Scuola Normale Superiore di Pisa

Tesi di Perfezionamento
in
Matematica per le tecnologie industriali

Stabilization of quantized linear systems: analysis, synthesis, performance and complexity

candidato

Bruno Picasso

relatori

Prof. Antonio Bicchi *Università di Pisa*

Prof. Patrizio Colaneri *Politecnico di Milano*

Prof. Fabio Fagnani *Politecnico di Torino*

Luglio 2008

Scuola Normale Superiore di Pisa

PhD Thesis
in
Mathematics for the industrial technology

Stabilization of quantized linear systems: analysis, synthesis, performance and complexity

CANDIDATE

Bruno Picasso
picasso.bruno@gmail.com

SUPERVISORS

Prof. Antonio Bicchi *Università di Pisa*
Prof. Patrizio Colaneri *Politecnico di Milano*
Prof. Fabio Fagnani *Politecnico di Torino*

July 22, 2008

Contents

Chapter 1

Introduction

1.1 Problem formulation and motivations

Quantized control systems are controlled dynamical systems with input and/or output maps taking values in discrete (say, *quantized*) sets. As a simple reference model, consider a discrete time system of the type

$$\begin{cases} x(t+1) = f\big(x(t), u(t)\big) \\ y(t) = q_{\mathcal{Y}}\big(x(t)\big) \\ x(t) \in \mathbb{R}^n\,,\ u(t) \in \mathcal{U} \subset \mathbb{R}^m\,,\ q_{\mathcal{Y}}(x) \in \mathcal{Y}\,, \end{cases} \tag{1.1}$$

where $\mathcal{U}$ and/or $\mathcal{Y}$ are finite sets and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is such that, $\forall\, z \in \mathbb{R}^m\,,\ f(\,\cdot\,, z)$ is a smooth function. Depending on the problem, the control set $\mathcal{U}$ and/or the output map $q_{\mathcal{Y}}$ may be assigned or be a design objective. In this framework, a *controller* is a system processing sequences of outputs $y(t)\,, y(t-1)\,, \ldots$ and returning an input $u(t) \in \mathcal{U}$ for the system with the purpose that the resulting trajectories $x(t)\,, x(t+1)\,, \ldots$ have desired behaviors.

Quantized control systems belong to the wider class of *hybrid* systems. The hybrid nature of the model is caused by the coexistence of continuous state variables with discrete input and/or output variables. In particular, in a quantized control problem, the overall system results to be organized into two levels reflecting its mixed logical/dynamical nature (see Fig. 1.1) : at the physical level, the plant is modelled by an equation like (1.1) ; at the logical level, the controller is a device mapping output strings to input strings from discrete alphabets. The designer operates at the logical level. The overall picture results in a highly *nonlinear* dynamical system. Actually, even in the seemingly easy situation in which the dynamics is described by a linear transformation, that is

$$f(x, u) = Ax + Bu\,,$$

the presence of discrete variables produces nonlinear closed loop dynamics which may exhibit features such as the presence of multiple isolated equilibria, limit cycles and chaotic behaviors

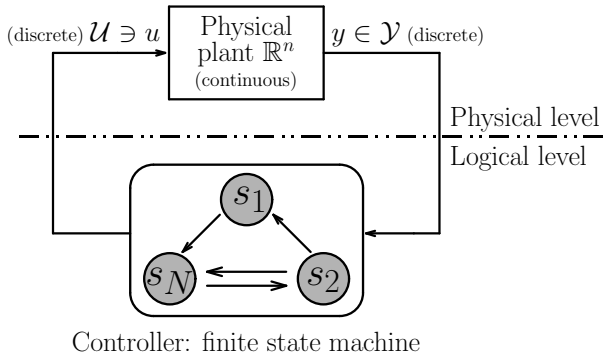

Figure 1.1: Graphical illustration of the hybrid structure of a quantized control system.

(see e.g., [63, 127, 29, 42]). This kind of models are referred to as *quantized linear systems* and are the subject of this thesis. From now on, we will tacitly refer to this class of systems.

The presence of quantization is traditionally believed to play adversely on control performance (if anything, it disrupts linearity of the system). However, its introduction has profound practical and technological motivations deeply rooted in the following two classes of problems: the "control with discrete sensors and/or actuators" and the "control under communication constraints".

Control with discrete sensors and/or actuators: digital controllers interact with the environment by means of Analog–to–Digital or Digital–to–Analog converters that have a finite resolution, which may even be very coarse. Think, for instance, of a stepper motor (where the allowed control actions are: "stand still", "one step forward" or "one step backward"; i.e., $\mathcal{U}$ is made up of three elements only) or a low resolution camera. More in general, think of a *low cost* sensing and actuation apparatus. In these cases, the input and/or output variables are inherently quantized. As a result, the set of control actions $\mathcal{U}$ in system (1.1) and the output map $q_{\mathcal{Y}}$ (modelling the sensors) are *fixed*. Prominent issues are in this case to sort out the achievable control goals and to synthesize a controller that maximizes performance.

Control under communication constraints: when a finite capacity communication link is present in the control loop, even if sensors and actuators have high resolution and can be reasonably modelled by means of continuous variables, the input and/or output signals have to be quantized and encoded into discrete–valued variables, suitable for the transmission (see Fig. 1.2). In a growing number of applications of modern technologies, one has to deal with large scale, possibly geographically distributed, complex and networked systems where multiple tasks are to be accomplished at the same time. Think, for instance, of the management of the control devices in a vehicle, of the remote control of a plant (being that a manufacturing process or a complex robotic device) or of the mobile telephony. In this kind of applications, the issues of control and of communication between the different components of the system cannot be separated. Indeed, although the overall communication capacity

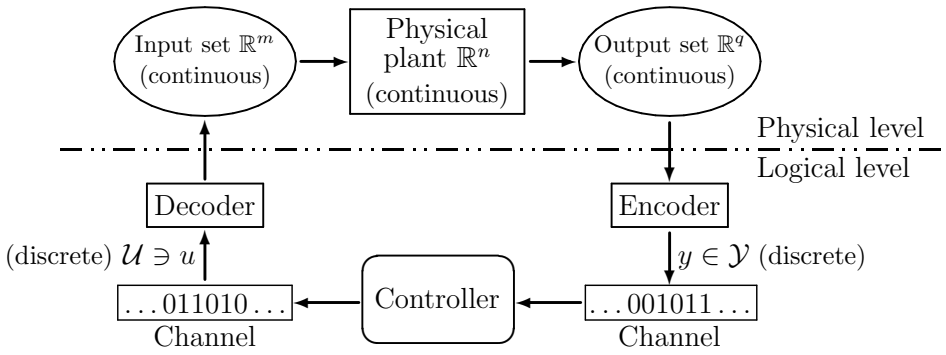

Figure 1.2: The basic scheme for the "control under communication constraints" problem.

may be high, due to the large amount of control tasks, only a small amount of it can be reserved for each subsystem. Not only this is the cause of performance deterioration, but even basic properties such as stability can be jeopardized. An efficient management of the limited communication resource requires that control and communication problem be jointly treated [131].

In terms of mathematical models, consider the control of a system whose input and output variables take values in continuous spaces (e.g., $u \in \mathbb{R}^m$ and $y \in \mathbb{R}^q$), but where the flow of information between the various components of the control loop is subject to communication constraints. Quantization has to be introduced so as to enable communication: thus, the quantized sets $\mathcal{U}$ and $\mathcal{Y}$, as well as the output map $q_{\mathcal{Y}}$, are *design* parameters chosen to accomplish the desired goals and to satisfy the communication constraint.

The two cases are often combined.

Since the beginning of the nineties, quantization has been studied under a new perspective. The underlying point of view has been that of regarding quantization as a useful tool rather than an undesirable phenomenon. For instance, in the control with discrete sensors and actuators, quantization is studied with a view to the possibility of attaining a control goal by means of the simplest and less expensive technology; in the control under communication constraints, quantization is introduced on purpose and the interest is for the coarsest quantization that preserve the possibility of achieving a desired control objective. The need for a deeper understanding of the role of quantization in systems dynamics has encouraged the different competencies of various scientific areas (such as Systems theory, Information and Communication theory as well as Theoretic Computer Science) to meet in a common mathematical background. Thus, it is not surprising that such a change of viewpoint has allowed to unveil undisclosed properties of quantized systems and to bring to light new opportunities offered by quantization.

Stability is the fundamental property required for a controlled system. Accordingly, stabilization is the master problem in control. As far as quantized linear systems are concerned,

the basic observation is that, if the control function $u(t)$ is constrained to take values from a finite set and the system is open loop unstable, then it is not possible either to achieve closed loop asymptotic stability or to confine the trajectories within arbitrarily small neighborhoods of the origin [29]. The first consequence of this fact is that the two examples presented above are quite different with regard to the stabilization problem.
In the control with discrete sensors and/or actuators, since the control set $\mathcal{U}$ is fixed and typically finite, then the problem has to be properly reformulated in terms of a weaker property than classical Lyapunov stability. For these models, *practical stability* notions are to be considered [29, 131, 45, 118]. Accordingly, the practical stabilization may consist in the synthesis of symbolic feedback controllers capable of steering the system to within sufficiently small neighborhoods Ω of the equilibrium, starting from large attraction basins X_0 (this property is referred to as (X_0, Ω)–stability). In this framework, the interesting issues to be studied are really varied and mutually dependent: beginning with, but not limited to, the characterization of the feasible pairs (X_0, Ω) such that (X_0, Ω)–stabilization is possible and the synthesis of a control law that accomplish the stabilization task.
In the control under communication constraints, instead, one can take advantage of the possibility of designing quantization to achieve asymptotic stabilization. This can be done by means of control policies where a finite set of control values is taken to be time–varying and adaptively chosen [17, 123]. In other words, the quantizer resolution is increased as the trajectories get close to the equilibrium but, at each stage, the control law takes values from a finite set (so that the communication constraint be satisfied). However, in order that stabilization can be achieved, at each stage it is necessary that a sufficiently large number of control values be available. The discovery of the minimal number of symbols to be transmitted and enabling the stabilization, as well as a stabilizing control policy (or, equivalently, the identification of a minimal value $R > 0$ for the capacity of the communication channel allowing for stabilization and a corresponding control and communication protocol), have been the most relevant contributions on the subject in the recent past [131, 5, 126, 89, 80].

Motivated by the emerging control applications and by the never–ending interest for large–scale systems, most of the efforts of the researchers in the control community have been addressed to the problem of the stabilization under communication constraints. In this thesis, instead, the main focus is on some mathematical issues stemming from the stabilization problem for systems with discrete sensors and/or actuators. In particular, the underlying assumption all along this work is that the input and/or output quantization is *assigned*. Although also some aspects concerned with the control under communication constraints will be considered, the preponderant role of quantization is that of representing a physical constraint on the system rather than a degree of freedom in design.
Let us briefly describe the main questions we deal with in this thesis. The (X_0, Ω)–stabilization is based on two main ingredients: a pair (X_0, Ω) of *controlled invariant* sets [11] (that is, sets within which it is possible to maintain confined the trajectories of the system) and a control law realizing convergence from X_0 to Ω. Thus, a preliminary step with respect to

the control synthesis consists in studying the controlled invariant neighborhoods of the equilibrium: this is referred to as the analysis stage. Because of the quantization constraint, both the invariance analysis and the control synthesis, which in turn is related with the reachability issue [2, 119, 10] , are quite involved problems and many questions are still open. Besides practical stabilization, also the study of closed loop performance is a central issue, as well as the study of the price one has to pay in terms of complexity of the quantization in order to achieve desired performance. More precisely, as far as closed loop performance are concerned, not only one is interested in the study of the *transient* behavior (which may be measured in terms of the speed of convergence towards the final set Ω) , but also the *steady–state* properties are a relevant parameter to be taken into account $\big($e.g., through some suitable notion of size for the final set Ω or through the amount of contraction $\text{Volume}\,(X_0)/\text{Volume}\,(\Omega)\,\big)$. As for the complexity of an (X_0, Ω)–stabilizing control law, it is meant as a measure modelling the cost to implement such a controller. According to the situation, complexity may be related with the sophistication of the sensing and control apparatus or, in a "control under communication constraints" framework, with the bandwidth required to transmit the control law. A feasible choice is that of considering a quantity related with the *entropy* of the controller (which in turn is related with the number of control values taken by the controller to accomplish the stabilization task) . In particular, the measure of complexity is depending on the geometric structure of the considered quantization. The problems we want to tackle consist in the analysis of the way closed loop performance vary with the complexity of the controller, in provide formal mathematical tools to quantify the existing trade off between performance and complexity, and in the identification of fundamental performance limits in quantized control.

To sum up, for a quantized linear system where $\mathcal{U} \subset \mathbb{R}^m$ is an *arbitrarily assigned* quantized input set (i.e, besides being quantized, there is not any further assumption on the structure of the assigned input set) and q_y is an arbitrarily assigned output map returning quantized measurements of the state of the system, we address the following questions and we propose innovative contributions (see next Section 1.3) :

1. **Analysis:** determine a family of controlled invariant sets. Among all controlled invariant sets, determine the *smallest* (in some proper sense) final neighborhood of the equilibrium Ω within which convergence of the trajectories can be ensured;

2. **Synthesis:** provide systematic tools for the design of controllers achieving practical stability properties;

3. **Performance and complexity:** introduce suitable measures of performance and complexity, analyze their mutual dependence and the limits on performance imposed by a fixed complexity.

To be precise, in the analysis and synthesis problems, the control set is always assumed to be quantized whereas the output of the system is either the full state x (absence of output quantization) or a quantized measurement of x . The quantized measurement case is divided

in two instances: state quantization (i.e., state space partitions made up of bounded sets) or output quantization (i.e., the output is of the type $y = q_y(x) = q_o(Cx)$, with $C \in \mathbb{R}^{q\times n}$ and q_o induces a partition of the output space $\mathbb{R}^q$ made up of bounded sets). In the study of performance and complexity, the focus is on state quantization.

1.2 Overview of the literature on quantized systems

Quantization has been studied ever since the late fifties (Kalman [63], Bertram [8]). Several years before the formal definition of chaos, Kalman showed the emergence of limit cycles and chaotic behaviors in dynamical systems under quantized control.
Sampled–data systems and signal processing have been the first field of interest for quantization: in this framework, there exists a considerable amount of literature on estimation problems and optimal control under quantized measurements (see the book of Curry [27] and references therein). In many of the pioneering works, quantization is modelled as an additive white noise perturbing a nominal system and techniques from Bayesian estimation and stochastic control are adopted. This assumption, however, is to be verified and results to be reasonable only if the quantizer has sufficiently high resolution. A remarkable exception is represented by the work of Curry, where a quantizer is modelled as a deterministic memoryless nonlinearity. Such an approach allows for a more qualitative study of the effects of quantization and keeps up with the novel interpretation, proposed by the author, of regarding a quantizer like a device providing limited information rather than a mere source of approximation. It is notable that the idea of modelling quantization as a deterministic nonlinearity is underlying many important contributions on the subject up to nowadays (e.g., Delchamps [29], Brockett and Liberzon [17], Fu and Xie [51]): this approach is taken also in this thesis.
Another interesting model for quantization is the set–membership description proposed by Schweppe [112] and by Bertsekas and Rhodes [9] to deal with state estimation problems under quantized measurements.
Digital implementation of control systems has been for long time the major field of interest in the framework of quantized systems (Moroney [87]). With regard to this problem, most of the efforts are spent to construct controllers (often quite involved) allowing for mitigation of the quantization effects. Ingenious analysis tools to study the effect of quantization on digital controlled systems are those based on norm–Lyapunov functions and presented by Michel et al. in [84, 85, 56] (the last of these references is concerned with nonlinear systems).

A renewed interest on quantization was brought by the seminal paper of Delchamps [29]. In that work, the idea of regarding quantization simply as a partial observation, as inspired by the work of Curry, is revived. Such an approach, and the tools provided by the ergodic theory [81, 71], has allowed the author to make a significant contribution towards the understanding of the role of quantization. Specifically, Delchamps has showed that asymptotic stabilization of an unstable plant is not possible in the presence of quantization; nevertheless,

if the plant is not too unstable, it is possible to control the state of the system as close as desired to the origin; if instead the condition of "mild" instability is not satisfied, controllability to the origin is not possible and chaotic behaviors arise. This work has given the first intuition on the existence of a critical rate of transmission, depending on the instability of the system, below which there is no control policy enabling for asymptotic stabilization of the system. For this reason, it has been the cornerstone for much of the subsequent developments in the studies on quantized control systems, especially in the framework of the control under communication constraints.

Further insight was brought by the work of Wong and Brockett [131], where the stabilization under communication constraints is studied by complementing the tools from the Systems theory with those from the Information theory. In that paper, the practical stability notion of *containability* is introduced and, by means of an information theoretic inequality (i.e., the Kraft inequality [26]), the first explicit necessary condition for stabilizability is carried out that relates the instability of the system with the available bit–rate for communication.

These results explicitly show that the presence of a communication constraint in the stabilization of a continuous time system has a profound impact. Indeed, if communication were possible at any rate, then stabilization would be always achieved by reducing the sampling period. In other words, it is possible to approximate any continuous time signal by switching very fast between discrete values (Raisch [109]). On the contrary, the presence of such a constraint induces a lower bound on the sampling period inhibiting this possibility.

The definition of a *dynamic* quantization scheme, where the range of the quantizer is time–varying and adaptively chosen so that the resolution is increased as the trajectories get close to the equilibrium (Brockett and Liberzon [17], Tatikonda [123]), has been one of the most significant methodology introduced in the recent literature (this approach is often referred to as the "zooming" technique and an embryonal version of it can be found in the paper of Sznaier and Sideris [120]). This policy is proved to guarantee asymptotic, rather than mere practical, stabilization. With reference to this technique, a precise mathematical formulation is given by Liberzon [75] to the idea that closed loop asymptotic stability is achievable provided that the expansion due to the instability of the open loop system can be outweighed by a sufficiently fine partition of the state (or output) space. When a finite capacity communication link is present in the control loop, this policy can be successfully implemented only if the capacity of the channel is sufficiently high. These ideas have been the basis to determine the necessary and sufficient condition on the rate of the channel that preserve asymptotic stabilizability (Tatikonda and Mitter [123, 126], Hespanha et al. [55], Nair and Evans [88]). For discrete time systems, the condition is

$$R > \sum_{|\lambda_i(A)|>1} \log_2 |\lambda_i(A)|, \tag{1.2}$$

where $\lambda_i(A)$ are the eigenvalues of the open loop system. A similar analysis, allowing for the possibility of packet losses, is offered by Ling and Lemmon [80]. In [89], Nair and Evans consider the case of linear systems under non–deterministic and unbounded noise:

the instability of the system is characterized in terms of the *entropy* rate of the open loop dynamics and the use of the entropy power inequality of Information theory [26] allowed the authors to show that inequality (1.2) is a necessary and sufficient condition also for mean square stabilizability. A thorough study of the dynamical properties of scalar systems operating near the data rate limit has been worked out by Baillieul in [5] and, in accordance with Delchamps' work, the emergence of chaotic behaviors is observed.
The crucial property ensuring that the zooming technique does indeed permit asymptotic stabilization is the input–to–state stability [62] of the ideal (i.e., non–quantized) underlying dynamics: this fact makes for the generalization of this technique to more general nonlinear systems (Liberzon and Hespanha [77] , De Persis [31]) . Moreover, it has been explicitly shown by Liberzon and Nĕsić [92, 78] that the zooming technique is also related with small–gain theory [67] : this has been the basis for extensions to input–to–state stabilization of quantized linear systems in the presence of an unknown disturbance (Liberzon and Nĕsić [79]) .

