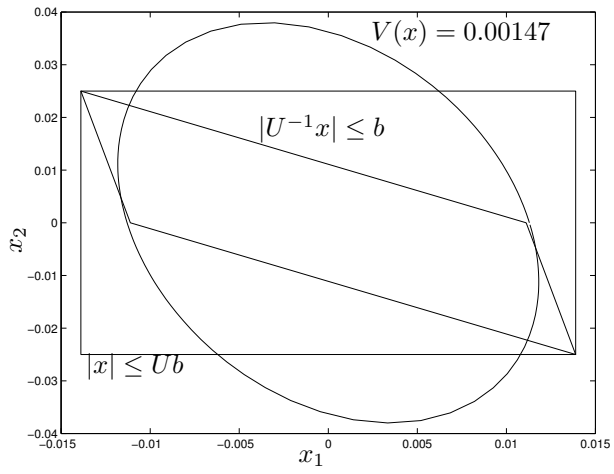


Figure 7.2: Different ultimate bounds in the discretised inverted pendulum

In this example, the ultimate bound obtained with the suggested method cannot be said to be tighter than the one obtained via Lyapunov analysis. However, one can combine the results obtained by both methodologies and compute an ultimate bound given by the intersection of the parallelogram and the ellipse shown in Figure 7.2. Moreover, for this example, the Lyapunov analysis method shows global convergence to the ultimate bound given by the ellipse. Thus, since our method guarantees convergence to the parallelogram shown in Figure 7.2 from the set $\{x \in \mathbb{R}^2 : |U^{-1}x| \leq |U^{-1}x_m|\}$ with $x_m = [5 \;\; 5]^T$, and this set contains the ellipse shown in Figure 7.2, then global convergence to the ultimate bound given by the intersection of the parallelogram and the ellipse is ensured. Thus, this example illustrates how the strengths of both methodologies can be combined to obtain tighter bounds and larger

regions of attraction.

7.5 Chapter Summary

In this chapter, we have presented a systematic method to obtain ultimate bounds for both continuous- and discrete-time systems. The method is based on componentwise analysis of the system in modal coordinates and thus exploits the system geometry as well as the perturbation structure without requiring calculation of a Lyapunov function. We have developed the method for linear perturbed systems with componentwise state-dependent perturbation bounds and then shown that the method may be applied to nonlinear systems by treating nonlinear terms as perturbations. The resulting ultimate bounds are given as simple expressions in terms of the solution of a fixed point problem which can be solved analytically or numerically. The method also provides an estimate of the region of attraction of the ultimate bound. We have shown, by means of examples, that the proposed method may offer a simple alternative to the classical Lyapunov-based analysis and may sometimes yield tighter bounds. In addition, the strengths of both methodologies can be combined to obtain even tighter bounds and/or larger regions of attraction.

Chapter 8

Summary and Future Work

8.1 General Overview

In this thesis, we have addressed different aspects of quantisation in feedback control systems. In particular, we have dealt with quadratic stabilisation via quantised static feedback (focusing on quantisation density) in Part I of this thesis, and with the derivation of componentwise ultimate bounds for perturbed systems, especially when quantisation is regarded as a perturbation, in Part II of this thesis. In this final chapter, we summarise the main results presented throughout the thesis. We also discuss some future research directions.

8.1.1 Quantisation Density

Throughout Part I of this thesis, we have dealt with the concept of quantisation density. The density of a quantiser is a measure of how efficient a quantiser is in the use of its levels.

In Chapter 2, we have briefly reviewed quadratic stabilisation of linear discrete-time systems and then focused on quantisation density in the context of multiple-input systems. We have generalised the definition of quantisation density of Elia and Mitter (2001) to multiple-input systems and derived several new results regarding quantisation density. We have also posed the problem of optimising quantisation density over all quantisers that quadratically stabilise a given multiple-input system and derived an important result that reveals the structure of quantisers that optimise density.

In Chapter 3, we have focused on the characterisation of quantisers that quadratically stabilise a given multiple-input system. As a first step toward this characterisation, we considered quantisers having a form that can be interpreted as the simplest possible in some appropriate sense. We derived necessary and sufficient conditions for these quantisers to quadratically stabilise a system. We did this by means of explicit geometric considerations. We thus developed a novel geometric approach to

quadratic stabilisation of multiple-input systems by means of quantisers. We also employed this geometric approach to design quantisers with finite density that can stabilise multiple-input systems having an arbitrary number of inputs.

In Chapter 4, we dealt with single-input systems. For these systems, we enhanced the geometric approach of Chapter 3 to explore quantiser coarseness from a state-space standpoint, as opposed to the standard input-space-based concept of quantisation density. We introduced a novel type of quantisers (CAQS quantisers) and analysed the relationships between CAQS quantisers and quantisers that minimise quantisation density in the standard sense. We also showed how to directly utilise CAQS quantisers to design static output feedback strategies that employ quantisers of infimum density. We concluded this chapter by showing how to recover a well-known result on infimum quantisation density by means of our approach.

In Chapter 5, we solved a special case of infimum quantisation density problem for multiple-input systems. Specifically, we derived the infimum density over all quantisers that quadratically stabilise the system and have levels in a one-dimensional subspace of the input space. We also showed that our result conflicts with a previously published result. We also provided a counterexample to the latter result.

In summary, we have addressed the infimum quantisation density problem for multiple- and single-input systems and have derived several new results contributing to this problem. The reader is referred to §1.3 in Chapter 1 for a list of references to the specific contributions of this part of the thesis.

8.1.2 Componentwise Ultimate Bounds for Perturbed Systems

Throughout Part II of this thesis, we have dealt with the derivation of componentwise ultimate bounds for perturbed systems. Deriving tight ultimate bounds for a perturbed system is important since ultimate bounds provide a measure of system performance in steady state.

In Chapter 6, we have derived ultimate bound expressions for discrete-time and sampled-data perturbed systems, especially when the perturbations arise from quantisation. We have considered a setting where each signal connecting controller and plant may have an independent scalar quantiser. Moreover, each scalar quantiser may be of a different type and have different features. Regarding a quantised variable as a perturbation on the corresponding unquantised variable turns the original quantised system into a perturbed system, where the perturbation has a natural componentwise bound. We have therefore derived ultimate bounds on the system states that explicitly take account of the componentwise structure of the perturbation bound. The ultimate bounds derived also have a componentwise form, and can be systematically computed without having to, for example, select a suitable Lyapunov function for the system. A very important feature of the derivations of this chapter is that they can directly accommodate feedback schemes where quantisers of different characteristics and/or types affect different signals in the same system. We have demonstrated the applicability and potential of the method by means of an

example taken from recent literature on the topic of control over communication networks.

In Chapter 7, we have extended the results of Chapter 6 to deal with perturbed systems where the perturbation bounds have more general componentwise forms. We have provided systematic methods to obtain componentwise ultimate bounds on the system state for continuous- and discrete-time perturbed systems. Since the perturbations are allowed to be bounded by state-dependent functions, the method could then be applied to nonlinear systems by regarding them as perturbed linear systems. The derived methods also provide an estimate of the region of attraction of the ultimate bound. We have shown by means of examples that these methods may offer a simple alternative to the classical Lyapunov-based analysis and may sometimes yield tighter bounds. In addition, the strengths of both methodologies can be combined to obtain even tighter bounds and/or larger regions of attraction.

In summary, we have developed systematic methods to obtain componentwise ultimate bounds on the system state for sampled-data, discrete- and continuous-time systems having componentwise perturbation bounds. The main features of these methods are their systematic nature and their flexibility in dealing with componentwise-structured perturbation bounds. The reader is again referred to §1.3 in Chapter 1 for a list of references to the specific contributions of this part of the thesis.

8.2 Future Research

8.2.1 Quantisation Density

In spite of the new results given in this thesis, the infimum density problem for multiple-input systems still remains largely open. The results that we have derived help, we believe, to more clearly identify the difficulties involved in solving the infimum density problem that are not manifested when dealing with a single-input system. We next comment on these difficulties and reveal some future research directions.

Given a discrete-time single-input open-loop-unstable system

$$x(k+1) = Ax(k) + Bu(k),$$

where the pair (A, B) is stabilisable, and a quadratic CLF $V(x) = x^T Px$, we highlight the following feature: the number of inputs of the system, namely $m = 1$, coincides with the minimum dimension of the input space that a stabilising feedback can utilise. Consequently, any stabilising feedback is forced to have values in a one-dimensional subspace of the input space, and this subspace coincides with the system input space.

On the other hand, multiple-input systems do not necessarily have the feature mentioned in the preceding paragraph. That is, the minimum dimension of the input space that a stabilising feedback can utilise does not necessarily coincide with the number of system inputs. For instance, in §3.5.3 in Chapter 3 we have seen an example of a system having 4 inputs, but where a subspace of the input space

of dimension 3 was sufficient for stabilisation. The absence of such feature gives rise to the following interesting question: can the infimum density always be achieved over quantisers that have levels in a subspace of the input space of minimum dimension?

In Chapter 5, we have considered the special case where the minimum dimension of the input space necessary for stabilisation is 1. We have shown that even in this simpler case, the answer to the above question is still unknown. Further research is needed to investigate this issue in more detail.

All of the derivations of Part I of the thesis were performed for a given quadratic CLF. Obviously, once the infimum density with respect to a given CLF was found, the next step would be to optimise the result over all quadratic CLFs. This optimisation has already been performed for single-input system but constitutes another topic for further research in the case of multiple-input systems.

More generally, future research topics may include the extension of results to systems involving uncertainty. Different sorts of uncertain systems could be considered. Examples would be systems with parametric uncertainty and systems having an additional perturbation term.

On a different level, we could consider different quantisation density measures. The quantisation density measure that we have considered in this thesis takes account of the separation in the magnitudes of the quantisation levels of a quantiser. We note that the density of a quantiser is not affected by the angular separation of its levels. Therefore, we could consider other quantisation density measures that might combine magnitude and angular separation of the quantisation levels of a quantiser. This would lead to density optimisation problems that could, in principle, be completely different from the one considered throughout this thesis.

8.2.2 Componentwise Ultimate Bounds for Perturbed Systems

The ultimate bounds derived in Chapter 6 are globally valid, as opposed to the ones derived in Chapter 7. The derivation of conditions under which the bounds in the latter chapter are globally valid thus constitutes a possible future research topic.

In Part II of this thesis, we have given examples where the application of our ultimate bound derivation methods yield acceptably tight bounds. However, our method is not *guaranteed* to provide bounds that are tighter than those obtained via quadratic Lyapunov functions. An interesting future research topic would thus be to investigate what type of systems cause our method to yield tighter bounds.

Application of our ultimate bound derivation method to nonlinear systems relies on regarding such systems as perturbed linear systems, having state-dependent perturbation bounds that satisfy a monotonicity condition. Note that useful information on the evolution of the system may be lost when bounding nonlinear terms. Therefore, another interesting future research topic would be the enhancement of our methods to exploit stabilising nonlinear terms. This enhancement could be achieved, perhaps, by combining Lyapunov-function-based methods with our componentwise approach.