Also in the framework of *static* quantizers, small–gain theory offers profitable tools. A fundamental contribution and a point of reference for the scientific community is provided by the work of Elia and Mitter [39] : in this paper, the authors are interested in the problem of understanding the minimum information needed to carry out a stabilization task. It is shown that, for a given stabilizable discrete time linear system, the minimum density of a static single–input quantizer ensuring that stabilizability is preserved is achieved by a *logarithmic* quantizer. This result is in accordance with the intuition that, when the system state is far off from the equilibrium, a coarse control action can be used, whereas when the state is approaching the equilibrium, "fine–grained" corrections are needed. A logarithmic quantizer, when modelled as a static nonlinearity, can be seen as an uncertainty included in a sector [67] . As a counterpart, the nonlinearity representing the corresponding quantization error is a standard example of a finite gain nonlinearity. Actually, the results in [39] can be interpreted in terms of small–gain conditions in the functional space H_∞ . Thus, by taking advantage of robust control techniques, the results of Elia and Mitter has been extended by Fu and Xie [51] to more general cases such as systems with multiple inputs or with output quantization. Moreover, it is possible to include performance requirements such as guaranteed quadratic cost under quantized outputs or H_∞ performance under quantized inputs. Some of the ideas on logarithmic quantization are extended by Ishii and Francis to continuous time models in [60, 61] .
In the aforementioned papers, the range of the control law $u(t)$ is made up of infinite points accumulating towards 0 (as it follows by the increase of the quantizer resolution while the trajectories get close to the equilibrium) and asymptotic stabilization is possible. This is not the case if instead the controller is forced to take only a finite number of values: in this case, 0 is an isolated point for the range of the quantized control law and only *practical stability*, such as (X_0, Ω)–stability, can be ensured.
An issue related with the work of Elia and Mitter, but referred to finite range quantizers, is the one considered by Bullo and Liberzon [18] . In this paper, for a given continuous time linear system and a feedback law guaranteeing closed loop stability in the absence of quan-

tization, a static state quantizer taking a prefixed number of values has to be found so that a "destabilizing measure" is minimized (thus, ensuring good practical stability properties for the closed loop dynamics in the presence of quantization). It is shown that the design can be conveniently cast as a locational optimization problem [36] and different algorithms to solve it are discussed.
The emergence of chaos in linear systems under quantized control bears an unsuspected design opportunity for constructing efficient practically stabilizing controllers (Fagnani and Zampieri [42, 43]). The idea is as much brilliant as it is simple: if the dynamics within X_0 is chaotic, then almost all the trajectories will eventually enter *any* neighborhood Ω of the equilibrium; in addition, the quantized control law can be defined so as to ensure the invariance of Ω, thus trapping the trajectories therein. Therefore, with the effort needed to guarantee the invariance of X_0, also the convergence towards Ω is ensured. As one expects, the smaller Ω is, the larger the mean time taken by the trajectories to enter Ω is. More in general, there exists a trade off between *performance* of the closed loop system and *complexity* of the controller. Fagnani and Zampieri report a detailed study of this trade off in [45]. As already mentioned in the previous section, the analysis is enriched by the necessity of defining two performance parameters: one accounting for the transient behavior, the other for the steady–state. This yields a variety of optimal quantized controllers depending on the relative weight given to these parameters. This kind of analysis is extended by Delvenne [30] to quantizers with more complex topological structure.

Many results have been developed also for stochastic models. In [110, 111], Sahai and Mitter consider the feedback stabilization over a *noisy* channel. They show that the classical characterization of a communication channel in terms of the Shannon capacity is not suitable to deal with the stabilization of unstable processes. It is proved, instead, that the right notion is that of *anytime capacity*, a function of the so called anytime reliability (namely, the exponential rate at which the probability of decoding error of a message is guaranteed to decay with time when communication happens at rate R). By establishing the equivalence between feedback stabilization and reliable communication in the anytime sense, these works have contributed to bring further insight on the relations between control and communication. Borkar and Mitter [14] and Matveev and Savkin [82] have studied LQG control under limited data rate. The former reference, in particular, deals with separation properties between coding and control, an issue which is further analyzed by Tatikonda et al. in [125].
Other references on the control under communication constraints include papers dealing with the control of networked systems having a topological structure which is more complex than a single feedback loop (Tatikonda [124], Nair et al. [90], Matveev and Savkin [83], Li and Baillieul [73]) and the special issue of the *IEEE Transactions on Automatic Control* on "Networked control systems" [1]. A comprehensive presentation of the main results from the literature can be found in the work of Nair et al. [91].

Most of the papers listed above presume that the quantization better suited to the control objective can be freely chosen. Much more limited is the list of papers dealing with the

control of systems under assigned input and/or output quantization. Moreover, in most of the contributions only uniform quantization is considered. For instance, the reachability of a neighborhood of 0 under uniform input or state quantization is studied by Sznaier et al. [119, 120] by means of the Minkowsky functional [12] . The work of Delchamps [29] deals with uniform state quantization and practical stability properties are studied in terms of standard quadratic Lyapunov functions. In addition to uniform quantizers, analysis results are provided by Michel et al. in [85] for logarithmic quantizers. Also the recent paper of Azuma and Sugie [4] deals with assigned uniformly quantized inputs, but dynamic quantization is proposed: namely, a dynamic quantizer is designed so as to minimize the norm of the difference between the given linear plant and the cascade interconnection of the quantizer with the given plant. The study is restricted to open loop behavior.

Different approaches are those based on optimal control techniques: in this case, the control law is the result of an optimization problem. Among these methods, a suitable framework is provided by the so called Model Predictive Control techniques (MPC) , for their ability to effectively deal with constraints in the control action. A detailed study of the application of MPC to quantized control systems can be found in the work of Quevedo et al. [108] . In this paper, the practically stabilizing control law is provided in closed form and analysis methods are included to determine the invariant sets for the corresponding closed loop dynamics. An extension of MPC based techniques for networked control systems is proposed by Goodwin et al. in [54] . Here, assigned finite input sets and quantized measurements are considered: the state estimation problem is studied either in terms of a set–valued observer or, for non–deterministic plants, via Kalman filtering; the control synthesis is then carried out with a receding horizon policy. Closed loop analysis is offered only for open loop stable plant.
The so called *dual mode* MPC scheme is considered by Picasso et al. in [98] : in this case, convergence to within a priori determined invariant sets can be enforced. For uniformly quantized single–input systems, Picasso and Bicchi present the corresponding feasibility analysis of the quantized MPC scheme in [101] .
In the contribution proposed by Su et al. [118] , the design of practically stabilizing controllers is converted into a nonlinear programming problem: such an approach can be applied to a wide class of hybrid systems including quantized input models as a particular case.

A synthesis of the main results on the stabilization of quantized systems, including technical details, can be also found in the contribution of Picasso et al. [106] .

1.3 Summary of the thesis and main contributions

1.3.1 Summary of the thesis

We focus on the practical stabilization problem for discrete time linear systems under assigned input and/or output quantization. The study has been organized into three steps: *analysis*, *synthesis* and study of the relations between *performance* of the closed loop system and *complexity* of the controller.

In **Chapter 2**, besides the main definitions, the most relevant issues addressed in the thesis are presented through simple scalar examples.
The analysis stage is presented in **Chapter 3**. This study is prior to the control synthesis and, in agreement with the final goal of (X_0, Ω)–stabilization, it is concerned with the search for controlled invariant neighborhoods of the equilibrium. Two methods are proposed: the first one provides controlled invariant hypercubes. This technique can be applied to reachable single–input systems only but it has optimality properties (i.e., it is capable of returning a final set Ω of *minimal* size, see Section 3.1.2) and it is suitable to handle the cases of systems under input and output quantization. In the second method, controlled invariant ellipsoids are considered which result from classic Lyapunov arguments for quadratic functions: this approach is feasible for any stabilizable system but results are often quite conservative and limited to input quantization. In Section 3.1.3, an extension is proposed to the controlled invariance analysis for multiple scalar systems sharing a limited capacity communication channel: in this case, the issues of the control under assigned quantization are combined with those from the control under communication constraints.
The control synthesis problem is faced in Chapters 4, 5 and 6.
It is shown in Chapter 3, that a quantized version of the classical deadbeat controller (the so called *qdb–controller*) is directly involved in the invariance analysis of hypercubes. Hence, in **Chapter 4**, (X_0, Ω)–stability of the closed loop dynamics generated by a qdb–controller is studied in terms of pairs (X_0, Ω) made up of hypercubes. This analysis is developed for various cases including systems with arbitrarily assigned single–input and single–output quantization (in this most general case, the study is performed in terms of the size of hypercubes bounding the initial condition, the state transient and the steady–state evolution, which is the so called (X_0, X_1, Ω)–stability).
Small–gain theorems [67] are the technical tool allowing us to deal with the general class of multi–input systems. The small–gain approach for the design of practically stabilizing control laws is presented in Chapters 5 and 6. It is an abstract methodology enabling to solve the control synthesis problem in a systematic way. In this framework, the system and the feedback controller are seen as input/output operators. Small–gain theorems provide conditions for the closed loop stability in terms of the norms of these operators. In this way, the control synthesis problem is converted into the satisfaction of suitable relations between the norms of the system and of the controller. A more detailed introduction to this approach is given in Section 5.1. In this part of the thesis, systems under arbitrarily assigned multi–input quantization are considered, while full state is assumed to be available (in particular, measurements are not quantized). Two small–gain approaches are fitted on quantized input systems so as to guarantee closed loop practical stability.
The first approach is presented in **Chapter 5** and consists of considering the input/output operators as elements of the so called Hardy's functional space H_∞. In this way, the control synthesis for practical stabilization is transformed into a particular control problem in H_∞ [132, 117]. The corresponding practical stability analysis is based on the Lyapunov arguments for quadratic functions illustrated in Chapter 3 and yields pairs (X_0, Ω) of invariant

ellipsoids.
The other method is proposed in **Chapter 6**: here, the input/output operators are associated with elements of the functional space ℓ_1, hence the control synthesis for practical stabilization is turned into a control problem in ℓ_1 [128, 34]. The corresponding practical stability analysis provides hypercubes within which the trajectories are proved to be confined. These results extend to multi–input systems those illustrated in Chapter 4 for the qdb–controller, which indeed are shown to be achievable as a particular case of the ℓ_1 control technique. For single–input systems, the control synthesis problem is entirely solved by means of the ℓ_1 approach whereas, in the multi–input case, the proposed methodology is complementary to the one relying on the H_∞ theory. Namely, the control synthesis is performed in the H_∞ framework, whilst a small–gain theorem in ℓ_1 enables us to obtain a less conservative steady–state analysis of the closed loop dynamics than that based on invariant ellipsoids provided in Chapter 5. Furthermore, a formulation of the control synthesis problem for practical stabilization is also proposed in terms of a mixed H_∞/ℓ_1 control problem. As far as the size of the final set Ω is concerned, the control synthesis in ℓ_1, and the corresponding closed loop analysis, appears to inherit the optimality properties of the analysis based on invariant hypercubes (in this respect, only numerical examples are given).
Finally, the analysis of the relations between the closed loop performance and the complexity of the controller is carried out in **Chapter 7**. The study is performed with reference to the coding issue rising in the "control under communication constraints" framework (thus, the complexity of the controller is related to the bandwidth needed for transmission of the control symbols). To this end, it is convenient to analyze the dynamics of the system in a probabilistic fashion. Namely, the state of the system at time t is represented by a probability distribution μ_t: the focus is on the evolution of the second moment of the distribution (i.e., the mean–square value, or energy, of the distribution). An entropy–like function of the energy is defined to measure the complexity of the state quantization induced by the controller. The asymptotic and the transient behavior of the system are described in terms of the asymptotic value of the energy of the evolving distributions and of the convergence rate towards the asymptotic value, respectively. The behaviors of the considered quantities for various types of quantizers are analyzed, including (but not limited to) the main quantizers considered in the previous chapters of the thesis. Moreover, fundamental relations between complexity and steady–state performance and between complexity and transient behavior are worked out by means of the mathematical tools offered by the Information theory. The study is restricted to the case of scalar linear systems, which already contains the basic difficulties one encounters for more complex systems.

A general picture of the thesis organization, including references to the various types of quantizations taken into consideration and to the relations between the different addressed topics, is given in Fig. 1.3.

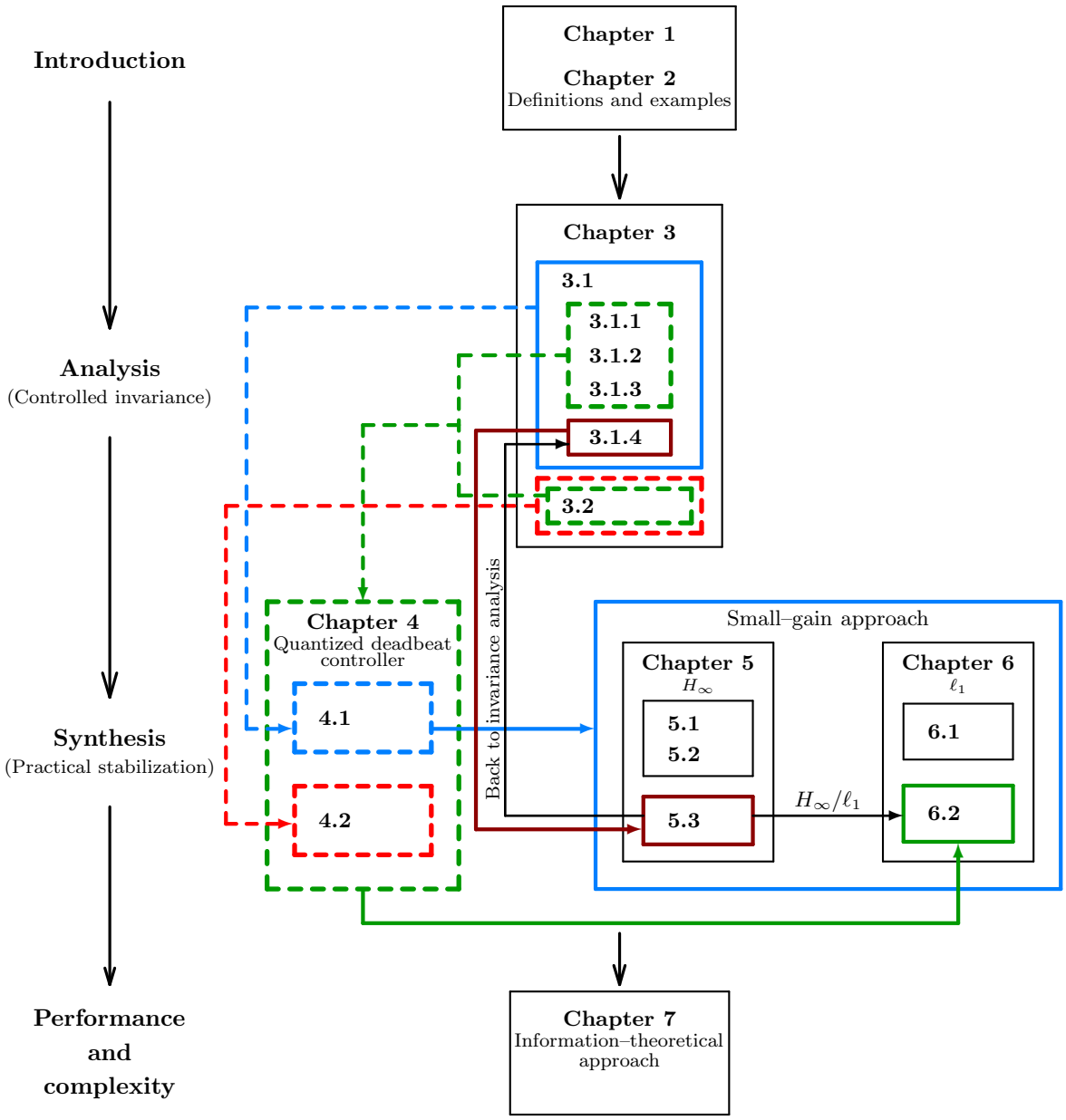

Figure 1.3: Thesis organization and reading paths. **Blue lines:** quantized input; **Red lines:** quantized input and quantized measurement; **Green lines:** hypercubes, **Brown lines:** ellipsoids; **Black arrows:** main logical relations; **Black arrows** with labels**:** other interrelations; Dashed lines: single–input; Full lines: multi–input.

1.3.2 Main contributions

A list of the main contributions of the thesis is the following:

1. The practical stabilization problem for discrete time linear systems under *arbitrarily assigned* input and/or output quantization has not received sufficient attention by the control community. For this problem, systematic analysis and control synthesis tools are offered. The cases where only practical stabilization is possible (e.g., when the quantizers are static and have a finite range) are considered.

2. For single–input systems, an original method is proposed to determine hypercubic controlled invariant sets. The approach is completely analytic (rather than algorithmic) and it is capable of dealing with systems under arbitrarily assigned quantization. One relevant contribution is the possibility of handling also output quantization. Moreover, it is proved that the considered family of invariant sets contains an element whose size is minimal with respect to *any* controlled invariant set. Convergence to such a set is ensured by a qdb–controller, whose practically stabilizing properties are studied in terms of invariant hypercubes. (See Chapters 3 and 4) .

3. A quantized version of a control law which is stabilizing in the absence of quantization $\big($as, for instance, $u(x) = q_{\mathcal{U}}(Kx)$, for some $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ and a *stabilizing* control gain $K \in \mathbb{R}^{m\times n}\big)$ is not guaranteed to achieve practical stabilization. The idea of resorting to robust control techniques to determine feasible choices for the non–quantized control law is not new, but a thorough study of the small–gain approach for the control synthesis of practically stabilizing quantized controllers under assigned multi–input quantization indeed is. In particular, the method based on the ℓ_1 theory is quite new and especially indicated to deal with practical stabilization. (See Chapters 5 and 6) .

4. A parallel issue, where quantization is not directly involved, has been the development of innovative tools for the analysis and the control synthesis in ℓ_1 . In this respect, interesting results have been worked out for the special class of *positive* systems. (See Section 6.1) .

5. An analysis of performance and complexity for static quantizers capable of dealing with distributions having unbounded support has not been previously considered in the literature and poses interesting theoretical issues. A powerful relation between the complexity of the controller (as measured by an entropy–like function) and the geometric structure of the corresponding quantization is worked out in terms of the Laplace transformation. Moreover, fundamental limits in performance, enforced by the given complexity of a controller, have been carried out by means of information theoretical inequalities relating the entropy of the control process to the behavior of the energy of the evolving distributions. (See Chapter 7) .

Let us discuss the items of the above list with detailed references to the existing literature.

1. As we have clarified in previous Section 1.2 , the control under communication constraints have been the predominant subject of the recent literature on quantized control systems. There are two main consequences of this fact: first, recently developed control tools (as for instance the robust control techniques in ℓ_1) have not been extensively applied to the practical stabilization problem under assigned quantization. Secondly, most of the methods introduced in the literature are tailored to the asymptotic stabilization and dynamic quantizers: for this reason, with few remarkable exceptions [39, 61] , they overlook the practical stabilization issue and, above all, they rarely provide the tools to solve the problem under arbitrarily assigned quantization (i.e, when no special assumptions are made on the structure of the assigned quantized sets) .