Appendix A

Proof of Theorems 3.10 and 4.7

Throughout this appendix, we will use the following notation. For any $x \in \mathbb{R}^n$, x_- and x_+ denote the vectors formed by the first $n-\ell$ and the last ℓ components of x, respectively, so that $x = [x_-^T \; x_+^T]^T$. The matrices $\tilde{D} \in \mathbb{R}^{n\times\ell}$ and $\Lambda \in \mathbb{R}^{n\times n}$ will be expressed as

$$\tilde{D} = \begin{bmatrix} \tilde{D}_- \\ \tilde{D}_+ \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -\Lambda_- & 0 \\ 0 & \Lambda_+ \end{bmatrix}, \tag{A.1}$$

where

$$\begin{aligned} &\tilde{D}_- \in \mathbb{R}^{n-\ell\times\ell}, && \Lambda_- = -\mathrm{diag}(\lambda_1,\ldots,\lambda_{n-\ell}) \in \mathbb{R}^{n-\ell\times n-\ell}, && \Lambda_- > 0, \\ &\tilde{D}_+ \in \mathbb{R}^{\ell\times\ell}, && \Lambda_+ = \mathrm{diag}(\lambda_{n-\ell+1},\ldots,\lambda_n) \in \mathbb{R}^{\ell\times\ell}, && \Lambda_+ > 0. \end{aligned}$$

Note that $\ell = 1$ for Theorem 4.7. For future reference, note that (A.1) implies

$$-\tilde{D}_-^T\Lambda_-^{-1}\tilde{D}_- + \tilde{D}_+^T\Lambda_+^{-1}\tilde{D}_+ = \tilde{D}^T\Lambda^{-1}\tilde{D}. \tag{A.2}$$

First, we need to prove the following five claims, which are used throughout this appendix.

Claim 1 *Suppose that $\tilde{D}_+$ is nonsingular and let $\tilde{p} \in \tilde{\mathcal{P}}$. Then, $\tilde{p} \in \tilde{X}(u)$ if and only if*

$$\tilde{p}_-^T N\tilde{p}_- + v^T\tilde{p}_- + \tilde{a}^T\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\tilde{a} + u^THu < 0, \tag{A.3}$$

where

$$N \triangleq -\Lambda_- + \tilde{D}_-\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\tilde{D}_-^T, \tag{A.4}$$

$$v^T \triangleq -2\,\tilde{a}^T\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\tilde{D}_-^T. \tag{A.5}$$

Also, $\tilde{p} \in \tilde{X}_0(u)$ if and only if (A.3) holds replacing '$<$' by '$\leq$'.

Proof of Claim 1. From (3.29), $\tilde{p} \in \tilde{X}(u)$ if and only if

$$\tilde{p}^T\Lambda\tilde{p} + u^THu < 0. \tag{A.6}$$

Since by assumption $\tilde{D}_+$ is nonsingular and $\tilde{p} \in \tilde{\mathcal{P}}$, then $\tilde{D}^T\tilde{p} = \tilde{a}$, and using (A.1) it follows that

$$\tilde{p}_+ = \tilde{D}_+^{-T}\left(\tilde{a} - \tilde{D}_-^T\tilde{p}_-\right). \tag{A.7}$$

Using (A.1), and substituting (A.7) into (A.6) yields

$$-\tilde{p}_-^T\Lambda_-\tilde{p}_- + \left(\tilde{a}^T - \tilde{p}_-^T\tilde{D}_-\right)\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\left(\tilde{a} - \tilde{D}_-^T\tilde{p}_-\right) + u^THu < 0, \tag{A.8}$$

The proof of the first part follows straightforwardly from (A.8). The proof for the case $\tilde{p} \in \tilde{X}_0(u)$ follows identical steps, replacing '$<$' by '$\leq$' in (A.6) and (A.8). This concludes the proof of Claim 1. □

Claim 2 *Suppose that $\tilde{\mathcal{P}} \subset \tilde{X}_0(u)$. Then, $\tilde{D}_+$ is nonsingular.*

Proof of Claim 2. For a contradiction, assume that $\tilde{D}_+$ is singular. Then, there exists $\tilde{p}_+ \neq 0$ such that $\tilde{D}_+\tilde{p}_+ = 0$. Since $\tilde{D}$ has linearly independent columns by assumption, note that $\tilde{\mathcal{P}}$ is nonempty. Then, let $\tilde{q} \in \tilde{\mathcal{P}}$ and consider, for all $\alpha \in \mathbb{R}$, the points

$$r_\alpha \triangleq \alpha\begin{bmatrix}0\\ \tilde{p}_+\end{bmatrix} + \tilde{q}.$$

Note that $r_\alpha \in \tilde{\mathcal{P}}$ and hence $r_\alpha \in \tilde{X}_0(u)$ for all $\alpha \in \mathbb{R}$. We have, using (A.1),

$$r_\alpha^T\Lambda r_\alpha = \tilde{p}_+^T\Lambda_+\tilde{p}_+\,\alpha^2 + 2[0\ \ \tilde{p}_+^T]\Lambda\tilde{q}\,\alpha + \tilde{q}^T\Lambda\tilde{q},$$

which is a quadratic polynomial in α, whose leading coefficient, namely $\tilde{p}_+^T\Lambda_+\tilde{p}_+$, is positive. Thus, we can make $r_\alpha^T\Lambda r_\alpha$ as large as desired by selecting α such that $|\alpha|$ is big enough. In particular, we can find some $\alpha' \in \mathbb{R}$, such that $r_{\alpha'}^T\Lambda r_{\alpha'} > -u^THu$. Hence, we have $r_{\alpha'}^T\Lambda r_{\alpha'} + u^THu > 0$ and thus $r_{\alpha'} \notin \tilde{X}_0(u)$, contradicting the fact that $r_\alpha \in \tilde{X}_0(u)$ for all $\alpha \in \mathbb{R}$. This concludes the proof of Claim 2. □

Claim 3 *Suppose that $\tilde{\mathcal{P}} \subset \tilde{X}_0(0)$. Then, $\tilde{a} = 0$.*

Proof of Claim 3. By Claim 2, $\tilde{D}_+$ is nonsingular. Then, the point $\tilde{p} = [0\ (\tilde{D}_+^{-T}\tilde{a})^T]^T$ satisfies $\tilde{p} \in \tilde{\mathcal{P}}$ and hence $\tilde{p} \in \tilde{X}_0(0)$. From (3.30), $\tilde{p}$ must satisfy $\tilde{p}^T\Lambda\tilde{p} \leq 0$. Note [see (A.1)] that $\tilde{p}^T\Lambda\tilde{p} = \tilde{a}^T\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\tilde{a} \geq 0$ because $\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T} > 0$ since $\Lambda_+ > 0$ and $\tilde{D}_+$ is nonsingular. Then, $\tilde{a}^T\tilde{D}_+^{-1}\Lambda_+\tilde{D}_+^{-T}\tilde{a} = 0$, which implies that $\tilde{a} = 0$, proving Claim 3. □

Claim 4 *Suppose that $\tilde{D}_+$ is nonsingular and consider the matrix N defined in (A.4). Then,*

1. *$N < 0$ if and only if $\tilde{D}^T\Lambda^{-1}\tilde{D} > 0$.*

2. $N \leq 0$ *if and only if* $\tilde{D}^T \Lambda^{-1} \tilde{D} \geq 0$.

Proof of Claim 4. Using (A.4), we have

$$N < 0 \Leftrightarrow -\Lambda_- + \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{D}_-^T < 0.$$

Recall that $\Lambda_-, \Lambda_+ > 0$. Hence,

$$\begin{aligned} N < 0 &\Leftrightarrow -\mathrm{I}_{n-\ell} + \Lambda_-^{-1/2} \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+^{1/2} \Lambda_+^{1/2} \tilde{D}_+^{-T} \tilde{D}_-^T \Lambda_-^{-1/2} < 0 \\ &\Leftrightarrow -\mathrm{I}_\ell + \Lambda_+^{1/2} \tilde{D}_+^{-T} \tilde{D}_-^T \Lambda_-^{-1/2} \Lambda_-^{-1/2} \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+^{1/2} < 0 \\ &\Leftrightarrow -\Lambda_+^{-1} + \tilde{D}_+^{-T} \tilde{D}_-^T \Lambda_-^{-1/2} \Lambda_-^{-1/2} \tilde{D}_- \tilde{D}_+^{-1} < 0 \\ &\Leftrightarrow -\tilde{D}_+^T \Lambda_+^{-1} \tilde{D}_+ + \tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- < 0 \\ &\Leftrightarrow -\tilde{D}^T \Lambda^{-1} \tilde{D} < 0 \Leftrightarrow \tilde{D}^T \Lambda^{-1} \tilde{D} > 0, \end{aligned}$$

where we have used (A.2). This proves part 1. The proof of part 2 follows identical steps, replacing '$<$' by '$\leq$' and '$>$' by '$\geq$'. This concludes the proof of Claim 4. □

Claim 5 *Suppose that* $\tilde{D}_+$ *is nonsingular and consider* N *and* v *as defined in (A.4) and (A.5), respectively. Suppose that* $N < 0$*. Then,*

$$\max_{x \in \mathbb{R}^{n-\ell}} x^T N x + v^T x + \tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a} = \tilde{a}^T \left(\tilde{D}^T \Lambda^{-1} \tilde{D} \right)^{-1} \tilde{a}. \tag{A.9}$$

Proof of Claim 5. Since $N < 0$, then $x^T N x + v^T x$ is maximised at the unique point where its gradient vanishes. It is then easy to check that

$$\max_{x \in \mathbb{R}^{n-\ell}} x^T N x + v^T x = -\frac{1}{4} v^T N^{-1} v. \tag{A.10}$$

From (A.4) and using a matrix inversion formula, we have

$$-N^{-1} = \Lambda_-^{-1} + \Lambda_-^{-1} \tilde{D}_- \left(\tilde{D}^T \Lambda^{-1} \tilde{D} \right)^{-1} \tilde{D}_-^T \Lambda_-^{-1}, \tag{A.11}$$

where we have used (A.2). Substituting (A.5) and (A.11) into (A.10) yields

$$\begin{aligned} -\frac{1}{4} v^T N^{-1} v = {} & \tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a} \\ & + \tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- \left(\tilde{D}^T \Lambda^{-1} \tilde{D} \right)^{-1} \tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a}. \end{aligned} \tag{A.12}$$

Note that $\tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} = (\tilde{D}_+^T \Lambda_+^{-1} \tilde{D}_+)^{-1}$ and hence, using (A.2),

$$\begin{aligned} \tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} &= \left(\tilde{D}_+^T \Lambda_+^{-1} \tilde{D}_+ - \tilde{D}^T \Lambda^{-1} \tilde{D} \right) \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \\ &= \mathrm{I}_\ell - \left(\tilde{D}^T \Lambda^{-1} \tilde{D} \right) \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T}. \end{aligned} \tag{A.13}$$

The result then follows by substituting (A.13) into (A.12) and adding the term $\tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a}$. □

Proof of Theorem 3.10

Proof of Theorem 3.10 part 1

Necessity. Note that $\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$ implies that $\tilde{\mathcal{P}} \subset \tilde{X}_0(0)$ [see (3.30)]. Then, Claim 3 proves (3.35). By Claim 2, $\tilde{D}_+$ is nonsingular and hence (A.7) holds for all $\tilde{p} \in \tilde{\mathcal{P}}$. Since $\tilde{a} = 0$, from (A.7) it follows that $\tilde{p} = 0$ if and only if $\tilde{p}_- = 0$. Also, for any $\tilde{p}_- \in \mathbb{R}^{n-\ell}$, there exists $\tilde{p}_+ \in \mathbb{R}^{\ell}$ such that $\tilde{p} \in \tilde{\mathcal{P}}$. Using Claim 1 and setting $\tilde{a} = 0$ and $u = 0$ in (A.3) and (A.5), we have that the expression

$$\tilde{p}_-^T N \tilde{p}_- \tag{A.14}$$

is negative for any nonzero $\tilde{p}_- \in \mathbb{R}^{n-\ell}$ and zero for $\tilde{p}_- = 0$, whence $N < 0$. Then, (3.34) follows from Claim 4, concluding the necessity part of the proof.