2. Although there exist a wide literature on controlled invariance (see [11] and references therein) , the problem for quantized input systems is not trivial. In fact, not only quantization disrupts linearity of a system, but also other properties, such as convexity of the control set, are lost. This fact inhibits the use of many classical results on controlled invariance under constrained control.
   The perturbation technique proposed in [108] contains some ideas similar to those presented in this thesis but it only allows to determine invariant sets for a specified closed loop dynamics and it is often computationally untractable. In [19] a method is proposed which employs geometric arguments and allows one to deal with multi–input systems and robustness issues. However, only quantization resulting from the uniform tiling of the state space can be handled. A quite general framework within which invariant sets can be studied, and allowing one to deal with quantization, is provided by the Viability theory [3] . We mention, inter alia, the "Controlled Invariance Kernel Algorithm" [113]. The main drawback of this technique is that it is computationally intensive because of the increasingly larger number of constraints needed to describe the sets generated at the various stages of the algorithm. Another interesting approach is the one proposed in [118] for switching systems (thus including quantized systems as a special case) where invariant Euclidean balls are algorithmically computed using nonlinear programming. This technique can efficiently handle the two stages of practical stabilization (i.e., invariance and convergence) but it has to cope with a non–convex optimization, hence the global optimum may not be found. Furthermore, in many cases, looking for invariant Euclidean balls is too restrictive as a quantized input system may have invariant sets but no invariant balls. Among classical methods for the analysis of controlled invariance, those based on Lyapunov theory and quadratic functions are the most easily adaptable to quantized systems. Although not always explicitly referred to the controlled invariance problem, this approach is common in the literature (see, for instance, [29, 17, 39]) .
   While all these techniques are of quite general application, on the other hand they typically yield conservative results. That is, they are not capable of providing information on the *minimal* invariant set for a system under assigned input quantization.

Moreover, the nature of the practical stabilization problem varies with the different types of output map $q_{\mathcal{Y}}$ considered. Results in the literature are mainly concerned with quantization of the state [29, 17, 80] or quantization of the innovation [39, 126]. Results for quantized outputs are given in [17, 51] but a detailed study on the synthesis of symbolic controllers for discrete time systems with arbitrarily assigned quantized outputs has not been addressed so far (to the best of our knowledge).

3. Robust control techniques have been already used to study the stabilization problem for quantized systems. The analysis techniques proposed in [84, 85] and employing norm–Lyapunov functions are similar to those based on quadratic functions illustrated in this thesis. However, they are not suitable to deal with control synthesis and are limited to SISO systems. In [61], sample–data systems and uniform or logarithmic quantizers are considered. The small–gain approach in H_∞ proposed in this thesis, instead, is closely related to the sector bound approach proposed in [51]. In the work of Fu and Xie, the H_∞ control tools are employed to design the coarsest quantizers preserving asymptotic stabilizability. In this way, logarithmic quantizers are considered for MIMO systems. The study does not involve quantitative analysis of practical stability and is restricted to particular types of quantizers (namely, logarithmic quantizers acting independently on each component of the multi–input or multi–output variable, thus giving rise to a quantized set in the form of a cartesian product). On the contrary, the formulation in terms of small–gain conditions proposed in this thesis can efficiently handle any arbitrarily assigned multi–input quantized set.
The functional space H_∞ is isomorphic to the space of the operators between input and output sequences belonging to ℓ_2, hence vanishing for $t \to +\infty$. It is deeply–rooted in this fact the reason why the H_∞ control tools are indicated for those cases where asymptotic stabilization is achievable (i.e., when the range of the control function $u(t)$ is allowed to have 0 as an accumulation point so that the quantization error is vanishing for $t \to +\infty$). This is not the case when, for instance, the control function is forced to take a finite number of control values. In this situation, the quantization error is a persistent (i.e., non–vanishing) disturbance which can be seen as a sequence in ℓ_∞. Therefore, the choice of studying the practical stabilization in the ℓ_1 functional space (whose elements define operators between input and output sequences belonging to ℓ_∞) is natural and it is not surprising that, as for the steady–state performance, the corresponding results appear to be less conservative than those based on H_∞ theory.
Although we are mainly interested on the stabilization under assigned quantization, the proposed techniques can be applied also to *design* the quantizers so as to achieve desired stability properties. In this case, the results based on ℓ_1 control turn out to be conservative in terms of bit–rate with respect to the tight bound proved in [126, 89, 80]. This is however inherently related with the faster convergence rate the proposed control strategy ensures (see Remark 16 in Section 4.2.2).

4. The design of a controller so that the ℓ_∞–gain[1] of the operator associated to the corresponding closed loop dynamics is below a desired threshold, or minimized, is called the ℓ_1–control problem. This problem has been introduced in [128] and has been the subject of a certain amount of literature [28, 34, 113, 65, 38, 7] . The main proposed approaches take advantage of the convex structure of the set of all stabilizing controllers and, either the problem is transformed into an infinite dimensional linear optimization, or a linear (or quadratic) programming formulation is presented. Hence, algorithmic procedures are carried out for numerical approximation of the solution. Differently from the case of dynamic controllers, the problem of minimizing the ℓ_∞–gain by means of static output feedback has been less investigated. In this case, the main difficulties rise from the fact that the set of stabilizing control gains is not convex.
   Assuming that a good synthesis methodology may be found if suitable analysis tools are available, we have first turned our attention to the problem of evaluating the ℓ_∞–gain of a *BIBO–stable* linear system. In this respect, the main results [6, 58] are still based on algorithmic procedures that do not appear to be practical for extension to control synthesis problems. The main contribution of Section 6.1 consists in providing an easy method for the computation of an upper bound for the ℓ_∞–gain of a BIBO–stable linear system. Although the proposed bound is not always feasible (i.e., it can be computed only for some particular systems) and often quite conservative, yet it is useful in some interesting cases. In particular, the bound is proved to be tight for single–input *positive* systems. Furthermore, the proposed method can be extended to deal with control synthesis: a sufficient criterion is provided that allows one to find a static output feedback $u = Ky$ so that the ℓ_∞–gain of the closed loop dynamics is below a desired threshold. This can be done by solving a system of linear inequalities.

5. The presence of a communication constraint in the control loop is the cause of performance deterioration mainly due to quantization and transmission delays. The suitable mathematical framework within which these phenomena can be analyzed is offered by the Information theory. This approach has been pursued ever since [131] and later on by [89, 44, 125, 30] . Motivated by the "control under communication constraints" problem, the analysis of the relations between closed loop performance and complexity of the controller has been first offered in [42, 45] . An information theoretical oriented study is given in [44, 30] . These works deal either with distributions having a bounded support or with dynamic quantizers. As clarified in [89] , to cope with distributions having an unbounded support, control laws taking infinite values are necessary. Because we consider static controllers (as opposed to time–varying techniques widely studied in the literature [123, 17, 125, 89, 80]) , infinite symbols are necessary to encode the control values. This poses technical problems concerned with transmission delays caused by the presence of arbitrarily long coding sequences, as well as theoretical questions on the definition of a proper complexity measure for a controller (indeed, a cardinality

[1] That is, the induced norm of an operator between sequences belonging to ℓ_∞ .

function as proposed in [45] is no more meaningful) . Not only the notion of entropy in Information theory is enabling us to deal with these problems, but it provides us with powerful technical aids for the analysis of the relations between performance and complexity. For instance, recently studied information theoretic relations between the entropy of continuous and discrete variables [46] allow us to determine a lower bound on the achievable asymptotic value of the energy under assigned quantization.

Further discussions can be found in the introductory parts of the various sections.

References [97, 100, 102, 48, 103, 105, 104, 106, 107] are a selection of the main publications where earlier versions of the contributions of this thesis can be found. The writing of other papers, on the most recently developed parts of this work, is in progress.

1.4 Acknowledgements

Many people have contributed in different ways to carry out this thesis and I am thankful to all of them (also to those I may forget to mention in this section) .
The main gratitude is directed to my advisors: Proff. Antonio Bicchi, Patrizio Colaneri and Fabio Fagnani. To work with such a team of professors has been an extremely valuable opportunity. Thanks to their different skills, they have offered me a great chance to gain a varied education. It is not of minor importance to emphasize that they all are really nice people to work with: this fact has had a positive influence on the realization of this work. Let me explicitly thank Prof. Antonio Bicchi for his ability in coordinating my PhD program which has been developed in a geographically distributed context.
I would like to acknowledge the institutions that gave me hospitality during the years of my PhD. First, the "Scuola Normale Superiore di Pisa", which offered me the best means to work out my research project. The research center "Centro Piaggio" of the "University of Pisa", where I found an exciting interdisciplinary group of young researchers. The "Politecnico di Milano", where I spent more than three years in touch with a keen scientific community. The three institutions are also gratefully acknowledged for their effective financial support.
During my PhD program, I have also had the chance to visit two foreign institutions. In May 2004 I have been visiting the "Royal Institute of Technology (KTH)" in Stockholm thanks to the invitation of Prof. K.H. Johansson. Parts of the contents of Section 3.1.3 are the result of this collaboration. At the KTH, I also met Alberto Speranzon, a really nice fellow with whom I had stimulating conversations.
During the fall semester 2006 I have been a visiting scholar at the "Massachusetts Institute of Technology (MIT)" in Cambridge. It has been a privilege for me to spend a period in such a prestigious institution, for this reason I am really grateful to Proff. S.K. Mitter and M.A. Dahleh. Prof. Dahleh gave me useful suggestions about the ℓ_1 control. With Prof. Mitter I had inspiring discussions that have guided me in the writing of some parts of this document. I also would like to acknowledge him for favoring the contacts with the members of his research team: the daily interactions with these people have been a precious source of learning. A

special mention is reserved to Mukul Agarwal who, besides telling me many things about Information theory, has also been a helpful presence during my sojourn in Boston.
The contents of Section 3.1.3 arise from an idea of Prof. Luigi Palopoli: to work with him has been a profitable chance to discuss many aspects of quantized control, for this reason the influences of this collaboration are not limited to that section.
Another source of inspiration has been Prof. Sandro Zampieri. I have to acknowledge him for suggesting to study the example reported in Section 7.5 . More in general, any talk with him has been the occasion to learn something about control or about the right way to approach research.
Some parts of this thesis are currently taught within control classes at the Politecnico di Milano: Proff. Marco Lovera and Patrizio Colaneri are acknowledged for the confidence placed in me.
I would like to show gratitude to my officemates in Politecnico di Milano: a group of highly qualified and friendly people which has contributed to create a pleasant and enthusiastic working atmosphere. Let me explicitly mention Marcello Farina and Mara Tanelli from whom I heard a lot about theoretic and applied control theory, and Annalisa Zappavigna who read this thesis since its draft versions and gave me suggestions to improve the quality of the presentation.
A special thank to my father who has collaborated to the realization of all the figures included in this thesis.
The patience and the support of my wife, Anna, has been an incomparable help during the years of the PhD and during the writing of this document. This thesis is dedicated to her and to our family.

1.5 Notation and terminology

Matrices and vectors:

Consider a given square matrix $A \in \mathbb{R}^{n\times n}$ and a vector $x \in \mathbb{R}^n$.

- Let

$$\begin{cases} \lambda(A) & \text{be an eigenvalue of } A \\ \mathcal{S}(A) := \{\lambda(A) \in \mathbb{C} \,|\, \lambda(A) \text{ be an eigenvalue of } A\} & \text{be the } \textit{spectrum} \text{ of } A \\ \rho(A) := \max\limits_{\lambda(A)\in\mathcal{S}(A)} |\lambda(A)| & \text{be the } \textit{spectral radius} \text{ of } A\,. \end{cases}$$

  If $\rho(A) < 1$, then the matrix A is said to be *Schur*.

- x' and A' stand for the transpose of the vector x and of the matrix A , respectively.

- e_i is the i–th vector of the canonical basis.

- $x_i := e_i'x$ is the i–th component of the vector x .
  The i–th component of the vector Ax is denoted by $(Ax)_i$.

- $A_{i,j} := e_i' x e_j$ is the (i,j)–th entry of the matrix A.
- I_h is the identity matrix in $\mathbb{R}^{h\times h}$. Sometimes, when the dimension is clear from the context, the subscript "h" is omitted.
- For $\Omega \subseteq \mathbb{R}^n$, $A\,\Omega := \{A\omega \,|\, \omega \in \Omega\}$.

Systems:

- A discrete time, time–invariant and strictly proper dynamical linear system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \\ x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \quad \text{and} \quad y \in \mathbb{R}^q \end{cases}$$

  is denoted by $\Sigma(A,B,C)$. In order to avoid useless redundancies, the matrices B and C are always supposed to be full rank.
- Either $x^+(t)$ or simply x^+ are often used to denote $x(t+1)$.

Norms:

- For $v \in \mathbb{R}^n$ and $p \in [1\,,\,\infty]$, we let:

$$\|v\|_p := \begin{cases} \left(\sum_{i=1}^n |v_i|^p\right)^{1/p} & \text{if} \quad p \in [1\,,\,\infty[\\ \max\limits_{i=1,\ldots,n} |v_i| & \text{if} \quad p = \infty\,. \end{cases}$$

- For $A \in \mathbb{R}^{n\times n}$ and $p \in [1\,,\,\infty]$, we let:

$$\|A\|_p := \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}\,.$$

  It holds that

$$\|A\|_p = \begin{cases} \sqrt{\rho(A'A)} = \sqrt{\rho(AA')} & \text{if} \quad p = 2 \\ \max\limits_{i=1,\ldots,n} \sum_{j=1}^n |A_{i,j}| & \text{if} \quad p = \infty\,. \end{cases}$$

Sets:

- The cardinality of a set S is denoted by $\#\,S$.
- A set S is said to be *finite* iff $\#\,S < +\infty$, whereas, it is said to be *countable* iff there exists an injective function $f : S \to \mathbb{N}$. Hence, finite sets are countable sets.
- Intervals are denoted by: $[a\,,\,b] := \{x \in \mathbb{R} \,|\, a \le x \le b\}$, $]a\,,\,b[:= \{x \in \mathbb{R} \,|\, a < x < b\}$, $[a\,,\,b[:= \{x \in \mathbb{R} \,|\, a \le x < b\}$, etc.
- Let $\mathbb{R}^+ := \mathbb{R} \cap [0\,,\,+\infty[$.

- For a given $\epsilon > 0$, let $\epsilon\mathbb{Z} := \{\epsilon z \,|\, z \in \mathbb{Z}\}$.
- Let $\mathcal{B}_R := \{x \in \mathbb{R}^n \,|\, \|x\|_2 \leq R\}$.
- For a given $\mathbb{R}^{n\times n} \ni P > 0$ and $r \in \mathbb{R}$, let
$$\mathcal{E}_{P,r^2} := \left\{x \in \mathbb{R}^n \,|\, x'Px \leq r^2\right\}.$$
- Let $E \subseteq \mathbb{R}^n$:
$\overline{E}$ denotes the closure of E;
$\chi_E(x)$ is the characteristic function of E, namely:
$$\chi_E(x) = \begin{cases} 1 & \text{if} \quad x \in E \\ 0 & \text{otherwise}; \end{cases}$$
${}^cE$ denotes the complementary set;
for $v \in \mathbb{R}^n$, let $E - v := \{x \in \mathbb{R}^k \,|\, x + v \in E\}$.
- Given two sets D and E, let $E \setminus D := \{e \in E \,|\, e \notin D\}$.
- For $E \subseteq \mathbb{R}$, $\operatorname{diam}(E) := \sup_{x,y\in E} |x - y|$ is the diameter of E.
- For $\Omega \subseteq \mathbb{R}^n$, let $\operatorname{Pr}_i \Omega := \{\omega_i \in \mathbb{R} \,|\, \exists\, \omega = (\omega_1, \ldots, \omega_i, \ldots, \omega_n) \in \Omega\} = \{\omega_i \,|\, \omega \in \Omega\}$;
$\operatorname{diam}_i \Omega := \operatorname{diam}\big(\operatorname{Pr}_i \Omega\big)$.
- Let
$$Q_n(\Delta) := \left[-\tfrac{\Delta}{2}\,,\, \tfrac{\Delta}{2}\right]^n = \left\{x \in \mathbb{R}^n \,|\, \|x\|_\infty \leq \tfrac{\Delta}{2}\right\}$$
be the closed hypercube of edge length Δ;
denote by $Q_n^o(\Delta) := \left]-\tfrac{\Delta}{2}\,,\, \tfrac{\Delta}{2}\right[^n$ the semi–open hypercube.

Probability:

- Probability spaces are simply denoted by the sample space Ω, both the σ–algebra and the probability measure are understood (to our purposes, it is enough to specify the distribution of the considered random variables).
- We denote by $\mathcal{P}r(\mathbb{R})$ the space of Borel probability measures on $\mathbb{R}$.
- For $m \in]0\,,\, 1]$ and $x \in \mathbb{R}$, $m\delta_x$ denotes the delta–measure having mass m concentrated at $x \in \mathbb{R}$. That is, if $A \subseteq \mathbb{R}$ is a Borel set, then
$$m\delta_x(A) = \begin{cases} m & \text{if} \quad x \in A \\ 0 & \text{otherwise}. \end{cases}$$
- Let X be a random variable taking values in $\mathbb{R}$: by $\mathbb{E}[X]$ and $\mathbb{V}\text{ar}[X]$ we mean the expectation and the variance of X.

- We denote by $\mathcal{P}r(\mathbb{Z})$ the space of discrete probability measures on $\mathbb{Z}$: its elements will be identified with sequences $\mathbf{p} = \{p_k\}_{k\in\mathbb{Z}}$ of non–negative real numbers such that $\sum_{k\in\mathbb{Z}} p_k = 1$.

Other notation:

- If not otherwise stated, the logarithms are in the base e.
- The floor function is $\lfloor x \rfloor := \max\{z \in \mathbb{Z} \,|\, z \le x\}$;
  the ceil function is $\lceil x \rceil := \min\{z \in \mathbb{Z} \,|\, z \ge x\}$.
- The Gamma function is defined by

$$\begin{cases} \Gamma(x) := \int_0^{+\infty} e^{-t} t^{x-1} dt \\ x > 0\,. \end{cases}$$

  For $n \in \mathbb{N}$, it holds that $\Gamma(n+1) = n!$.

- The sign function is defined by

$$\text{sign}\,(x) := \begin{cases} \frac{|x|}{x} & \text{if } x \neq 0 \\ 1 & \text{otherwise}\,. \end{cases}$$

- Let $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{Z}$:
  for $Y \subseteq \mathcal{Y}$, $f^{-1}(Y) := \{x \in \mathcal{X} \,|\, f(x) \in Y\}$ denotes the inverse image of Y;
  $g \circ f$ denotes the composite function, $(g \circ f)(x) := g\big(f(x)\big)$.
- Consider a function $\phi : \mathcal{X} \to \mathbb{R}$ and assume that $\mu := \min\limits_{x\in\mathcal{X}} \phi(x)$ is well–defined. We let,

$$\operatorname*{argmin}_{x\in\mathcal{X}} \phi(x) := \{x \in \mathcal{X} \,|\, \phi(x) = \mu\}\,.$$

  We write $y = \operatorname*{argmin}\limits_{x\in\mathcal{X}} \phi(x)$ to mean any $y \in \operatorname*{argmin}\limits_{x\in\mathcal{X}} \phi(x)$. Symmetric assumptions are considered for $\operatorname*{argmax}\limits_{x\in\mathcal{X}} \phi(x)$.