Sufficiency. Since $\tilde{D}^T \Lambda^{-1} \tilde{D} = -\tilde{D}_-^T \Lambda_-^{-1} \tilde{D}_- + \tilde{D}_+^T \Lambda_+^{-1} \tilde{D}_+ > 0$ and $\Lambda_- > 0$, then $\tilde{D}_+$ is nonsingular. Then Claim 4 shows that $N < 0$, and (A.14) is negative for all nonzero $\tilde{p}_- \in \mathbb{R}^{n-\ell}$, whence Claim 1 proves that $\tilde{\mathcal{P}} \setminus \{0\} \subset \tilde{X}(0)$. This concludes the proof of part 1.

Proof of Theorem 3.10 part 2

We begin by proving that $u \neq 0$. For a contradiction, assume that $u = 0$. Then, $\tilde{\mathcal{P}} \subset \tilde{X}(0) \subset \tilde{X}_0(0)$. Claim 3 then proves that $\tilde{a} = 0$, whence $0 \in \tilde{\mathcal{P}}$, implying that $0 \in \tilde{X}(0)$. But (3.29) shows that this is a contradiction and thus we have proved that $u \neq 0$.

Since $\tilde{\mathcal{P}} \subset \tilde{X}(u) \subset \tilde{X}_0(u)$, Claim 2 shows that $\tilde{D}_+$ is nonsingular. For a contradiction, suppose that the matrix N defined in (A.4) satisfies $N \not\leq 0$. Note that N is symmetric and hence all its eigenvalues are real. Since $N \not\leq 0$, it has a positive eigenvalue. Let μ denote a positive eigenvalue of N and let $w \neq 0$ satisfy $Nw = \mu w$ and $w^T w = 1$. By (A.7) and Claim 1, we have that (A.3) holds for all $\tilde{p}_- \in \mathbb{R}^{n-\ell}$. In particular, for $\tilde{p}_- = \alpha w$, for any $\alpha \in \mathbb{R}$, we have

$$\mu\alpha^2 + v^T w \alpha + \tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a} + u^T H u < 0. \tag{A.15}$$

This is a contradiction since the left-hand side of (A.15) is a quadratic polynomial in α with leading coefficient $\mu > 0$ and hence we can always find $\alpha \in \mathbb{R}$ so that (A.15) is not satisfied. Therefore, we have shown by contradiction that $N \leq 0$ and using Claim 4 we establish (3.36). This concludes the proof of part 2.

Proof of Theorem 3.10 part 3

Note that $\tilde{D}^T \Lambda^{-1} \tilde{D} > 0$ implies that $\tilde{D}_+$ is nonsingular and then (A.7) holds for all $\tilde{p} \in \tilde{\mathcal{P}}$. Also, Claim 4 shows that the matrix N defined in (A.4) satisfies $N < 0$.

Necessity. By Claim 1 and (A.7), (A.3) holds for all $\tilde{p}_- \in \mathbb{R}^{n-\ell}$. Since $N < 0$, then the supremum of the left-hand side of (A.3) over all $\tilde{p}_- \in \mathbb{R}^{n-\ell}$ is a maximum and by (A.3) is negative. Using Claim 5, then

$$\tilde{a}^T \left(\tilde{D}^T \Lambda^{-1} \tilde{D}\right)^{-1} \tilde{a} + u^T H u < 0, \tag{A.16}$$

which is equivalent to (3.37). This concludes the necessity part of the proof.

Sufficiency. By (3.37), (A.16) holds. From (A.16) and Claim 5, we have

$$\max_{x \in \mathbb{R}^{n-\ell}} x^T N x + v^T x + \tilde{a}^T \tilde{D}_+^{-1} \Lambda_+ \tilde{D}_+^{-T} \tilde{a} + u^T H u < 0. \tag{A.17}$$

Hence, (A.3) holds for all $\tilde{p}_- \in \mathbb{R}^{n-\ell}$ and then for any $\tilde{p} \in \tilde{\mathcal{P}}$, Claim 1 shows that also $\tilde{p} \in \tilde{X}(u)$, proving that $\tilde{\mathcal{P}} \subset \tilde{X}(u)$. This concludes the proof of part 3.

Proof of Theorem 3.10 part 4

The proof of part 4 is identical to that of part 3, replacing '$<$' by '$\leq$' in (A.16) and (A.17).

The proof of Theorem 3.10 is now complete.

Proof of Theorem 4.7

Note that the assumptions of Theorem 4.7 are a special case of those of Theorem 3.10, with $\ell = m = 1$. We will thus utilise Claims 1 to 5 with $\tilde{D} = \tilde{d} \in \mathbb{R}^n$, $\tilde{D}_+ = \tilde{d}_+ \in \mathbb{R}$ and $\tilde{D}_- = \tilde{d}_- \in \mathbb{R}^{n-1}$. Also, note that $\Lambda_+ = \lambda_n \in \mathbb{R}$.

Proof of Theorem 4.7 part 1

Since $\tilde{d}^T \Lambda^{-1} \tilde{d} = -\tilde{d}_-^T \Lambda_-^{-1} \tilde{d}_- + \lambda_n^{-1} \tilde{d}_+^2 = 0,\ \tilde{d} \neq 0,\ \Lambda_- > 0,$ and $\lambda_n > 0$, then $\tilde{d}_+ \neq 0$. Then, Claim 4 shows that the matrix N in (A.4) satisfies $N \leq 0$.