- Let f and g be two real functions defined on a half–line $[x_0\,, +\infty[\subset \mathbb{R}$ and such that $\lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} g(x) = +\infty$. By the notation

$$\text{for} \quad x \to +\infty\,, \quad f(x) \sim g(x)$$

  we mean that $\lim_{x\to+\infty} \frac{f(x)}{g(x)} = 1$ or, equivalently, that $f(x) = g(x)\big(1 + h(x)\big)$ with $\lim_{x\to+\infty} h(x) = 0$.

Chapter 2

Definitions and examples

This thesis is concerned with the *practical* stabilization problem for a discrete time and time–invariant dynamical system of the type

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = q_{\mathcal{Y}}\big(x(t)\big) \\ x \in \mathbb{R}^n, \quad u \in \mathcal{U} \subset \mathbb{R}^m, \quad y \in \mathcal{Y} \\ A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m}, \quad t \in \mathbb{N}\,. \end{cases} \tag{2.1}$$

It is assumed that $0 \in \mathcal{U}$, thus $(x = 0\,, u = 0)$ is an equilibrium pair, and $\mathcal{U}$ is an assigned *quantized* set. Also the output map $q_{\mathcal{Y}} : \mathbb{R}^n \to \mathcal{Y}$ is assigned and, according to the considered problem, either is the identity map $q_{\mathcal{Y}}(x) = x$ (full state available) or $\mathcal{Y}$ is a countable set (*quantized measurement*).
Some clarifications are in order: first, the definition of quantized set is needed; secondly, precise assumptions on the output map $q_{\mathcal{Y}}$ are to be specified; finally it must be clarified what a controller is and which the proper notion of practical stability/stabilization for system (2.1) is. This is the subject of the next few sections. The review of these points allows us to emphasize some features of system (2.1) that raise some of the main questions on the stabilization problem which are studied in this thesis.
Our first concern is the case where the input set $\mathcal{U}$ and the output map $q_{\mathcal{Y}}$ are given. No special assumptions are done on the structure of $\mathcal{U}$. Particular cases, such as *uniform* or *logarithmic* quantizations, will be considered in examples only, whilst the theory can be applied to very general input sets. Similarly, the assumptions specified for $q_{\mathcal{Y}}$ will be really mild. Most of the proposed results can be extended to the case where the designer has the freedom to choose the input set and/or the output map: references to this case will be given in remarks and examples.

2.1 Quantized sets and locally finite partitions

We assume that the reader is familiar with basic notions of general topology, such as those of *closed set*, *compact set*, *isolated point* and *accumulation point* for a set. These concepts can be revised in [66] .

Definition 1 *A set $\mathcal{U}$ endowed with a topology is said to be* discrete *iff all its points are isolated.*

Definition 2 *A set $\mathcal{U} \subset \mathbb{R}^m$ is said to be* quantized *iff it is closed and discrete with respect to the standard Euclidean topology.*

Elementary examples of quantized sets are the following: any finite set $\mathcal{U} \subset \mathbb{R}^m$ is a quantized set; an easy case of a quantized set having infinite cardinality is $\mathcal{U} = \mathbb{Z} \subset \mathbb{R}$.

Lemma 1 (Basic properties of quantized sets)
ı) The following properties are equivalent:

1- *$\mathcal{U} \subset \mathbb{R}^m$ is quantized;*
2- *$\mathcal{U}$ has no accumulation points;*
3- *$\forall\, \mathcal{S} \subset \mathbb{R}^m$ such that $\mathcal{S}$ is bounded, it holds that $\#\,(\mathcal{U} \cap \mathcal{S}) < +\infty$.*

ıı) If $\mathcal{U} \subset \mathbb{R}^m$ is quantized, then $\mathcal{U}$ is countable.

Proof. See in Appendix A.1.1 . ■

While quantized sets are countable, the contrary is not true: e.g., the set of rational numbers $\mathbb{Q} \subset \mathbb{R}$ is countable but it is not a quantized set. With an expressive sentence, we may say that the notion of quantized set is the closest concept generalizing that of a finite set.
A relaxation of the notion of quantized set, where the accumulation of values towards one point (which is taken to be 0) is allowed, is provided by the following

Definition 3 *A set $\mathcal{U} \subset \mathbb{R}^m$ is said to be* quantized *in the* generalized *sense (generalized quantized set) iff $\mathcal{U}$ is a closed set and $\mathcal{U} \setminus \{0\}$ is discrete.*

If $\mathcal{U}$ is a quantized set, then it is also a generalized quantized set (but not vice versa, see next Example 1) . Hence, the properties stated for generalized quantized sets hold also for quantized sets.

Example 1 *The set*

$$\mathcal{U} := \{0\} \cup \left\{ \pm \frac{1}{n+1} \,|\, n \in \mathbb{N} \right\}$$

is a generalized quantized set but it is not a quantized set. In fact, 0 is an accumulation point for $\mathcal{U}$. In other words, $\mathcal{U}$ is closed but it is not discrete . ♣

We shall see in Section 2.3 that the stabilization problem for system (2.1) is significantly different depending on the fact that the control set $\mathcal{U}$ is assumed to be quantized or quantized in the generalized sense.
Let $\mathcal{U}_{\rm g} \subset \mathbb{R}^m$ be a generalized quantized set. A quantized set can be obtained by the *truncation* of $\mathcal{U}_{\rm g}$: consider a neighborhood Ω_0 of 0, then $\mathcal{U} := \{0\} \cup \left(\mathcal{U}_{\rm g} \cap {}^c\Omega_0\right)$ is a quantized set which is said to be obtained by truncation of $\mathcal{U}_{\rm g}$ within Ω_0. In a broad sense, the term truncation will be used to allude to the fact that 0 is an isolated point.

Definition 4 *Let $\mathcal{U} \subset \mathbb{R}^m$ be a generalized quantized set: a* quantizer *is a map*

$$q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U} .$$

When $\mathcal{U}$ is a control set, $q_{\mathcal{U}}$ is called an input quantizer.
For a given quantizer, the corresponding quantization error *is the map defined by*

$$\begin{array}{rccc} q_e := q_{\mathcal{U}} - I : & \mathbb{R}^m & \to & \mathbb{R}^m \\ & y & \mapsto & q_{\mathcal{U}}(y) - y . \end{array}$$

The following special class of quantizers is relevant for our study:

Definition 5 *Let $\mathcal{U} \subset \mathbb{R}^m$ be a generalized quantized set: a quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ is said to be a* nearest neighbor *quantizer iff $\forall\, y \in \mathbb{R}^m$, $q_{\mathcal{U}}(y)$ is an element of $\mathcal{U}$ minimizing the Euclidean distance from y.*

The nearest neighbor quantizer is well–defined because $\mathcal{U}$ is a closed set. For a given set $\mathcal{U}$, there are several nearest neighbor quantizers but they only differ for the values taken in correspondence of those $y \in \mathbb{R}^m$ such that

$$\# \operatorname*{argmin}_{u \in \mathcal{U}} \|y - u\|_2 \geq 2 .$$

Notice also that, if $q_{\mathcal{U}}$ is a nearest neighbor quantizer, then the associated quantization error is such that the function

$$\begin{array}{rccc} \|q_e\|_2 : & \mathbb{R}^m & \to & \mathbb{R}^+ \\ & y & \mapsto & \|q_e(y)\|_2 \end{array}$$

is continuous.

Definition 6 *A* partition *of $\mathbb{R}^m$ is a family of subsets of $\mathbb{R}^m$, say $\{\mathcal{C}_i\}_{i \in \mathcal{I}}$ (for some set of indices $\mathcal{I}$), such that $\mathbb{R}^m = \bigcup_{i \in \mathcal{I}} \mathcal{C}_i$ and $\forall\, i_1, i_2 \in \mathcal{I}$, $i_1 \neq i_2$, $\mathcal{C}_{i_1} \cap \mathcal{C}_{i_2} = \emptyset$.*
A partition is said to be locally finite *iff for any bounded set $\mathcal{S} \subset \mathbb{R}^m$, it holds that*

$$\#\left\{i \in \mathcal{I} \,|\, \mathcal{S} \cap \mathcal{C}_i \neq \emptyset\right\} < +\infty .$$

For any given map $f : \mathbb{R}^m \to \mathcal{I}$, the partition of $\mathbb{R}^m$ defined by

$$\begin{cases} \mathbb{R}^m = \bigcup_{i \in \mathcal{I}} \mathcal{C}_i \\ \mathcal{C}_i := \left\{y \in \mathbb{R}^m \,|\, f(y) = i\right\} \end{cases}$$

is referred to as the partition induced by *f.*

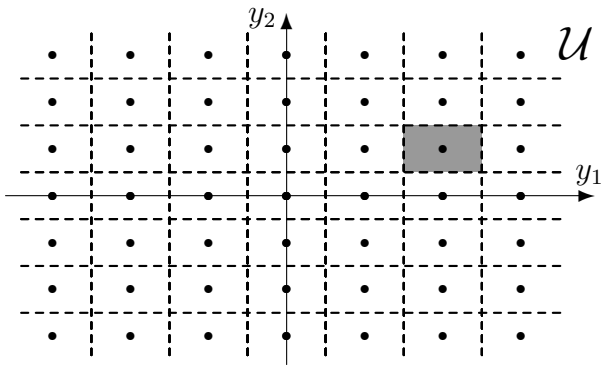

Figure 2.1: Uniform quantization of $\mathbb{R}^2$ with $u_{02} = \frac{3}{5} \cdot u_{01}$. Broken lines define the partition of $\mathbb{R}^2$ induced by $q_{\mathcal{U}}$.

Let us discuss the relations between locally finite partitions and the partitions induced by a quantizer.
The partition induced by a nearest neighbor quantizer is the so called *Voronoi* partition generated by $\mathcal{U}$ (see [94]). If $\mathcal{U}$ is a quantized set, then the Voronoi partition generated by $\mathcal{U}$ is locally finite.
Given a locally finite partition $\{\mathcal{C}_i\}_{i\in\mathcal{I}}$ of $\mathbb{R}^m$, $\forall\, i \in \mathcal{I}$, pick[1] $u_i \in \mathcal{C}_i$. It is easy to see that the set $\mathcal{U} := \{u_i \,|\, i \in \mathcal{I}\}$ is a quantized set. Moreover, the quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ defined by $q_{\mathcal{U}}(y) = u_i$ if and only if $y \in \mathcal{C}_i$ is such that its induced partition is the given one and $\forall\, u \in \mathcal{U}$, $u \in q_{\mathcal{U}}^{-1}(u)$. Vice versa, it is easy to construct examples of quantizers $q_{\mathcal{U}}$ taking values in a quantized set $\mathcal{U}$ (not in the generalized sense) and such that $\forall\, u \in \mathcal{U}$, $u \in q_{\mathcal{U}}^{-1}(u)$ but the induced partition is not locally finite (see Example 30 in Appendix A.1.1): of course, such a quantizer is not a nearest neighbor one.

Let us provide some typical examples of quantizers.

Definition 7 (Uniform quantization of $\mathbb{R}^m$) *Let $u_0 > 0$. A set $\mathcal{U} \subset \mathbb{R}$ is said to be* uniformly quantized with parameter u_0 *iff $\mathcal{U} = u_0\mathbb{Z}$. By a* uniform quantization of $\mathbb{R}$ with parameter u_0 *we mean a nearest neighbor quantizer $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U} := u_0\mathbb{Z}$.*
For $i = 1, \dots, m$, let $u_{0i} > 0$. By a uniform quantization of $\mathbb{R}^m$ with parameters $(u_{01}, \dots, u_{0m})$ *we mean a nearest neighbor quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U} := u_{01}\mathbb{Z} \times \cdots \times u_{0m}\mathbb{Z}$ (see Fig. 2.1).*

It is straightforward to see that

Lemma 2 *If $q_{\mathcal{U}}$ is a uniform quantization of $\mathbb{R}^m$ with parameters $(u_{01}, \dots, u_{0m})$ and q_e is the corresponding quantization error, then $\forall\, y \in \mathbb{R}^m$, it holds that*

$$\|q_e(y)\|_2 \leq \frac{1}{2}\sqrt{\sum_{i=1}^{m} u_{0i}^2}\,. \qquad \square$$

[1] Notice that, for the construction, we are using the Zermelo's axiom of choice [86].

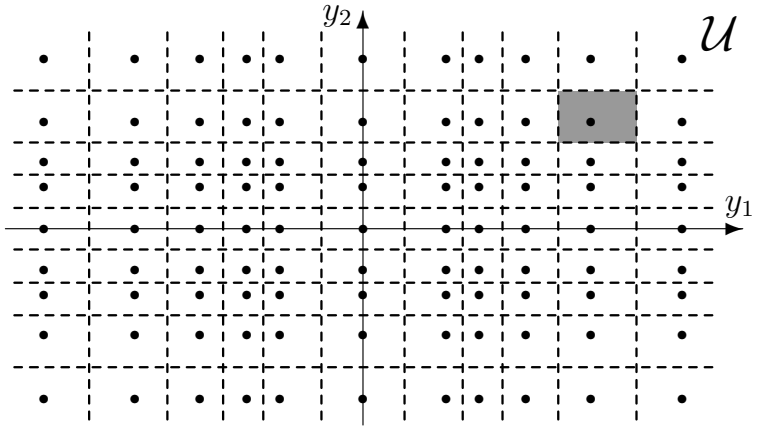

Figure 2.2: Componentwise logarithmic quantization of $\mathbb{R}^2$ with $u_{02} = \frac{1}{2} \cdot u_{01}$, $\theta_1 = 1.4$ and $\theta_2 = 1.6$. Broken lines define the partition of $\mathbb{R}^2$ induced by $q_{\mathcal{U}}$.

Definition 8 (Componentwise logarithmic quantization of $\mathbb{R}^m$) *Let $u_0 > 0$ and $\theta > 1$. A set $\mathcal{U} \subset \mathbb{R}$ is said to be* logarithmically quantized with parameters (u_0, θ) *iff*

$$\mathcal{U} = \{0\} \cup \{\pm u_0 \theta^h \,|\, h \in \mathbb{N}\}.$$

If instead $h \in \mathbb{Z}$, then $\mathcal{U}$ is a generalized quantized set and it is said to be logarithmically quantized *in the* generalized *sense.*
By a (generalized) logarithmic quantization of $\mathbb{R}$ with parameters (u_0, θ) *we mean a nearest neighbor quantizer $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$, where $\mathcal{U}$ is a (generalized) logarithmically quantized set of parameters (u_0, θ).*
For $i = 1 \dots m$, let $u_{0i} > 0$, $\theta_i > 1$ and $\mathcal{U}_i$ be a logarithmically quantized set of parameters (u_{0i}, θ_i). By a componentwise logarithmic quantization of $\mathbb{R}^m$ with parameters $\big((u_{01}, \theta_1), \dots, (u_{0m}, \theta_m)\big)$ *we mean a nearest neighbor quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U} := \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ (see Fig. 2.2).*

The quantizer

$$\begin{array}{rccl} q_{\mathcal{U}} : & \mathbb{R}^m & \to & \mathcal{U}_1 \times \cdots \times \mathcal{U}_m \\ & y & \mapsto & \big(q_{\mathcal{U}_1}(y_1), \cdots, q_{\mathcal{U}_m}(y_m)\big), \end{array}$$

where $\forall\, i = 1, \dots, m$, $q_{\mathcal{U}_i} : \mathbb{R} \to \mathcal{U}_i$ is a logarithmic quantization of $\mathbb{R}$, is a particular componentwise logarithmic quantization of $\mathbb{R}^m$.
Notice that the cartesian product of generalized logarithmically quantized sets is not a generalized quantized set: in fact, fix an index $i \in \{1 \dots, m\}$ and, $\forall\, j \neq i$, fix an integer k_j, then the sequence

$$\mathcal{U} \supset \{u_k\}_{k \in \mathbb{Z}} = \big\{ (\, u_{01}\theta_1^{k_1} \;\cdots\; u_{0i}\theta_i^{k} \;\cdots\; u_{0m}\theta_m^{k_m}\,) \big\}_{k \in \mathbb{Z}}$$

is accumulating towards

$$u = (u_{01}\theta_1^{k_1} \ \cdots \ u_{0i-1}\theta_{i-1}^{k_{i-1}} \ \ 0 \ \ u_{0i+1}\theta_{i+1}^{k_{i+1}} \ \cdots \ u_{0m}\theta_m^{k_m}) .$$

Hence, there are infinite accumulations points different from 0.
For logarithmic quantizations of $\mathbb{R}^m$, the quantization error is not bounded. Nevertheless, the *relative* quantization error is bounded:

Lemma 3 *Consider a generalized logarithmic quantization of $\mathbb{R}$ with parameters (u_0, θ), then $\forall\, y \neq 0$,*

$$\frac{|q_e(y)|}{|y|} \leq \frac{\theta - 1}{\theta + 1} .$$

Proof. See in Appendix A.1.1. ■

The relative error of a logarithmic quantization of $\mathbb{R}$ differs from that of a generalized logarithmic quantization only in a neighborhood of 0. Because of the truncation, around 0 the relative error is larger than in the generalized logarithmic quantization case and it reaches its maximal value equal to 1 for $0 < |y| \leq u_0/2$. The details will be illustrated in Example 19 of Chapter 5 (see also Fig. A.2 in Appendix A.4.1). In Example 20 of the same chapter, also the behavior of the relative error of componentwise logarithmic quantizations of $\mathbb{R}^2$ is studied.

In the above examples, the quantized sets $\mathcal{U} \subset \mathbb{R}^m$ are all in the form of a cartesian product, that is $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$, with $\mathcal{U}_i \subset \mathbb{R}$. In this case, it is possible to deal with quantizers $q_{\mathcal{U}}$ acting separately on each component of the vector $y \in \mathbb{R}^m$. This is not possible for more general quantized sets $\mathcal{U} \subset \mathbb{R}^m$:

Definition 9 (Joint radial logarithmic quantization of $\mathbb{R}^2$) *Let $\mathbb{N} \ni N \geq 3$, $u_0 > 0$ and $\theta > 1$. Consider*

$$\ell_k := \Big\{ (x_1, x_2) = \big(\lambda \cos(2\pi k/N), \lambda \sin(2\pi k/N)\big) \in \mathbb{R}^2 \,|\, \lambda \geq 0 \Big\} , \quad k = 0, 1, \ldots, N-1 ,$$

$$c_h := \{ x \in \mathbb{R}^2 \,|\, \|x\|_2 = u_0 \theta^h \} , \quad h \in \mathbb{N} .$$

Let $\mathcal{L} := \bigcup_{k=0}^{N-1} \ell_k$ and $\mathcal{C} := \bigcup_{h \in \mathbb{N}} c_h$. A set $\mathcal{U} \subset \mathbb{R}^2$ is said to be radially logarithmically quantized with parameters (N, u_0, θ) *iff*

$$\mathcal{U} = \{0\} \cup \big(\mathcal{L} \cap \mathcal{C}\big)$$

(see Fig. 2.3). If instead, in the definitions of c_h and $\mathcal{C}$, we let $h \in \mathbb{Z}$, then $\mathcal{U}$ is a generalized quantized set and it is said to be radially logarithmically quantized *in the* generalized *sense.*
By a (generalized) radial logarithmic quantization of $\mathbb{R}^2$ with parameters (N, u_0, θ) *we mean a nearest neighbor quantizer $q_{\mathcal{U}} : \mathbb{R}^2 \to \mathcal{U}$, where $\mathcal{U}$ is a (generalized) radially logarithmically quantized set of parameters (N, u_0, θ).*

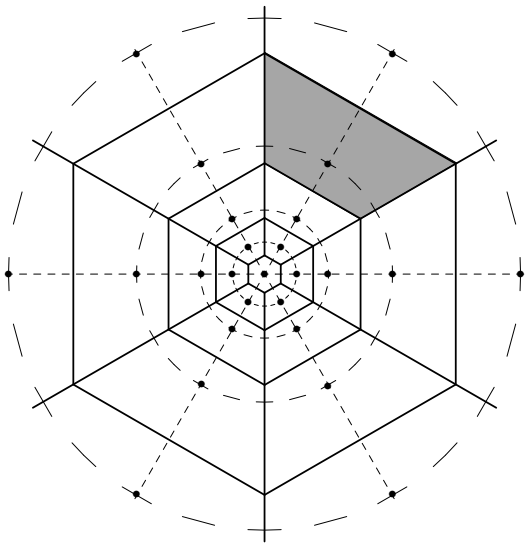

Figure 2.3: Radial logarithmic quantization of $\mathbb{R}^2$ with $N = 6$ and $\theta = 2$. Full lines define the partition of $\mathbb{R}^2$ induced by $q_{\mathcal{U}}$.

Also for the generalized radial logarithmic quantization the relative quantization error is bounded.

Lemma 4 *Consider a generalized radial logarithmic quantization of* $\mathbb{R}^2$ *with parameters* (N, u_0, θ), *then* $\forall\, y \in \mathbb{R}^2 \setminus \{0\}$,

$$\frac{\|q_e(y)\|_2}{\|y\|_2} \leq \sqrt{1 - \frac{4\theta \cos^2(\pi/N)}{(\theta+1)^2}}\,.$$

Proof. The joint radial logarithmic quantization will be studied in details in Chapter 5 and the proof of this lemma is reported at the end of Appendix A.4.1. ■

Considerations similar to those we made for the relative error of a logarithmic quantization of $\mathbb{R}$ can be done on the behavior of the relative error of a radial logarithmic quantization of $\mathbb{R}^2$ (in particular, its maximal value is equal to 1 in a neighborhood of 0). The details will be illustrated in Example 21 of Chapter 5.

In the particular case of quantized sets $\mathcal{U} \subset \mathbb{R}$, it is useful to introduce the following quantities.

Definition 10 *Let* $\mathcal{U} \subset \mathbb{R}$ *be a quantized set such that* $\mathcal{U} \neq \{0\}$. *The* resolution at 0 *of* $\mathcal{U}$ *is defined by*

$$u_0 := \min_{u \in \mathcal{U} \setminus \{0\}} |u|\,.$$

The definition of resolution at 0 can be generalized to a notion taking into account the overall control set or, more in general, a portion of it:

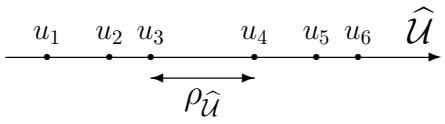

Figure 2.4: Representation of the dispersion of a quantized set $\widehat{\mathcal{U}} = \{u_1, \dots, u_6\} \subset \mathbb{R}$.

Definition 11 *Consider a quantized set $\mathcal{U} \subset \mathbb{R}$, let $\widehat{\mathcal{U}} \subseteq \mathcal{U}$ and denote by $\widehat{\mathcal{U}}^{\,\mathrm{ch}}$ the convex hull of $\widehat{\mathcal{U}}$. The* dispersion *of $\widehat{\mathcal{U}}$ is defined as*

$$\rho_{\widehat{\mathcal{U}}} := \begin{cases} \sup\left\{ b-a \,\middle|\,]a,b[\subseteq \widehat{\mathcal{U}}^{\,\mathrm{ch}} \quad \text{and} \quad]a,b[\cap \widehat{\mathcal{U}} = \emptyset \right\} & \text{if } \#\widehat{\mathcal{U}} > 1 \\ +\infty & \text{otherwise.} \end{cases} \tag{2.2}$$

In plain words, the dispersion of $\widehat{\mathcal{U}}$ is the maximal gap between consecutive elements of $\widehat{\mathcal{U}}$ (see Fig. 2.4).

2.2 The output map $q_{\mathcal{Y}}$

The output map $q_{\mathcal{Y}} : \mathbb{R}^n \to \mathcal{Y}$ of system (2.1) is characterized by its induced state space partition:

$$\begin{cases} \mathbb{R}^n = \bigcup_{y \in \mathcal{Y}} \mathcal{C}_y \\ \mathcal{C}_y := q_{\mathcal{Y}}^{-1}(y). \end{cases} \tag{2.3}$$

We consider three cases:

1. **Full state:** it is the case in which quantization is on inputs only while full state is available. Namely, $q_{\mathcal{Y}}(x) = x$ and $\mathcal{Y} = \mathbb{R}^n$ (clearly, $\forall\, y \in \mathcal{Y}$, $\mathcal{C}_y = \{y\}$). In this case, system (2.1) is referred to as *full state* or as *quantized input* (depending on the context) and it is often denoted by $\Sigma(A, B, \mathcal{U})$.

The following two instances are generically referred to as *quantized measurement* case. In both cases, the induced state space partition is assumed to be locally finite.

2. **State quantization:** it is the case in which, at least in a sufficiently large neighborhood of the equilibrium, the state space partition is made of bounded sets. Namely, there exists a sufficiently large $r > 0$ such that $\forall\, y \in q_{\mathcal{Y}}(\mathcal{B}_r)$, $\mathcal{C}_y$ is bounded. In this case, system (2.1) is referred to as *quantized input* and *quantized state* (or quantized state, for short).

3. **Output quantization:** it is the case in which the output map is of the type $q_{\mathcal{Y}} = q_o \circ C$, with $C \in \mathbb{R}^{q \times n}$ ($q < n$) and the map $q_o : \mathbb{R}^q \to \mathcal{Y}$ induces a locally finite partition of $\mathbb{R}^q$. It is assumed that (A, C) is an observable pair. As $q < n$, the state space partition is made of unbounded sets. In this thesis, we only deal with single–output

systems (i.e., $q = 1$) and it is assumed that $\forall\, y \in \mathcal{Y}$, $q_o^{-1}(y) \subseteq \mathbb{R}$ is a connected set $\big($thus, $q_o^{-1}(y) \subseteq \mathbb{R}$ is either an interval of finite length or a half–line$\big)$. This case of system (2.1) is referred to as *quantized input* and *quantized output* (or quantized output, for short).

The quantized measurement case is a model for situations where, although the dynamics is deterministic, only partial information about the state of the system are available.

2.3 The stabilization problem for quantized linear systems

The basic fact on the stabilization problem for system (2.1) is that, if $\mathcal{U}$ is a quantized set and the system is open loop unstable, then stabilization in the Lyapunov sense is not possible (see also [29]). Namely, for any control law $u(\,\cdot\,)$ taking values in a quantized set $\mathcal{U}$, 0 is not a stable equilibrium for the closed loop dynamics (regardless of the argument of u). This gives reasons for the need of introducing a weaker notion of stability to be considered for quantized systems, that is the so called *practical stability*. The following example is useful to gain insight on this fact:

Example 2 *Consider the quantized input system*

$$\begin{cases} x(t+1) = ax(t) + u(t) \\ |a| > 1 \\ u \in \mathcal{U} \subset \mathbb{R}\,, \end{cases} \tag{2.4}$$

where $\mathcal{U} \neq \{0\}$ is a quantized set containing 0. It holds that any control law $u(\,\cdot\,)$ taking values in $\mathcal{U}$ does not stabilize the system. In fact, assume that $u(\,\cdot\,)$ is such that 0 is a stable equilibrium for the closed loop system: by definition [67], this means that

$$\forall\, \epsilon > 0 \ \text{ and } \ \forall\, t_0 \geq 0\,, \quad \exists\, \delta(\epsilon, t_0) > 0 \quad \text{such that} \quad |x(t_0)| < \delta(\epsilon, t_0) \ \Rightarrow\ |x(t)| < \epsilon \quad \forall\, t \geq t_0\,.$$

We claim that, if such a control law exists, then $u(t) = 0$ whenever $|x(t)| < \frac{u_0}{1+|a|}$, where u_0 is the resolution at 0 of the control set $\mathcal{U}$. This is a contradiction because, in the neighborhood $\Omega_0 := \big\{x \in \mathbb{R} \,\big|\, |x| < \frac{u_0}{1+|a|}\big\}$, the closed loop system coincides with the open loop dynamics $x(t+1) = ax(t)$ which is unstable.
The proof of the claim is trivial, it is sufficient to show that, for $\epsilon = \frac{u_0}{1+|a|}$, it holds that, if $|x(t)| < \epsilon$ and $|x(t+1)| < \epsilon$, then $u(t) = 0$: $|x(t+1)| < \epsilon$ if and only if $-\epsilon - ax(t) < u(t) < \epsilon - ax(t)$. From this inequality it follows that if $|x(t)| < \epsilon = \frac{u_0}{1+|a|}$, then $|u(t)| < u_0$. By definition of u_0, this implies that $u(t) = 0$. ♣

The key point of the example is that the value $u = 0$ is an isolated point of $\mathcal{U}$ (i.e., $u_0 > 0$). The same arguments can be hence easily extended to show that, for any open loop unstable quantized linear system (2.1) and any control law taking values in a quantized set, if $x = 0$ is an equilibrium for the closed loop dynamics, then such an equilibrium is unstable.
Next Example 3 shows that, if instead the control set is quantized in the generalized sense, then closed loop asymptotic stability can be achieved.

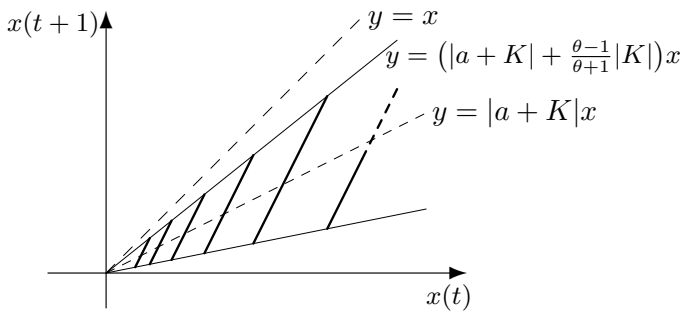

Figure 2.5: Closed loop dynamics in Example 3 ($a = 2$, $K = -3/2$ and $\theta = 3/2$) .

Example 3 *Consider again system* (2.4) *but assume that* $\mathcal{U}$ *is a logarithmically quantized set in the generalized sense with parameters* (u_0, θ) *. Let us consider a control law of the type* $u(x) = q_{\mathcal{U}}(Kx)$ *, where* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *is a nearest neighbor quantizer and* $K \in \mathbb{R}$ *is such that* $|a+K| < 1$ *(i.e.,* K *is a stabilizing control gain in the ideal case of absence of quantization) . The corresponding closed loop dynamics is*

$$x(t+1) = (a+K)x(t) + q_e\big(Kx(t)\big) \ :$$

this is the feedback interconnection of the asymptotically stable linear system $\Sigma(a+K, 1, K)$ *with the nonlinearity* q_e *representing the quantization error.*
If K *is such that* $\forall\, x \neq 0$, $|x^+| < |x|$, *then* $V(x) := x^2$ *is a Lyapunov function for the closed loop system and* $x = 0$ *is a globally asymptotically stable equilibrium. We look for a condition on* K *ensuring that this property holds irrespective of the choice of the particular nearest neighbor quantizer (recall that, given* $\mathcal{U}$ *, a nearest neighbor quantizer is not unique) . It is easy to show that a necessary and sufficient condition in order that this happen is*

$$|a+K| + \frac{\theta-1}{\theta+1}|K| < 1 \tag{2.5}$$

(see Fig. 2.5 and Lemma 25 in Appendix A.1.2) . With $\gamma_*(\theta) := \frac{\theta-1}{\theta+1}$ *and* $\gamma_s(K) := \frac{|K|}{1-|a+K|}$ *, under the assumption that* $|a+K| < 1$ *, condition* (2.5) *is equivalent to*

$$\gamma_s(K) \cdot \gamma_*(\theta) < 1 \,. \tag{2.6}$$

Inequality (2.6) *is called* small–gain *condition: this issue will be investigated in Chapters 5 and 6 , here we tell in advance that* $\gamma_s(K)$ *is the* ℓ_2*–gain (or, equivalently, the* H_∞*–norm) of the linear system* $\Sigma(a+K, 1, K)$ *, whereas* $\gamma_*(\theta)$ *is the* ℓ_2*–gain of the nonlinearity* q_e *. The allowed choices for* K *ensuring that stabilization in the presence of quantization is achievable are less than in the ideal case of absence of quantization. In fact, let*

$$\mathcal{K}_{\text{sg}} := \{K \in \mathbb{R} \,|\, \gamma_s(K) \cdot \gamma_*(\theta) < 1\}$$

and

$$\mathcal{K} := \{K \in \mathbb{R} \,|\, |a+K| < 1\}\,,$$

by condition (2.5)*,*

$$\mathcal{K}_{\rm sg} \subset \mathcal{K}\,.$$

Moreover, if θ *is too large* $\big($*i.e., if* $\gamma_*(\theta)$ *is too close to* $1\big)$*,* $\mathcal{K}_{\rm sg}$ *may be empty. Indeed, as* $K \in \mathbb{R}$ *varies so that* $|a+K| < 1$*,* $\gamma_s(K)$ *is minimized for* $K = -a$*, accordingly* $\gamma_s(-a) = |a|$*. Hence,* $\exists\, K \in \mathbb{R}$ *such that, for any nearest neighbor quantizer,* $V(x) = x^2$ *is a Lyapunov function for the closed loop dynamics (or, equivalently,* $\mathcal{K}_{\rm sg} \neq \emptyset$*) if and only if* $|a|\frac{\theta-1}{\theta+1} < 1$*, that is*

$$\theta < \frac{|a|+1}{|a|-1}\,. \tag{2.7}$$

If instead one has the freedom to choose a generalized quantized set $\mathcal{U}$ *(i.e.,* $\mathcal{U}$ *is not assigned), then the problem is easier and stabilization is always achievable. Namely,* $\forall\, |a| > 1$ *and for any control gain* $K \in \mathcal{K}$*, there exists a logarithmically quantized set in the generalized sense such that the system is stabilized by* $u(x) = q_{\mathcal{U}}(Kx)$*. In fact, it is sufficient to choose the parameter* θ *so that* $\gamma_*(\theta) < 1/\gamma_s(K)$*. Therefore, any choice of* θ *within the non–empty interval*

$$\Big]1\,,\, \frac{\gamma_s(K)+1}{\gamma_s(K)-1}\Big[$$

is feasible. With $K = -a$*,* $\gamma_s(K)$ *is minimized and* $\frac{\gamma_s(K)+1}{\gamma_s(K)-1}$ *achieves its maximum value equal to* $\frac{|a|+1}{|a|-1}$*: since for* $|a| > 1$*,* $\frac{|a|+1}{|a|-1}$ *is a decreasing function of* $|a|$*, then the more unstable the open loop system is, the smaller the allowed values for* θ *are. As it is clear by Definition 8, smaller values of* θ *correspond to more* densely *quantized sets (see [39] for a formal definition of density of quantization).* ♣

Exactly because stabilization is possible with generalized quantized sets whereas it is not possible in the presence of quantization (unless the system is already open loop stable), our interest is focused on the latter case. Moreover, generalized quantized control sets with values accumulating towards the origin are an idealization where the peculiar features of quantization are lost. In fact, infinite control values belong to any bounded neighborhood of 0 and this situation is really different from the property stated in Lemma 1.*i* that makes quantized sets closely related to finite sets.

2.3.1 Practical stability

A notion of stability suited to quantized systems comes out naturally by the discussion of the following example.

Example 4 *Consider again system* (2.4)*. Let the control law be* $u(x) = q_{\mathcal{U}}(-ax)$*, where* $q_{\mathcal{U}}$ *is a nearest neighbor quantizer. The corresponding closed loop dynamics is*

$$x(t+1) = q_e\big(-ax(t)\big)\,.$$

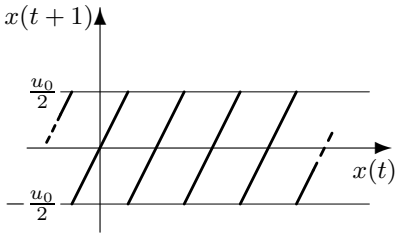

Figure 2.6: Closed loop dynamics in case 1 of Example 4 ($a = 2$) .

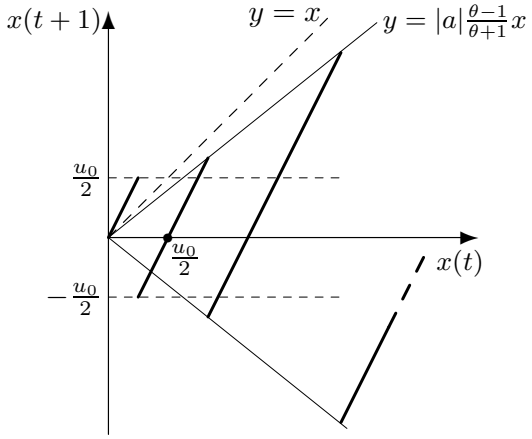

Figure 2.7: Closed loop dynamics in case 2 of Example 4 ($a = 2$ and $\theta = 7/3$) .

Let us analyze this dynamics for three different quantized sets.

Case 1: *assume that* $\mathcal{U} = u_0\mathbb{Z}$. *By Lemma 2, it holds that* $\forall\, x \in \mathbb{R}$, $|x^+| \leq u_0/2$ *(see Fig. 2.6). Namely, for* $t \geq 1$, *all the trajectories are confined inside the interval* $\Omega := \left[-\frac{u_0}{2} , \frac{u_0}{2} \right]$.

Case 2: *assume that* $\mathcal{U}$ *is a logarithmically quantized set of parameters* (u_0, θ) *and* θ *is such that inequality* (2.7) *holds. In this case, for* $|x| \leq u_0/2$, *it holds that* $|x^+| = |q_e(-ax)| \leq u_0/2$, *whereas for* $|x| > u_0/2$, $|x^+| \leq |a|\frac{\theta-1}{\theta+1} \cdot |x| < |x|$ *by inequality* (2.7) *(see Fig. 2.7). Thus, with* $\Omega := [-\frac{u_0}{2} , \frac{u_0}{2}]$, *it holds that* $\forall\, x \in \Omega$, $x^+ \in \Omega$ *and* $\forall\, x(0) \in \mathbb{R}$, $\exists\, t \in \mathbb{N}$ *such that* $x(t) \in \Omega$ *: also in this case all the trajectories eventually enters a neighborhood* Ω *of the equilibrium and remain confined therein.*

Case 3: *assume that* $\mathcal{U} = \{0\} \cup \{\pm u_0 , \pm 2u_0 , \pm 4u_0 , \pm 6u_0\}$ *(thus,* $\mathcal{U}$ *is a finite set) and* $a = 2$. *In this case, for* $|x| \leq u_0/2$, *it holds that* $|x^+| = |q_e(-2x)| \leq u_0/2$ *; for* $u_0/2 < |x| < 6u_0$, $|x^+| < |x|$ *and for* $|x| \geq 6u_0$, $|x^+| \geq |x|$ *(see Fig. 2.8). Thus, with* $\Omega := [-\frac{u_0}{2} , \frac{u_0}{2}]$ *and* $X_0 := [-\frac{\Delta_0}{2} , \frac{\Delta_0}{2}]$ $\big($*for any* $\Delta_0 \in [0 , 12u_0[\big)$, *it holds that* $\forall\, x \in \Omega$, $x^+ \in \Omega$ *and* $\forall\, x(0) \in X_0$, $\exists\, t \in \mathbb{N}$ *such that* $x(t) \in \Omega$. *That is, all the trajectories starting from* X_0 *eventually enters a neighborhood* Ω *of the equilibrium and remain confined therein.* ♣

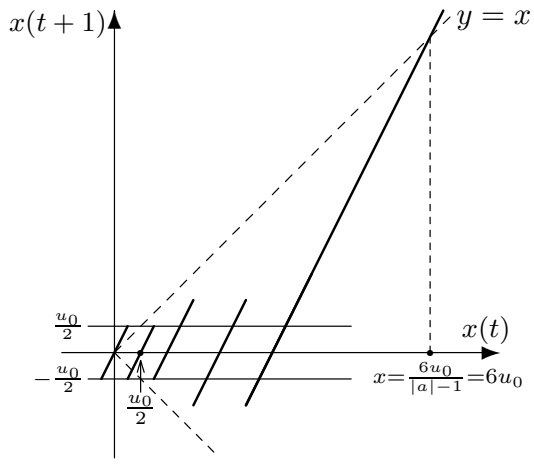

Figure 2.8: Closed loop dynamics in case 3 of Example 4 .