Necessity. Note that N necessarily has one zero eigenvalue or else $N < 0$, and Claim 4 would establish that $\tilde{d}^T \Lambda^{-1} \tilde{d} > 0$, contradicting the assumption. Therefore, let $w \in \mathbb{R}^{n-1}$ satisfy $w \neq 0$ and $Nw = 0$. For every $\alpha \in \mathbb{R}$, let $\tilde{p}_- = \alpha w$, and let $\tilde{p}_+$ satisfy (A.7), so that $\tilde{p} \in \tilde{\mathcal{P}}$. Since $\tilde{\mathcal{P}} \subset \tilde{X}(u)$ and $\tilde{D}_+ = \tilde{d}_+ \neq 0$, then Claim 1 shows that (A.3) holds. From (A.3) and since $\tilde{p}_- = \alpha w$ and $Nw = 0$, then

$$\alpha v^T w + \tilde{d}_+^{-2} \lambda_n \tilde{a}^2 + H u^2 < 0,$$

for all $\alpha \in \mathbb{R}$. Therefore, $v^T w = 0$. From (A.5), then either $\tilde{a} = 0$ or $\tilde{d}_-^T w = 0$. From (A.4), we have

$$Nw = -\Lambda_- w + \lambda_n \tilde{d}_+^{-2} \tilde{d}_- \tilde{d}_-^T w. \tag{A.18}$$

Since $\Lambda_- > 0$, $w \neq 0$ and $Nw = 0$, it follows from (A.18) that $\tilde{d}_-^T w \neq 0$. Hence, $\tilde{a} = 0$.

Sufficiency. Since $N \leq 0$, $u \neq 0$ and $H < 0$, then

$$\tilde{p}_-^T N \tilde{p}_- + Hu^2 < 0 \tag{A.19}$$

for all $\tilde{p}_- \in \mathbb{R}^{n-1}$. Since $\tilde{a} = 0$, then v in (A.5) satisfies $v = 0$. Therefore, given any $\tilde{p} \in \tilde{\mathcal{P}}$, then $\tilde{p}_-$ satisfies (A.19), and since $v = 0$ and $\tilde{a} = 0$, then $\tilde{p}_-$ also satisfies (A.3). Then, using Claim 1 we show that given any $\tilde{p} \in \tilde{\mathcal{P}}$, then $\tilde{p} \in \tilde{X}(u)$, establishing that $\tilde{\mathcal{P}} \subset \tilde{X}(u)$.

Proof of Theorem 4.7 part 2

Necessity. We first establish that if there exists $\tilde{p} \in \tilde{\mathcal{P}}$, $\tilde{p} \notin \tilde{X}(u)$ and $\tilde{\mathcal{P}} \setminus \{\tilde{p}\} \subset \tilde{X}(u)$, then $\tilde{p}$ satisfies (4.18). Choose a nonzero $z \in \mathbb{R}^n$ such that $\tilde{d}^T z = 0$ and consider the line $\mathcal{L} \triangleq \{y \in \mathbb{R}^n : y = \alpha z + \tilde{p}, \text{ for some } \alpha \in \mathbb{R}\}$. Note that $\tilde{p} \in \mathcal{L} \subset \tilde{\mathcal{P}}$. Since $\tilde{\mathcal{P}} \setminus \{\tilde{p}\} \subset \tilde{X}(u)$ and $\tilde{p} \notin \tilde{X}(u)$, we have $y^T \Lambda y + Hu^2 < 0$ for all $y \in \mathcal{L} \setminus \{\tilde{p}\}$ and $\tilde{p}^T \Lambda \tilde{p} + Hu^2 \geq 0$. Therefore,

$$z^T \Lambda z \alpha^2 + 2 z^T \Lambda \tilde{p} \alpha + \tilde{p}^T \Lambda \tilde{p} + Hu^2 < 0, \tag{A.20}$$

only if $\alpha \in \mathbb{R} \setminus \{0\}$. Since the left-hand side of (A.20) is a continuous function of α, it follows that $\tilde{p}^T \Lambda \tilde{p} + Hu^2 = 0$, proving (4.18).

Hence, it follows that $\tilde{\mathcal{P}} \subset \tilde{X}_0(u)$ and Claim 2 shows that $\tilde{d}_+ \neq 0$. Then, from Claim 1, $\tilde{p}_-$ is the only point in $\mathbb{R}^{n-1}$ that does not satisfy (A.3) and from Claim 1, we have

$$\tilde{p}_-^T N \tilde{p}_- + v^T \tilde{p}_- + \lambda_n \, \tilde{d}_+^{-2} \, \tilde{a}^{\,2} + Hu^2 = 0. \tag{A.21}$$

Therefore, it follows that $\tilde{p}_-$ is the unique maximiser, over $\mathbb{R}^{n-1}$, of the left-hand side of (A.3) and (A.21). Then, N must be negative definite and Claim 4 establishes (4.16). We have

$$\max_{x \in \mathbb{R}^{n-1}} x^T N x + v^T x + \lambda_n \, \tilde{d}_+^{-2} \, \tilde{a}^{\,2} + Hu^2 = 0, \tag{A.22}$$

and from Claim 5,

$$\frac{\tilde{a}^{\,2}}{\tilde{d}^T \Lambda^{-1} \tilde{d}} + Hu^2 = 0. \tag{A.23}$$

Then (4.17) follows by solving for $\tilde{a}^{\,2}$ and recalling (4.15). This concludes the necessity proof.

Sufficiency. By assumption, $\tilde{d}^T \Lambda^{-1} \tilde{d} > 0$, which implies that $\tilde{d}_+ \neq 0$. Then, Claim 4 proves that $N < 0$ and Claim 5 shows that

$$\max_{x \in \mathbb{R}^{n-1}} x^T N x + v^T x + \lambda_n \, \tilde{d}_+^{-2} \, \tilde{a}^{\,2} = \frac{\tilde{a}^{\,2}}{\tilde{d}^T \Lambda^{-1} \tilde{d}} = -Hu^2, \tag{A.24}$$

where we have used (4.17) and (4.15). Therefore,

$$\max_{x \in \mathbb{R}^{n-1}} x^T N x + v^T x + \lambda_n \, \tilde{d}_+^{-2} \, \tilde{a}^{\,2} + Hu^2 = 0. \tag{A.25}$$

Since $N < 0$, (A.25) has only one maximiser, which we denote $\bar{x}$. The expression

$$x^T N x + v^T x + \lambda_n \, \tilde{d}_+^{-2} \, \tilde{a}^{\,2} + H u^2 \tag{A.26}$$

is negative for all $x \in \mathbb{R}^{n-1} \setminus \{\bar{x}\}$ and is equal to zero only when $x = \bar{x}$. Since $\tilde{d}_+ \neq 0$, we can always find $\tilde{p} \in \tilde{\mathcal{P}}$ such that $\tilde{p}_- = \bar{x}$ and also $\tilde{p}$ is unique [see (A.7)]. By Claim 1 then $\tilde{p} \notin \tilde{X}(u)$. For any $y \in \tilde{\mathcal{P}}$, $y \neq \tilde{p}$, expression (A.26) is negative for $x = y_-$ and Claim 1 proves that $y \in \tilde{X}(u)$. The proof of Theorem 4.7 is now complete.

Bibliography

P. Antsaklis and J. Baillieul, Guest Eds. Special issue on networked control systems. *IEEE Trans. on Automatic Control*, 49(9), 2004.

K. J. Åström and B. Wittenmark. *Computer controlled systems: Theory and design*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 1997.

J. Baillieul. Feedback designs in information-based control. In B. Pasik-Duncan, editor, *Stochastic Theory and Control*, number 280 in LNCIS, pages 35–57. Springer-Verlag, Berlin Heidelberg, 2002. Proceedings of the Workshop held at the University of Kansas on October 18-20.

J. E. Bertram. The effect of quantization in sampled-feedback systems. *Trans. Amer. Inst. Elec. Engrs.*, 77, pt. 2:177–182, 1958.

R. W. Brockett and D. Liberzon. Quantized feedback stabilization of linear systems. *IEEE Trans. on Automatic Control*, 45(7):1279–1289, 2000.

L. J. Corwin and R. H. Szczarba. *Calculus in Vector Spaces*. Monographs and textbooks in pure and applied mathematics ; 189. Marcel Dekker, Inc., New York-Basel-Hong Kong, 2nd edition, 1995.

R. E. Curry. *Estimation and Control with Quantized Measurements*. M. I. T. Press, Cambridge, Massachusetts, and London, England, 1970.

D. F. Delchamps. Stabilizing a linear system with quantized state feedback. *IEEE Trans. on Automatic Control*, 35:916–924, 1990.

N. Elia. Coarsest quantizer density for quadratic stabilization of two-input linear systems. In *Proc. of the 10th Mediterranean Conf. on Control and Automation - MED2002, Lisbon, Portugal*, 2002.

N. Elia and E. Frazzoli. Quantized stabilization of two-input linear systems: a lower bound on the minimal quantization density. In C. J. Tomlin and M. R. Greenstreet, editors, *Hybrid Systems and Control, HSCC*, number 2289 in Lecture Notes in Computer Science, pages 179–193. Springer: Berlin-Heidelberg, 2002.

N. Elia and S. K. Mitter. Stabilization of linear systems with limited information. *IEEE Trans. on Automatic Control*, 46(9):1384–1400, 2001.

J. A. Farrell and A. N. Michel. Estimates of asymptotic trajectory bounds in digital implementations of linear feedback control systems. *IEEE Trans. on Automatic Control*, 34(12):1319–1324, 1989.

I. Flügge-Lotz and C. F. Taylor. Synthesis of a nonlinear control system. *IRE Trans. on Automatic Control*, 1(1):3–9, 1956.

M. Fu and S. Hara. Quantized feedback control for sampled-data systems. In *16th IFAC World Congress, Prague, Czech Republic*, Prague, Czech Republic, 2005.

M. Fu and L. Xie. On control of linear systems using quantized feedback. In *Proc. American Control Conference, Denver, CO, USA*, pages 4567–4572, 2003.

M. Fu and L. Xie. The sector bound approach to quantized feedback control. *IEEE Trans. on Automatic Control*, 50(11):1698–1711, 2005.

G. C. Goodwin, H. Haimovich, D. E. Quevedo, and J. S. Welsh. A moving horizon approach to networked control system design. *IEEE Trans. on Automatic Control*, 49(9):1427–1445, 2004.

G. C. Goodwin, M. M. Seron, and J. A. De Doná. *Constrained Control and Estimation: An Optimisation Approach*. Springer-Verlag, London, 2005.

B. D. Green and L. E. Turner. New limit cycle bounds for digital filters. *IEEE Trans. on Circuits and Systems*, 35(4):365–374, 1988.

J. W. Grizzle, K. L. Dobbins, and J. A. Cook. Individual cylinder air-fuel ratio control with a single EGO sensor. *IEEE Trans. on Vehicular Technology*, 40(1):280–286, 1991.

H. Haimovich. Stabilizing static output feedback via coarsest quantizers. In *16th IFAC World Congress, Prague, Czech Republic*, 2005.

H. Haimovich and M. M. Seron. On infimum quantization density for multiple-input systems. In *Proc. 44th IEEE Conf. on Decision and Control, Seville, Spain*, pages 7692–7697, 2005.