In the three cases, although $x = 0$ is an unstable equilibrium for the closed loop dynamics, it is possible to make the trajectories non–divergent (in the finite case, only those starting from a sufficiently small neighborhood X_0 of the equilibrium) and, even better, to make them convergent to a bounded neighborhood Ω of the equilibrium. This property is not stability in the Lyapunov sense, but it is the closest behavior to stability that one can aim to obtain by the quantized control of an open loop unstable linear system. This is an example of the typical behavior of the so called *practically stable* systems.

We are ready to introduce the formal notion of practical stability. Many definitions for this property have been introduced in the literature on quantized systems, all of them are related with the requirement of ultimate boundedness of the trajectories and with the notion of *invariant* set [11] .
Let us consider a time invariant dynamical system

$$\begin{cases} x^+ = f(x) \\ x \in \mathbb{R}^n. \end{cases} \tag{2.8}$$

Definition 12 *A set* $\Omega \subseteq \mathbb{R}^n$ *is said to be* positively invariant *for system* (2.8) *iff* $\forall\, x \in \Omega$, $x^+ \in \Omega$.

We consider the following notions of practical stability:

Definition 13 (Practical stability) *Let* Ω , X_0 *and* X_1 *be bounded subsets of* $\mathbb{R}^n$ *such that* Ω *and* X_0 *are neighborhoods of the origin,* $\Omega \subseteq X_1$ *and* $X_0 \subseteq X_1$.
ı) System (2.8) *is said to be* (X_0, X_1, Ω)–stable *iff* $\forall\, x(0) \in X_0$, $x(t) \in X_1 \ \ \forall\, t \geq 0$ *and* $\exists\, \bar{t} \in \mathbb{N}$ *such that* $\forall\, t \geq \bar{t}$, $x(t) \in \Omega$.
ıı) System (2.8) *is said to be* (X_0, Ω)–stable *iff both* X_0 *and* Ω *are positively invariant and* $\forall\, x(0) \in X_0 \ \ \exists\, \bar{t} \in \mathbb{N}$ *such that* $\forall\, t \geq \bar{t}$, $x(t) \in \Omega$.

Remark 1 ((X_0, X_0, Ω)–stability vs (X_0, Ω)–stability) *Consider the particularization of (X_0, X_1, Ω)–stability to the case $X_1 = X_0$, that is (X_0, X_0, Ω)–stability. If system (2.8) is (X_0, Ω)–stable, then it is (X_0, X_0, Ω)–stable. In general, the contrary is not true because, although the trajectories starting from X_0 will eventually remain confined within Ω, the set Ω is not guaranteed to be positively invariant. Thus, (X_0, X_0, Ω)–stability, is a weaker stability notion than (X_0, Ω)–stability.*

Let us introduce the practical stabilization problem for system (2.1). First, the notion of positive invariance is extended to controlled systems:

Definition 14 *A set $\Omega \subseteq \mathbb{R}^n$ is said to be $q_{\mathcal{Y}}$–*controlled invariant *for system (2.1) iff $\forall\, y \in q_{\mathcal{Y}}(\Omega)$, $\exists\, u \in \mathcal{U}$ such that $\forall\, x \in q_{\mathcal{Y}}^{-1}(y) \cap \Omega$, $x^+ = Ax + Bu \in \Omega$.*
Controlled invariance is the particularization of the notion of $q_{\mathcal{Y}}$–controlled invariance to quantized input systems (i.e., $q_{\mathcal{Y}}$ is the identity map): a set $\Omega \subseteq \mathbb{R}^n$ is said to be controlled invariant *for system $\Sigma(A, B, \mathcal{U})$ iff $\forall\, x \in \Omega$, $\exists\, u \in \mathcal{U}$ such that $x^+ = Ax + Bu \in \Omega$.*

Namely, $\forall\, x \in \Omega$ it must be possible to select a control, as a function of the available measurement $q_{\mathcal{Y}}(x)$ only, such that $x^+ \in \Omega$.
Given the most general form of system (2.1) (i.e., in the quantized input and quantized output case), a controller is a *machine* (see [114]) that, based on quantized output measurements $y \in \mathcal{Y}$ and on the internal state $w \in \mathcal{W}$, selects a quantized control value $u \in \mathcal{U}$. In formulae, the controller is described by the following system defined on some set $\mathcal{W}$:

$$\begin{cases} w(t+1) = \gamma\big(w(t), y(t), t\big) \\ u(t) = \bar{k}\big(w(t), y(t), t\big) \\ t \in \mathbb{N}, \end{cases} \tag{2.9}$$

where $\gamma : \mathcal{W} \times \mathcal{Y} \times \mathbb{N} \to \mathcal{W}$ and $\bar{k} : \mathcal{W} \times \mathcal{Y} \times \mathbb{N} \to \mathcal{U}$ (the map $\bar{k}$, when it is not explicitly depending on the time, is simply denoted by k). The closed loop dynamics induced by the feedback interconnection of such a controller with system (2.1) is:

$$\begin{cases} x(t+1) = Ax(t) + B\bar{k}\Big(w(t), q_{\mathcal{Y}}\big(x(t)\big), t\Big) \\ w(t+1) = \gamma\left(w(t), q_{\mathcal{Y}}\big(x(t)\big), t\right). \end{cases} \tag{2.10}$$

In some cases (i.e., full state or quantized state systems), it will be sufficient to consider the subclass of controllers (2.9) made of static and time–invariant controllers: namely, $\mathcal{W} = \{w\}$ and $\bar{k}$ is not depending on t. In this case, the control law is a static state feedback $u(\cdot)$ of the type $k \circ q_{\mathcal{Y}} : \mathbb{R}^n \to \mathcal{U}$, for some $k : \mathcal{Y} \to \mathcal{U}$, and the closed loop dynamics

$$x^+ = Ax + B(k \circ q_{\mathcal{Y}})(x) \tag{2.11}$$

is of the type in equation (2.8).
The goal is to design a controller so that the corresponding closed loop dynamics have practical stability properties. The controllers considered in this thesis are all defined on a quantized

set $\mathcal{W}$ and practical stability properties are referred only to the state variable x of the controlled system $\big($i.e., to the continuous component of the overall state (x, w) of the closed loop system (2.10)$\big)$:

Definition 15 (Practical stabilization) *(Cf. [45]) Let Ω, X_0 and X_1 be bounded subsets of $\mathbb{R}^n$ such that Ω and X_0 are neighborhoods of the origin, $\Omega \subseteq X_1$ and $X_0 \subseteq X_1$. The controller (2.9) is said to be (X_0, X_1, Ω)–*stabilizing *iff the corresponding closed loop dynamics (2.10) is so that $\forall\, x(0) \in X_0$ and $\forall\, w(0) \in \mathcal{W}$, $x(t) \in X_1 \;\; \forall\, t \geq 0$ and $\exists\, \bar{t} \in \mathbb{N}$ such that $\forall\, t \geq \bar{t}$, $x(t) \in \Omega$. In this case, system (2.10) is said to be (X_0, X_1, Ω)–*stable.
*A static and time–invariant controller $k \circ q_{\mathcal{Y}}$ is said to be (X_0, Ω)–*stabilizing *iff the closed loop system (2.11) is (X_0, Ω)–stable.*
The set Ω is often referred to as the final *set.*

Remark 2 *To be really precise, the definition of (X_0, X_1, Ω)–stabilizing controller should take into account the fact that the controller (2.9), and hence system (2.10), is time–varying. Nevertheless, most of the controllers we consider in this thesis are time–invariant. The only time–varying controllers are such that $\exists\, T < n$ so that, for $t \geq T$,*

$$\begin{cases} w(t+1) = \gamma\big(w(t), y(t)\big) \\ u(t) = k\big(w(t), y(t)\big). \end{cases}$$

Namely, after a few time instants the controller behaves like a time–invariant system. For this reason, we find unnecessary to introduce a more general definition of stabilizing controller.

Remark 3 (On the steady–state behavior) *Because an open loop unstable system (2.1) with quantized input set $\mathcal{U}$ is not stabilizable in the classical sense, then it is not possible to confine the trajectories within arbitrarily* small *final sets Ω. Thus, for a given system (2.1), it is interesting to evaluate the* optimal *closed loop steady–state behavior that can be achieved or, in other words, to find the* minimal *(in some proper sense) final set Ω. This issue will be investigated with particular attention, starting from the analysis of the minimality properties for controlled invariant neighborhoods of the origin (see Section 3.1.2).*

If $k \circ q_{\mathcal{Y}}$ is (X_0, X_0, Ω)– or (X_0, Ω)–stabilizing, then X_0 (and, in the latter case, also Ω) is positively invariant for the closed loop system. The following lemma makes explicit the relation between positive invariance and $q_{\mathcal{Y}}$–controlled invariance. Thus, with regard to the practical stabilization problem, it clarifies the importance of searching for $q_{\mathcal{Y}}$–controlled invariant sets for system (2.1).

Lemma 5 *Let $X_0 \subset \mathbb{R}^n$, there exists a controller of the type $k \circ q_{\mathcal{Y}}$ such that X_0 is positively invariant for the closed loop system $x^+ = Ax + B(k \circ q_{\mathcal{Y}})(x)$ if and only if X_0 is $q_{\mathcal{Y}}$–controlled invariant.*

Proof. It is a trivial consequence of the definition of $q_{\mathcal{Y}}$–controlled invariance. ■

We conclude this section with the definition of a controller that is often encountered in this thesis and that, especially for single–input systems, plays a central role for the practical stabilization problem. It is the quantized version of the classical *deadbeat* controller:

Definition 16 (The qdb–controller) *Given a quantized input system* $\Sigma(A, B, \mathcal{U})$ *, if* $K \in \mathbb{R}^{m\times n}$ *is such that all the eigenvalues of the matrix* $A + BK$ *are in* 0 *and* $q_{\mathcal{U}}$ *is a nearest neighbor quantizer, then the feedback law*

$$\begin{aligned} k : \ \mathbb{R}^n &\to \mathcal{U} \\ x &\mapsto q_{\mathcal{U}}(Kx) \end{aligned}$$

is called quantized deadbeat controller *(qdb–controller).*

This is exactly the controller we analyzed in Example 4 and, except for the fact that $\mathcal{U}$ was a generalized quantized set, in Example 3 .

2.3.2 Nonlinear behaviors of quantized linear systems

Although the dynamics of the state of system (2.1) is described by a linear transformation, a quantized linear system is a nonlinear system in every respect. Certainly, the input/output relation is nonlinear if measurements are quantized. If full state is available but inputs are quantized, even if the "superposition principle" is still valid, nonlinearity arises from the fact that the inputs are restricted to take values in a nonlinear space (in other words, the operator mapping the input to the state of the system is the restriction of a linear operator to a nonlinear domain) . Therefore, the closed loop dynamics of a quantized linear system is nonlinear and may exhibit typical features of nonlinear dynamics such as the presence of *multiple isolated equilibria*, *limit cycles* and *chaotic behaviors*. For instance, when K in Example 3 is chosen so that $\gamma_s(K) \cdot \gamma_*(\theta) = 1$, the closed loop dynamics has multiple isolated equilibria and/or limit cycles (depending on the sign of a and $a + K$) . Whereas, when $a \in \mathbb{Z}$ ($|a| > 1$) , the closed loop behavior within Ω in case 1 of Example 4 is the prototype of discrete time chaotic dynamics (see [71, 81]) .

2.4 Complexity vs performance

Let us direct our attention towards case 1 and 2 of Example 4 . In both cases, the trajectories converge to the same final set Ω : this property can be expressed by saying that the closed loop dynamics have the same *steady–state performance*. On the other hand, the two closed loop behaviors are different: under the uniform quantization it holds that convergence to the final set is achieved in time $t = 1$; in the case of logarithmic quantization, instead, the time required to converge into Ω is not constant. That is, the closed loop dynamics have different *performance* in the *transient* behavior.
Closed loop performance depend on the quantization scheme. It is intuitive that the more dense (in some proper sense) the control set $\mathcal{U}$ is, the better closed loop performance can

be achieved. Namely, there exists a trade off between *performance* and what will be referred to as the *complexity* of the quantizer. Besides practical stabilization of system (2.1), we are interested in a quantitative study of the relations between quantization and performance. To this end, we need to introduce suitable parameters to measure performance and to characterize quantizers.

Performance can be measured in several ways. For instance, the transient behavior may be evaluated through the decaying rate of the norm of the state of the system or in terms of the mean time taken by the trajectories to reach the final set (in the latter case, see [45]). The steady–state performance can be measured through the *size* of the final set Ω or through the amount of *contraction* realized by the control law[2]. On the contrary, it is less evident how to properly introduce a quantitative description of a quantizer. As a starting point, let us consider the following function: for a given quantizer $q_{\mathcal{U}} : \mathbb{R}^n \to \mathcal{U}$, let $\{\mathcal{C}_u\}_{u\in\mathcal{U}}$ be the partition induced by $q_{\mathcal{U}}$ and assume that it is locally finite; let

$$\begin{aligned} N : \mathbb{R}^+ &\to \mathbb{N} \\ r &\mapsto \#\{u \in \mathcal{U} \,|\, \mathcal{C}_u \cap \mathcal{B}_r \neq \emptyset\}. \end{aligned}$$

We call this function, the *complexity* function associated to the quantizer $q_{\mathcal{U}}$.

The motivation to consider this kind of function and the reason why we associate it to the idea of complexity of a quantizer have an explanation in the framework of networked systems. In fact, consider a quantized input system controlled by a state feedback $q_{\mathcal{U}} : \mathbb{R}^n \to \mathcal{U}$ and assume that the state of the system is transmitted to the controller through a digital communication link. To this end, the state x has to be properly encoded. Actually, to select the right control value, the controller needs only to know the element $\mathcal{C}_u$ of the partition the current state belongs to. Hence, it is sufficient to encode the elements $\{\mathcal{C}_u\}_{u\in\mathcal{U}}$ of the partition rather than the continuous variable x. Moreover, if the state of the system is known to be confined in a bounded set, say $x \in \mathcal{B}_r$ (for some $r \geq 0$), then one has to encode only the elements $\mathcal{C}_u$ intersecting $\mathcal{B}_r$: the function $N(r)$ exactly returns the number of elements to encode.

Example 5 (Complexity vs Performance for logarithmic quantizers) *As an illustrative example, let us consider again case 2 of Example 4. Since the final set is* $\Omega = \left[-\frac{u_0}{2}, \frac{u_0}{2}\right]$, *we let* u_0 *be the parameters measuring the* steady–state performance*: the smaller is* u_0 *the better are the steady–state performance. As for the transient behavior, we have seen that, if* $x(t) \notin \Omega$, *then* $|x(t)| \leq \left(|a|\frac{\theta-1}{\theta+1}\right)^t \cdot |x(0)|$. *Hence, the norm of the state is exponentially decreasing at a rate not smaller than*

$$\mathcal{T}(\theta) := \log \frac{\theta+1}{|a|(\theta-1)}, \tag{2.12}$$

[2]Formal definitions for these concepts will be given when necessary. As an explanatory example, in case 3 of Example 4, $\forall\, \Delta_0 < 12u_0$, the closed loop dynamics is $\left([-\frac{\Delta_0}{2}, \frac{\Delta_0}{2}], [-\frac{u_0}{2}, \frac{u_0}{2}]\right)$–stable: in this case we may say that the size of the final set is u_0 or that the contraction is equal to 12.

that is

$$\text{if} \quad x(t) \notin \Omega\,, \quad |x(t)| \leq e^{-\mathcal{T}(\theta)t}|x(0)|\,.$$

We let $\mathcal{T}(\theta)$ *be the measure of* performance *in the* transient *behavior. The decaying rate* $\mathcal{T}(\theta)$ *is a decreasing function of* θ*: thus, the increase of* θ *is the cause of deterioration of performance in the transient. It holds that*

$$\begin{cases} \lim_{\theta\to 1^+} \mathcal{T}(\theta) = +\infty \\ \lim_{\theta\to\left(\frac{|a|+1}{|a|-1}\right)^-} \mathcal{T}(\theta) = 0\,. \end{cases}$$

Let us compute the complexity *function associated to the logarithmic quantizer* $u(x) = q_{\mathcal{U}}(x)$ *of parameters* (u_0, θ)*. The state space partition induced by* $q_{\mathcal{U}}$ *is made of intervals such that the set of the extremes of these intervals is*

$$S(u_0,\theta) := \left\{ \pm\frac{u_0}{2|a|} \right\} \cup \left\{ \pm\frac{u_0(\theta+1)}{2|a|}\theta^h \,\middle|\, h \in \mathbb{N} \right\} \subset \mathbb{R}\,.$$

By the expression of $S(u_0,\theta)$*, it is an easy computation to see that, for sufficiently large* r*, the qualitative behavior of the function* $N(r)/2$ *is*[3]

$$\frac{N(r)}{2} \simeq \frac{\log\left(\frac{2|a|r}{u_0(\theta+1)}\right)}{\log\theta} := \widetilde{N}(r)\,.$$

Inverting equation (2.12) *and substituting for* θ *in the expression of* $\widetilde{N}$*, we obtain*

$$\widetilde{N}(\mathcal{T}\,, u_0) = \frac{\log\frac{(|a|e^{\mathcal{T}}-1)r}{u_0 e^{\mathcal{T}}}}{\log\frac{|a|e^{\mathcal{T}}+1}{|a|e^{\mathcal{T}}-1}}\,. \tag{2.13}$$

This equation establishes the connection and the trade off between the complexity of the quantizer and the performance in both the steady–state and the transient behavior. Let us comment the main features of equation (2.13)*: assume that* $r > 0$ *is fixed and sufficiently large,*

1. *Given* u_0 *and* $|a|$*,* $\widetilde{N}$ *is an increasing function of* $\mathcal{T}$*. Moreover,* $\lim_{\mathcal{T}\to+\infty}\widetilde{N}(\mathcal{T}) = +\infty$ *and*

   $$\text{for} \quad \mathcal{T}\to+\infty\,, \quad \widetilde{N}(\mathcal{T}) \sim C_1(u_0,|a|,r)\cdot e^{\mathcal{T}}\,,$$

   where $C_1(u_0,|a|,r) = \frac{|a|}{2}\log\left(\frac{|a|r}{u_0}\right)$*. That is, the complexity grows as performance in the transient improve and, asymptotically, is exponentially divergent.*

2. *Given* $\mathcal{T}$ *and* $|a|$*,* $\widetilde{N}$ *is a decreasing function of* u_0*. Moreover,* $\lim_{u_0\to 0^+}\widetilde{N}(u_0) = +\infty$ *and*

   $$\text{for} \quad u_0\to 0^+\,, \quad \widetilde{N}(u_0) \sim -C_2(\mathcal{T},|a|)\cdot\log u_0\,,$$

   where $C_2(\mathcal{T},|a|) = 1/\log\frac{|a|e^{\mathcal{T}}+1}{|a|e^{\mathcal{T}}-1}$*. That is, complexity grows as performance in the steady–state improve and, at the vanishing of* u_0*, is logarithmically divergent.*

[3]The function $N(r)/2$ contains all the information because the partition is symmetric with respect to the origin. The choice of considering $N(r)/2$ is consistent with the treatment of this topic that will be developed in Chapter 7.