H. Haimovich, G. C. Goodwin, and D. E. Quevedo. Moving horizon Monte Carlo state estimation for linear systems with output quantization. In *Proc. 42nd IEEE Conf. on Decision and Control, Maui, HI, USA*, pages 4859–4864, 2003a.

H. Haimovich, T. Perez, and G. C. Goodwin. On optimality and certainty equivalence in output feedback control of constrained uncertain linear systems. In *Proc. European Control Conference, Cambridge, UK*, 2003b.

H. Haimovich, M. M. Seron, G. C. Goodwin, and J. C. Agüero. A neural approximation to the explicit solution of constrained linear MPC. In *Proc. European Control Conference, Cambridge, UK*, 2003c.

H. Haimovich, G. C. Goodwin, and J. S. Welsh. Set-valued observers for constrained state estimation of discrete-time systems with quantized measurements. In *Proc. 5th Asian Control Conference, Melbourne, Australia*, pages 1947–1955, 2004.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

H. Ishii and T. Başar. Remote control of LTI systems over networks with state quantization. *Systems and Control Letters*, 54:15–31, 2005.

H. Ishii and B. A. Francis. Quadratic stabilization of sampled-data systems with quantization. *Automatica*, 39:1793–1800, 2003.

H. Ishii and B. A. Francis, editors. *Limited data rate in control systems with networks*. Number 275 in Lecture Notes in Control and Information Sciences. Springer-Verlag Berlin Heidelberg, 2002a.

H. Ishii and B. A. Francis. Stabilizing a linear system by switching control with dwell time. *IEEE Trans. on Automatic Control*, 47(12):1962–1973, 2002b.

H. Ishii, T. Başar, and R. Tempo. Randomized algorithms for quadratic stability of quantized sampled-data systems. *Automatica*, 40:839–846, 2004.

Z. P. Jiang and Y. Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6): 857–869, 2001.

R. E. Kalman. Nonlinear aspects of sampled-data control systems. In *Proceedings of the Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, USA*, pages 273–313, 1956.

C.-Y. Kao and S. R. Venkatesh. Stabilization of linear systems with limited information – multiple input case. In *Proc. American Control Conference*, pages 2406–2411, Anchorage, AK, USA, 2002.

H. Khalil. *Nonlinear Systems*. Prentice-Hall, New Jersey, 3rd edition, 2002.

E. Kofman. Non conservative ultimate bound estimation in LTI perturbed systems. *Automatica*, 41(10), 2005.

K. Li and J. Baillieul. Robust quantization for digital finite communication bandwidth (DFCB) control. *IEEE Trans. on Automatic Control*, 49(9):1573–1584, 2004.

D. Liberzon. Hybrid feedback stabilisation of systems with quantized signals. *Automatica*, 39:1543–1554, 2003a.

D. Liberzon. On stabilization of linear system with limited information. *IEEE Trans. on Automatic Control*, 48(2):304–307, 2003b.

D. Liberzon and J. P. Hespanha. Stabilization of nonlinear systems with limited information feedback. *IEEE Trans. on Automatic Control*, 50(6):910–915, 2005.

R. K. Miller, M. S. Mousa, and A. N. Michel. Quantization and overflow effects in digital implementations of linear dynamic controllers. *IEEE Trans. on Automatic Control*, 33(7):698–704, 1988.

R. K. Miller, A. N. Michel, and J. A. Farrell. Quantizer effects on steady-state error specifications of digital feedback control systems. *IEEE Trans. on Automatic Control*, 34(6):651–654, 1989.

G. N. Nair and R. J. Evans. Exponential stabilisability of finite-dimensional linear systems with limited data rates. *Automatica*, 39:585–593, 2003.

G. N. Nair and R. J. Evans. Stabilizability of stochastic linear systems with finite feedback data rates. *SIAM J. Control and Optimization*, 43(2):413–436, 2004.

T. Perez, H. Haimovich, and G. C. Goodwin. On optimal control of constrained linear systems with imperfect state information and stochastic disturbances. *International Journal of Robust and Nonlinear Control*, 14:379–393, 2004.

R. Raji. Smart networks for control. *IEEE Spectrum*, pages 49–55, June 1994.

J. B. Slaughter. Quantization errors in digital control systems. *IEEE Trans. on Automatic Control*, 9(1): 70–74, 1964.

E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Number 6 in Texts in applied mathematics. Springer-Verlag New York, 1998.

E. D. Sontag. Smooth stabilization implies coprime factorization. *IEEE Trans. on Automatic Control*, 34:435–443, 1989.

E. D. Sontag and Y. Wang. On characterizations of the input-to-state stability property. *Systems and Control Letters*, 24:351–359, 1995.

S. Tatikonda and N. Elia. Communication requirements for networked control. In S. Tarbouriech et al., editor, *Advances in Communication Control Networks*, number 308 in LNCIS, pages 303–326. Springer-Verlag, Berlin Heidelberg, 2004.

S. Tatikonda and S. Mitter. Control under communication constraints. *IEEE Trans. on Automatic Control*, 49(7):1056–1068, 2004a.

S. Tatikonda and S. Mitter. Control over noisy channels. *IEEE Trans. on Automatic Control*, 49(7): 1196–1201, 2004b.

G. C. Walsh and H. Ye. Scheduling of networked control systems. *Control Systems Magazine*, 21(1): 57–65, February 2001.

W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth constraints—II: stabilization with limited information feedback. *IEEE Trans. on Automatic Control*, 44(5):1049–1053, 1999.

S. Yakowitz and S. R. Parker. Computation of bounds for digital filter quantization errors. *IEEE Trans. on Circuit Theory*, CT-20(4):391–396, 1973.

W. Zhang, M. S. Branicky, and S. M. Phillips. Stability of networked control systems. *Control Systems Magazine*, 21(1):84–99, February 2001.