3. *Given* $\mathcal{T}$ *and* u_0, $\widetilde{N}$ *is an increasing function of* $|a|$. *Moreover,* $\lim_{|a|\to+\infty} \widetilde{N}(|a|) = +\infty$ *and*

$$\text{for} \quad |a| \to +\infty\,, \quad \widetilde{N}(a) \sim \frac{e^{\mathcal{T}}}{2} \cdot |a| \log |a|\,.$$

That is, complexity grows with the instability of the open loop system and, asymptotically, is quasi–linearly divergent. ♣

The definition of the complexity function N has been motivated with reference to the coding issue. Actually, it is relevant not only the number of symbols to be encoded, but also the length of the coding sequences: in fact, if the communication link in the control loop can transmit at a finite rate, then long coding sequences are the cause of delay and hence of performance deterioration. If the model includes a statistics on the cells, then the average length of the coding sequences can be minimized by encoding the most probable cells with the shortest sequences, and vice versa. Hence, a more suited complexity function should be defined by taking into account the underlying statistics. To this end, the right mathematical framework is that of the *Information theory* and the right notion to be considered is that of *entropy*. This issue, and the corresponding analysis of complexity vs performance, will be addressed in Chapter 7.

Chapter 3

Analysis

In this chapter, the problem of the search for q_y–controlled invariant sets for system (2.1) is considered. Although there exists a wide literature on controlled invariance (see [11] and references therein), the problem for quantized input systems is not trivial. Indeed, most of the results on constrained control are limited to bounded convex sets and hence do not apply to the quantized control case. The main results available for quantized systems are based on algorithmic procedures: we mention, inter alia, the "Controlled Invariance Kernel Algorithm" [113] in the framework of the Viability theory [3], the "Inclusion Principle" [108] and methods based on nonlinear programming [118]. While these approaches are of quite general application, on the other hand they are affected by the limitations due to computational complexity and, above all, they typically yield conservative results. That is, in a (X_0, Ω)–stabilization problem, it is desirable to find a *small* final invariant set Ω but the aforementioned methods are not capable of providing information on the *minimal* invariant set for a system under assigned input quantization.

In this chapter, we propose two analytical methods. The first one is completely original and deals with the controlled invariance analysis of hypercubes: although it can be applied to a restricted class of systems (i.e., reachable single–input systems), on the other hand it has some features making this approach quite appealing. In fact, the analysis is really simple to handle and, for this reason, it is suitable to deal with cases that, in the current literature, have been faced only marginally: namely, the analysis (and, in next Chapter 4, also the synthesis) for systems under arbitrarily assigned input and output quantization. Moreover, in many interesting cases (e.g., in the presence of uniform or logarithmic input quantization), the family of invariant hypercubes contains an element which is, in a precise sense, the smallest one with respect to any other controlled invariant neighborhood of the equilibrium.
The second approach relies on classical tools derived by stability analysis based on quadratic Lyapunov functions and returns controlled invariant ellipsoids: this method can be applied to the more general class of stabilizable multi–input systems but only the quantized input case has been considered. Moreover, results are often quite conservative. Anyhow, this approach turns out to be very useful also in the context of control synthesis for practical stabilization

(see Chapter 5) .

The chapter is organized as follows: Section 3.1 deals with the controlled invariance analysis for quantized input systems in the full state case. The analysis based on hypercubes is introduced in Section 3.1.1 ; in subsequent Section 3.1.2 , the minimality properties of hypercubes are proved; Section 3.1.3 is devoted to an extension of the results presented in the previous sections to a problem in the framework of the *control under communication constraints*. The invariance analysis based on ellipsoid is described in Section 3.1.4 . In Section 3.2 , the analysis of controlled invariant hypercubes is extended to the case of systems with quantized single–input and quantized measurements: Section 3.2.1 deals with the quantized state case; in Section 3.2.2 , quantized outputs are considered.

3.1 Controlled invariance: quantized input

In this section we study the controlled invariance problem for system (2.1) in the full state case. Hence, we consider a system $\Sigma(A, B, \mathcal{U})$ where only the input set is quantized.

3.1.1 Controlled invariant hypercubes: single–input

Let us suppose that system (2.1) is single–input and that the pair (A, B) is reachable. In this case, under the general assumption that $\mathcal{U}$ is a quantized set, it is possible to find controlled invariant sets within a particularly simple class of polytopes. Namely, hypercubes in the *controller form* coordinates.
Because of the reachability assumption, throughout this section it is assumed without loss of generality (see [114]) that the system is represented in the controller form coordinates, that is:

A0)

$$A = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}, \; B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{3.1}$$

where $z^n - a_n z^{n-1} - \cdots - a_2 z - a_1$ is the characteristic polynomial of A .
In the controller form coordinates, $\|A\|_\infty = \max\limits_{i=1,\ldots,n} \sum_{j=1}^{n} |A_{i,j}| = \max\{1 \, , \, \sum_{i=1}^{n} |a_i|\}$. Let

$$\alpha := \sum_{i=1}^{n} |a_i| \, .$$

If $\alpha \leq 1$, then the system is open loop stable: in fact, $\forall\, x \in \mathbb{R}^n$, $\|Ax\|_\infty \leq \|x\|_\infty$. In particular, $\forall\, \Delta > 0$, the hypercube $Q_n(\Delta)$ is controlled invariant. Hence, the interesting case to study the invariance problem is $\alpha > 1$. Nevertheless, in anticipation of the stabilization problem, it is convenient to state some results under the more general assumption $\alpha \geq 1$.

Remark 4 (The qdb–controller) *Notice that* $K = (\,-a_1\ \cdots\ -a_n\,)$ *is the unique* $K \in \mathbb{R}^{1\times n}$ *such that the controller* $k(x) := q_{\mathcal{U}}(Kx)$, *where* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *is a nearest neighbor input quantizer, is a qdb–controller. Thus, for single–input reachable systems* $\Sigma(A, B, \mathcal{U})$ *in the controller form coordinates, a qdb–controller takes the form*

$$k(x) = q_{\mathcal{U}}\big(-\textstyle\sum_{i=1}^{n} a_i x_i\big) = q_{\mathcal{U}}\big(-(Ax)_n\big) \tag{3.2}$$

and the closed loop dynamics $x^+ = Ax + Bk(x)$ *is such that*

$$\begin{cases} x_i^+ = x_{i+1} \quad \textit{for} \quad i = 1, \dots, n-1 \\ x_n^+ = q_{\mathcal{U}}\big(-(Ax)_n\big) + (Ax)_n = q_e\big(-(Ax)_n\big)\,. \end{cases} \tag{3.3}$$

The analysis of controlled invariance for hypercubes $Q_n(\Delta)$ in controller form coordinates is particularly simple. In fact, given $\Delta > 0$, let $x \in Q_n(\Delta)$ and $u \in \mathcal{U}$: by the controller form of (A, B),

$$x^+ = \big(x_2, \dots, x_n, \textstyle\sum_i a_i x_i + u\big) \in Q_n(\Delta) \;\Leftrightarrow\; \big|\textstyle\sum_i a_i x_i + u\big| \le \frac{\Delta}{2}\,. \tag{3.4}$$

Thus, for hypercubes $Q_n(\Delta)$ in controller form coordinates, invariance can be tested considering the n–th component only. We seek a characterization of controlled invariant hypercubes in terms of algebraic relations between quantities related to the dynamics of system (2.1) and to the control set $\mathcal{U}$.
To this end, we first notice that the controlled invariance of a given hypercube $Q_n(\Delta)$ only depends on a bounded subset of the whole control set $\mathcal{U}$, indeed:

Lemma 6 *Consider system* (2.1), *assume* **A0** *and that* $\alpha \ge 1$. *If* $x \in Q_n(\Delta)$ *and* u *is such that* $x^+ \in Q_n(\Delta)$, *then* $u \in \big[-\frac{\Delta}{2}(\alpha+1)\,,\ \frac{\Delta}{2}(\alpha+1)\big]$.

Proof. Since $\|A\|_\infty = \alpha$ and A is in controller form, then

$$\mathrm{Pr}_n\big(AQ_n(\Delta)\big) = \big[-\tfrac{\Delta}{2}\,\alpha\,,\ \tfrac{\Delta}{2}\,\alpha\big]\,. \tag{3.5}$$

Now, $x^+ \in Q_n(\Delta)$ implies that $-\frac{\Delta}{2} \le (Ax)_n + u \le \frac{\Delta}{2}$, namely $-\frac{\Delta}{2} - (Ax)_n \le u \le \frac{\Delta}{2} - (Ax)_n$ which implies $-\frac{\Delta}{2}(\alpha+1) \le u \le \frac{\Delta}{2}(\alpha+1)$ because of (3.5). ■

Hence, the set of control values that are relevant to ensure the invariance of $Q_n(\Delta)$ is

$$\mathcal{U}(\Delta) := \mathcal{U} \cap \big[-\tfrac{\Delta}{2}(\alpha+1)\,,\ \tfrac{\Delta}{2}(\alpha+1)\big]\,. \tag{3.6}$$

It holds that $\mathcal{U}(\Delta) \neq \emptyset$ (in fact, it contains 0) and, by Lemma 1.*i*, $\mathcal{U}(\Delta)$ is a finite set. The analysis of controlled invariant hypercubes is carried out in terms of the following scalar functions of the edge length Δ: let

$$\begin{cases} m(\Delta) := \ \min \mathcal{U}(\Delta) \\ M(\Delta) := \ \max \mathcal{U}(\Delta) \end{cases} \tag{3.7}$$

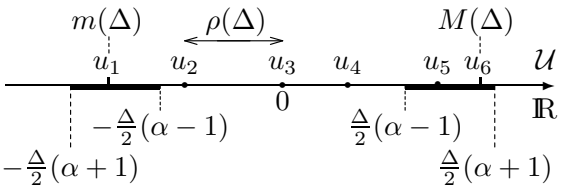

Figure 3.1: $\mathcal{U}(\Delta) = \{m(\Delta) = u_1, u_2, u_3, u_4, u_5, u_6 = M(\Delta)\}$: $\rho(\Delta) = u_3 - u_2$, the thicker segments represent the intervals where $m(\Delta)$ and $M(\Delta)$ satisfy inequalities (3.11a–b).

and, according to (2.2), let

$$\rho(\Delta) := \rho_{\mathcal{U}(\Delta)} = \begin{cases} \sup \quad \big\{ b-a \,\big|\,]a,b[\subseteq [m(\Delta), M(\Delta)] \quad \text{and} \\ \qquad\quad]a,b[\cap \mathcal{U}(\Delta) = \emptyset \big\} \quad \text{if } \#\mathcal{U}(\Delta) > 1 \\ +\infty \quad \text{otherwise} \end{cases} \tag{3.8}$$

be the dispersion of $\mathcal{U}(\Delta)$ (see Fig. 3.1). The three functions $m(\Delta)$, $M(\Delta)$, $\rho(\Delta)$ depend on the dynamics of the system only through the infinity norm of A. The function $M(\Delta)$ is piecewise constant and non–decreasing with Δ, whilst $m(\Delta)$ is piecewise constant and non–increasing. Given $\mathcal{U} \neq \{0\}$, let

$$\bar{\Delta} := \frac{2u_0}{\alpha+1} \tag{3.9}$$

(where u_0 is the resolution at 0 of $\mathcal{U}$, see Definition 10), then: for $\Delta < \bar{\Delta}$, $\rho(\Delta) = +\infty$; for $\Delta \geq \bar{\Delta}$, $\rho(\Delta)$ is piecewise constant, right continuous and non–decreasing with Δ. Since $\rho(\bar{\Delta}) = u_0$ and $u_0 \geq \bar{\Delta}$, then

$$\rho : [\bar{\Delta}, +\infty[\to [\bar{\Delta}, +\infty[. \tag{3.10}$$

Further details on the behavior of these functions are given later on in Remark 5.

Theorem 1 (Controlled invariant hypercubes) *Consider system $\Sigma(A,B,\mathcal{U})$, assume* **A0** *and that $\alpha = \|A\|_\infty > 1$. For $\Delta > 0$, $Q_n(\Delta)$ is controlled invariant if and only if*

$$\begin{cases} m(\Delta) \leq -\dfrac{\Delta}{2}\big(\alpha-1\big) & \text{(3.11a)} \\ M(\Delta) \geq \dfrac{\Delta}{2}\big(\alpha-1\big) & \text{(3.11b)} \\ \rho(\Delta) \leq \Delta\,. & \text{(3.11c)} \end{cases}$$

The following result is useful to prove the theorem and clarifies the role of the qdb–controller for the invariance problem.

Proposition 1 *Consider system $\Sigma(A,B,\mathcal{U})$, assume* **A0** *and let $k : \mathbb{R}^n \to \mathcal{U}$ be a qdb–controller. For $\Delta > 0$, $Q_n(\Delta)$ is controlled invariant if and only if it is positively invariant for the closed loop system $x^+ = Ax + Bk(x)$.*

Proof. By condition (3.4), $Q_n(\Delta)$ is controlled invariant if and only if $\forall\, x \in Q_n(\Delta)$, $\exists\, u \in \mathcal{U}$ such that $\big|(Ax)_n + u\big| \le \frac{\Delta}{2}$. By definition of nearest neighbor quantizer,

$$\operatorname*{argmin}_{u\in\mathcal{U}} \big|(Ax)_n + u\big| = q_{\mathcal{U}}\big(-(Ax)_n\big)\,.$$

As $q_{\mathcal{U}}\big(-(Ax)_n\big) = k(x)$, the thesis follows. ■

Proof of Theorem 1. Let us prove the necessity of (3.11a): let $x \in Q_n(\Delta)$ be such that $(Ax)_n = \frac{\Delta}{2}\alpha$ (see equation (3.5)). If $Q_n(\Delta)$ is controlled invariant, then $\exists\, u \in \mathcal{U}$ such that $\frac{\Delta}{2} \ge \big|(Ax)_n + u\big| = \big|\frac{\Delta}{2}\alpha + u\big|$. That is, $\exists\, u \in \mathcal{U}$ such that $-\frac{\Delta}{2}(\alpha+1) \le u \le -\frac{\Delta}{2}(\alpha-1)$: by definition of $m(\Delta)$, this means that $m(\Delta) \le -\frac{\Delta}{2}(\alpha-1)$ (see Fig. 3.1).
The necessity of (3.11b) can be proved similarly by considering $x \in Q_n(\Delta)$ such that $(Ax)_n = -\frac{\Delta}{2}\alpha$.
To prove the necessity of (3.11c) we argue by contradiction: if $\rho(\Delta) = +\infty$, then $\mathcal{U}(\Delta) = \{0\}$ but $\|A\|_\infty = \alpha > 1$ implies $AQ_n(\Delta) \not\subseteq Q_n(\Delta)$ which contradicts the invariance of $Q_n(\Delta)$. If instead $\Delta < \rho(\Delta) < +\infty$, then $\exists\, u_1 \in \mathcal{U}(\Delta)$ and $u_2 \in \mathcal{U}(\Delta)$ such that $u_2 - u_1 > \Delta$ and $]u_1\,,\, u_2[\cap \mathcal{U} = \emptyset$. Let $w := \frac{u_1+u_2}{2}$, $w \in \mathrm{Pr}_n\big(AQ_n(\Delta)\big) = \big[-\frac{\Delta}{2}\,\alpha\,,\, \frac{\Delta}{2}\,\alpha\big]$ because $u_2 - u_1 > \Delta$ and, by equation (3.6), $u_1 \ge -\frac{\Delta}{2}(\alpha+1)$ and $u_2 \le \frac{\Delta}{2}(\alpha+1)$. Hence, $\exists\, \tilde{x} \in Q_n(\Delta)$ such that $(A\tilde{x})_n = -w$. Let $q_{\mathcal{U}}$ be a nearest neighbor quantizer: by construction, $|q_{\mathcal{U}}(w) - w| > \frac{\Delta}{2}$, but

$$\big|q_{\mathcal{U}}(w) - w\big| = \big|q_{\mathcal{U}}\big(-(A\tilde{x})_n\big) + (A\tilde{x})_n\big| \overset{\text{(a)}}{=} |\tilde{x}_n^+| \overset{\text{(b)}}{\le} \frac{\Delta}{2}\,,$$

where equality (a) holds by equation (3.3) and inequality (b) follows by Proposition 1.
Finally, let us show that the validity of inequalities (3.11) is a sufficient condition for the controlled invariance of $Q_n(\Delta)$. According to condition (3.4) and Proposition 1, let us show that $\forall\, x \in Q_n(\Delta)$, $\big|q_{\mathcal{U}}\big(-(Ax)_n\big) + (Ax)_n\big| \le \frac{\Delta}{2}$. If x is such that $-(Ax)_n \in \big[m(\Delta)\,,\, M(\Delta)\big]$, then $\big|q_{\mathcal{U}}\big(-(Ax)_n\big) + (Ax)_n\big| \le \frac{\rho(\Delta)}{2}$ by definition of $\rho(\Delta)$. If instead $-(Ax)_n < m(\Delta)$, then

$$\big|q_{\mathcal{U}}\big(-(Ax)_n\big) + (Ax)_n\big| \overset{\text{(c)}}{\le} \big|m(\Delta) + (Ax)_n\big| = m(\Delta) + (Ax)_n \overset{\text{(d)}}{\le} -\tfrac{\Delta}{2}(\alpha-1) + \tfrac{\Delta}{2}\alpha = \tfrac{\Delta}{2}\,,$$

where in inequality (c) we used the fact that $\operatorname*{argmin}_{u\in\mathcal{U}} \big|u + (Ax)_n\big| = q_{\mathcal{U}}\big(-(Ax)_n\big)$ and in inequality (d) we used inequality (3.11a) and the fact that $(Ax)_n \le \frac{\Delta}{2}\alpha$ (see equation (3.5)).
The case $-(Ax)_n > M(\Delta)$ is similar. ■

Before presenting some examples, let us briefly describe the way to compute the functions $\rho(\Delta)$, $M(\Delta)$ and $m(\Delta)$.

Remark 5 (Computation of $\rho(\Delta)$, $M(\Delta)$ and $m(\Delta)$) *For the sake of simplicity, we consider the case of a control set $\mathcal{U}$ symmetric with respect to the origin, the extension to the general case is straightforward. Consider a system $\Sigma(A,B,\mathcal{U})$ in the controller form coordinates with $\alpha \ge 1$ and*

$$\mathcal{U} = \{0\} \cup \{\pm u_0\,, \pm u_1\,, \pm u_2\,, \dots\}\,,$$

where $0 < u_0 < u_1 < u_2 < \cdots$. *Let* $\mathbb{K} := \{k \in \mathbb{N} \,|\, u_k \in \mathcal{U}\}$ *be the set of indices for the elements of* $\mathcal{U} \setminus \{0\}$ *(thus, either* $\mathbb{K} = \{0, 1, \dots, N\}$ *or* $\mathbb{K} = \mathbb{N}$*, depending on whether* $\mathcal{U}$ *is a finite set or not).*
First, compute $\alpha = \|A\|_\infty$.

• **Computation of $M(\Delta)$ and $m(\Delta)$:** *since* $\mathcal{U}$ *is symmetric with respect to the origin,*

$$\forall\, \Delta > 0\,, \quad m(\Delta) = -M(\Delta)\,.$$

Let us compute $M(\Delta)$*: this function is piecewise constant, non–decreasing and right continuous. According to equations* (3.6) *and* (3.7)*, the set of the discontinuity points of* $M(\Delta)$ *is* $\mathcal{J}_M := \left\{\Delta > 0 \,|\, \frac{\Delta}{2}(\alpha+1) = u_k \text{ for some } k \in \mathbb{K}\right\} = \left\{\frac{2u_k}{\alpha+1} \,|\, k \in \mathbb{K}\right\}$. *Therefore, for* $\Delta > 0$,

$$M(\Delta) = \begin{cases} 0 & \text{if} \quad \Delta \in \left]\, 0, \frac{2u_0}{\alpha+1} \,\right[\\ u_k & \text{if} \quad \Delta \in \left[\, \frac{2u_k}{\alpha+1}, \frac{2u_{k+1}}{\alpha+1} \,\right[\end{cases} \tag{3.12}$$

$\big($*if* $\mathcal{U}$ *is finite and* $\mathbb{K} = \{0, 1, \dots, N\}$*, then* $M(\Delta) = u_N \;\; \forall\, \Delta \geq \frac{2u_N}{\alpha+1}\big)$.

• **Computation of $\rho(\Delta)$:** *also* $\rho(\Delta)$ *is piecewise constant, non–decreasing and right continuous but, while the set* $\mathcal{J}_M$ *of the discontinuity points of* $M(\Delta)$ *is depending on all the positive values of* $\mathcal{U}$*, the set* $\mathcal{J}_\rho$ *of the discontinuity points of* $\rho(\Delta)$ *only depends on a subset of the positive control values (which are referred to as* critical *control values). In fact,* $\rho(\Delta)$ *is discontinuous in correspondence of those values of* Δ *such that* $\mathcal{U}(\Delta)$ *includes a new control value* u_{k_c} *whose distance from* u_{k_c-1} *is larger than the dispersion of* $\mathcal{U} \cap [-u_{k_c-1}, u_{k_c-1}]$.
In formulae, let $u_{-1} := 0$ *and*

$$\begin{aligned} \mathbb{K}_c := \;& \{0\} \cup \{k \in \mathbb{K} \setminus \{0\} \,|\, \forall\, i < k\,,\; u_k - u_{k-1} > u_i - u_{i-1}\} \\ := \;& \{k_0\,, k_1\,, k_2\,, \dots\} \subseteq \mathbb{K} \end{aligned}$$

be the set of indices corresponding to the critical control values $(0 = k_0 < k_1 < k_2 < \cdots)$. *Then,* $\mathcal{J}_\rho = \left\{\frac{2u_k}{\alpha+1} \,|\, k \in \mathbb{K}_c\right\}$. *Therefore, for* $\Delta > 0$,

$$\rho(\Delta) = \begin{cases} +\infty & \text{if} \quad \Delta < \bar{\Delta} = \frac{2u_0}{\alpha+1} \\ u_0 & \text{if} \quad \Delta \in \left[\, \frac{2u_0}{\alpha+1}, \frac{2u_{k_1}}{\alpha+1} \,\right[\\ u_{k_i} - u_{k_i-1} & \text{if} \quad \Delta \in \left[\, \frac{2u_{k_i}}{\alpha+1}, \frac{2u_{k_{i+1}}}{\alpha+1} \,\right[\end{cases} \tag{3.13}$$

$\big($*if* $\mathbb{K}_c$ *is a finite set, say* $\mathbb{K}_c = \{k_0\,, \dots,\, k_N\}$*, then* $\rho(\Delta) = u_{k_N} - u_{k_N-1} \;\; \forall\, \Delta \geq \frac{2u_{k_N}}{\alpha+1}\big)$.
Notice that, $\forall\, \Delta_0 > \bar{\Delta}$*, the function* ρ *takes only a finite number of values over the interval* $[\bar{\Delta}\,, \Delta_0]$ *(it is a consequence of the fact that* $0 \in \mathcal{U}$ *is an isolated point).*

Example 6 (Finite control set) *Consider the quantized input system*

$$\begin{cases} x^+ = \begin{pmatrix} 0 & 1 \\ 5/4 & 1/4 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ u \in \mathcal{U}\,, \end{cases}$$

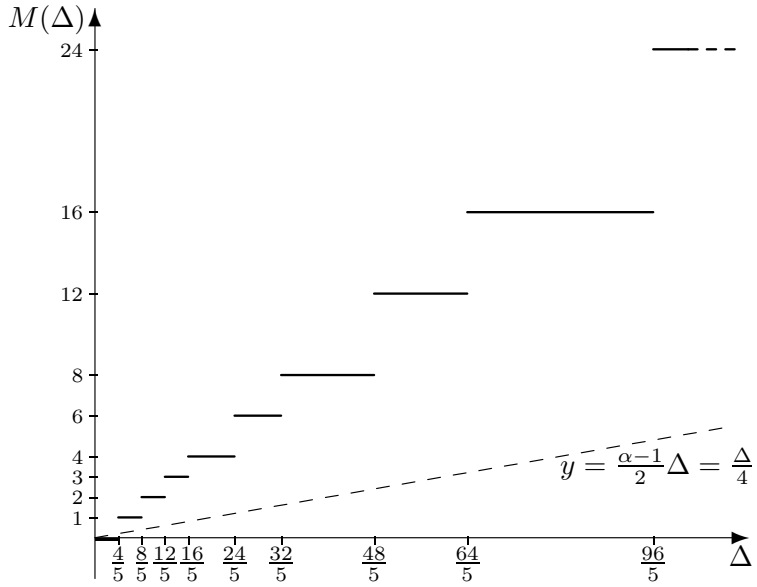

Figure 3.2: Plot of $M(\Delta)$ for the system in Example 6 .

where $\mathcal{U} = \{0\,,\pm1\,,\pm2\,,\pm3\,,\pm4\,,\pm6\,,\pm8\,,\pm12\,,\pm16\,,\pm24\}$. *Let us determine the controlled invariant hypercubes by applying Theorem 1 . Notice that, since* $|\det A| = 5/4 > 1$, *the system is open loop unstable.*
Let us compute the functions $M(\Delta)$ *and* $\rho(\Delta)$ $\big(m(\Delta) = -M(\Delta)\big)$.
The infinity norm of A *is* $\alpha = \frac{3}{2}$.
The function $M(\Delta)$ *is determined by simple substitution of the numerical values of* α *and* u_k *in equation* (3.12) . *The plot of* $M(\Delta)$ *is reported in Fig. 3.2 , here we report only its discontinuity points:* $\mathcal{J}_M = \left\{\frac{2u_k}{\alpha+1}\,|\,k \in \mathbb{K}\right\} = \left\{\frac{4}{5}\,,\frac{8}{5}\,,\frac{12}{5}\,,\frac{16}{5}\,,\frac{24}{5}\,,\frac{32}{5}\,,\frac{48}{5}\,,\frac{64}{5}\,,\frac{96}{5}\right\}$. *Conditions (3.11a–b) are satisfied for* $\Delta \in \left[\,\frac{4}{5}\,,96\,\right]$.
As far as $\rho(\Delta)$ *is concerned, the set of the critical control values is* $\{1\,,6\,,12\,,24\} \subset \mathcal{U}$. *Thus* $\mathcal{J}_\rho = \left\{\frac{4}{5}\,,\frac{24}{5}\,,\frac{48}{5}\,,\frac{96}{5}\right\}$ *and, by equation* (3.13) ,

$$\rho(\Delta) = \begin{cases} +\infty & \textit{if} \quad \Delta < \bar{\Delta} = \frac{4}{5} \\ 1 & \textit{if} \quad \Delta \in \left[\,\frac{4}{5}, \frac{24}{5}\,\right[\\ 2 & \textit{if} \quad \Delta \in \left[\,\frac{24}{5}, \frac{48}{5}\,\right[\\ 4 & \textit{if} \quad \Delta \in \left[\,\frac{48}{5}, \frac{96}{5}\,\right[\\ 8 & \textit{if} \quad \Delta \geq \frac{96}{5}\,, \end{cases} \tag{3.14}$$

see the plot in Fig. 3.3 . In particular, condition (3.11c) *is satisfied* $\forall\,\Delta \geq 1$.
Therefore, $Q_2(\Delta)$ *is controlled invariant if and only if* $1 \leq \Delta \leq 96$. ♣

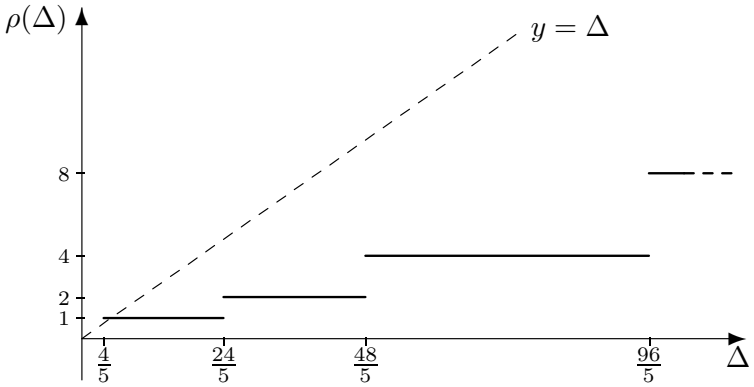

Figure 3.3: Plot of $\rho(\Delta)$ for the system in Example 6 .

Example 7 (Uniformly quantized controls) *Given a system* $\Sigma(A, B, \mathcal{U})$ *in the controller form coordinates with* $\alpha = \|A\|_\infty > 1$, *if* $\mathcal{U}$ *is uniformly quantized with parameter* u_0 , *then* $\forall\, \Delta \geq u_0$, *the hypercube* $Q_n(\Delta)$ *is controlled invariant. To show this, let us apply Theorem 1 .*
In this case, a closed formula for $M(\Delta)$ *can be provided, in fact: for* $\Delta > 0$, $M(\Delta) = ku_0$, *where* k *is the largest integer such that* $ku_0 \leq \frac{\Delta}{2}(\alpha + 1)$ *(see equations* (3.6) *and* (3.7)*) . Hence,*

$$M(\Delta) = -m(\Delta) = \left\lfloor \frac{\Delta(\alpha+1)}{2u_0} \right\rfloor \cdot u_0 \,.$$

Since $M(\Delta) > \left(\frac{\Delta(\alpha+1)}{2u_0} - 1\right)u_0 = \frac{\Delta\alpha+\Delta-2u_0}{2}$, *then*

$$\forall\, \Delta \geq u_0 \,, \quad M(\Delta) > \frac{\Delta}{2}(\alpha - 1)$$

and both inequalities (3.11a) *and* (3.11b) *are satisfied.*
As for $\rho(\Delta)$, *the unique critical control value is* u_0 , *therefore*

$$\rho(\Delta) = \begin{cases} +\infty & \text{if} \quad \Delta < \bar{\Delta} = \frac{2u_0}{\alpha+1} \\ u_0 & \text{if} \quad \Delta \geq \frac{2u_0}{\alpha+1} \,. \end{cases}$$

Since $\alpha > 1$, *then* $\bar{\Delta} < u_0$ *and inequality* (3.11c) *is satisfied* $\forall\, \Delta \geq u_0$.
The plots of $M(\Delta)$ *and* $\rho(\Delta)$ *are reported in Fig. 3.4 .* ♣

Remark 6 *Notice that the commonly used inequality* $\lfloor x \rfloor > x - 1$ *is* tight. *That is, for any arbitrarily small* $\epsilon > 0$, *it holds that* $\lfloor x \rfloor - (x - 1) = \epsilon$ *whenever* $x = z - \epsilon$, *with* $z \in \mathbb{Z}$ *and* $0 < \epsilon < 1$. *The same holds for the inequality* $\lceil x \rceil < x + 1$.

Example 8 (Logarithmically quantized controls) *Given a system* $\Sigma(A, B, \mathcal{U})$ *in the controller form coordinates with* $\alpha = \|A\|_\infty > 1$, *if* $\mathcal{U}$ *is logarithmically quantized with*

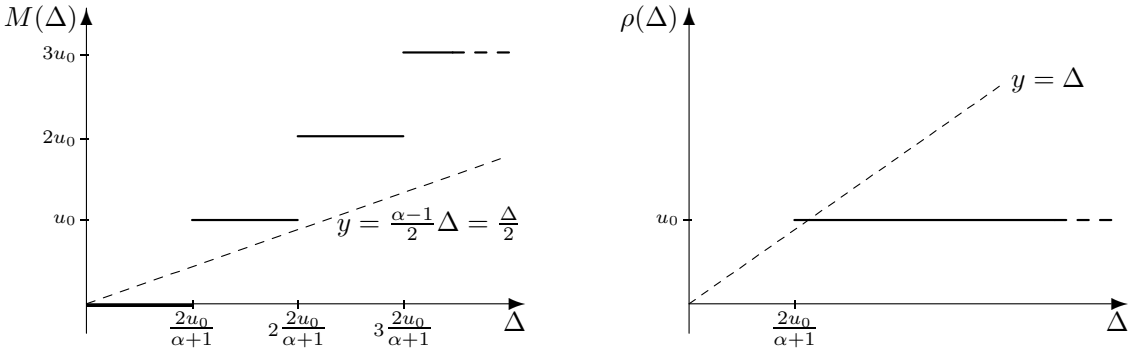

Figure 3.4: Plot of $M(\Delta)$ and of $\rho(\Delta)$ for a system with $\alpha = 2$ and uniformly quantized controls (see Example 7).

parameters (u_0, θ) *and* $1 < \theta \leq \frac{\alpha+1}{\alpha-1}$, *then* $\forall\, \Delta \geq u_0$, $Q_n(\Delta)$ *is controlled invariant. To show this, let us apply Theorem 1.*
We first compute the functions $M(\Delta)$, $m(\Delta)$ *and* $\rho(\Delta)$ *in the general case* $\theta > 1$. *Also in this case, a closed formula for* $M(\Delta)$ *can be provided, in fact: according to equation* (3.12), $M(\Delta) = 0$ *for* $0 < \Delta < \frac{2u_0}{\alpha+1}$, *whilst for* $\Delta \geq \frac{2u_0}{\alpha+1}$, $M(\Delta) = u_0\theta^k$, *where* k *is the largest integer such that* $u_0\theta^k \leq \frac{\Delta}{2}(\alpha+1)$ *(see equations* (3.6) *and* (3.7)*). Hence,*

$$M(\Delta) = -m(\Delta) = \begin{cases} 0 & \text{if} \quad 0 < \Delta < \frac{2u_0}{\alpha+1} \\ u_0 \cdot \theta^{\left\lfloor \log_\theta\left(\frac{(\alpha+1)\,\Delta}{2u_0}\right)\right\rfloor} & \text{if} \quad \Delta \geq \frac{2u_0}{\alpha+1}\,. \end{cases}$$

As far as $\rho(\Delta)$ *is concerned,*

$$\rho(\Delta) = \begin{cases} +\infty & \text{if} \quad \Delta < \bar{\Delta} = \frac{2u_0}{\alpha+1} \\ u_0 & \text{if} \quad \frac{2u_0}{\alpha+1} \leq \Delta < \frac{2u_0}{\alpha+1}\,\theta^{\left\lfloor \log_\theta\left(\frac{\theta^2}{\theta-1}\right)\right\rfloor} \\ u_0\frac{\theta-1}{\theta} \cdot \theta^{\left\lfloor \log_\theta\left(\frac{(\alpha+1)\,\Delta}{2u_0}\right)\right\rfloor} & \text{if} \quad \Delta \geq \frac{2u_0}{\alpha+1}\,\theta^{\left\lfloor \log_\theta\left(\frac{\theta^2}{\theta-1}\right)\right\rfloor} \end{cases} \tag{3.15}$$

(the details of the computations are reported in Appendix A.2.1).
Now, let us assume that $1 < \theta \leq \frac{\alpha+1}{\alpha-1}$ *and let us show that conditions* (3.11) *are satisfied* $\forall\, \Delta \geq u_0$ *: for* $\Delta \geq \frac{2u_0}{\alpha+1}$,

$$M(\Delta) > u_0 \cdot \theta^{\log_\theta\left(\frac{(\alpha+1)\,\Delta}{2u_0}\right)-1} = \frac{\alpha+1}{2\,\theta}\,\Delta \geq \frac{\Delta}{2}(\alpha-1)$$

because $\theta \leq \frac{\alpha+1}{\alpha-1}$ *and* $\alpha > 1$. *In particular, conditions (3.11a–b) are satisfied for* $\Delta \geq u_0$.
As for condition (3.11c), *for* $\Delta \geq \frac{2u_0}{\alpha+1}\,\theta^{\left\lfloor \log_\theta\left(\frac{\theta^2}{\theta-1}\right)\right\rfloor}$, *by removing the floor function in the expression of* $\rho(\Delta)$, *we have*

$$\rho(\Delta) \leq \frac{(\theta-1)(\alpha+1)}{2\,\theta}\,\Delta \leq \Delta$$

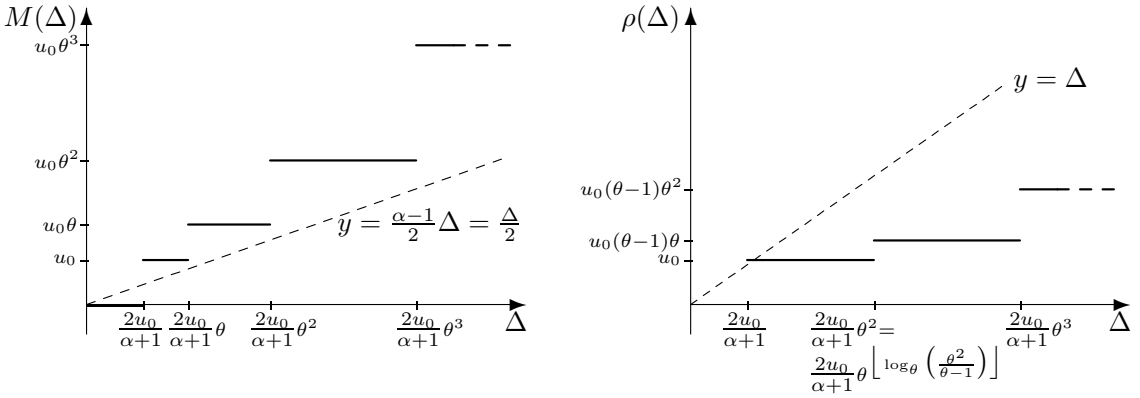

Figure 3.5: Plot of $M(\Delta)$ and of $\rho(\Delta)$ for a system with $\alpha = 2$ and logarithmically quantized controls with parameter $(u_0, \theta = 1.8)$ (see Example 8).

because, for $\theta > 0$ *and* $\alpha > 1$, $\frac{(\theta-1)(\alpha+1)}{2\theta} \leq 1 \Leftrightarrow \theta \leq \frac{\alpha+1}{\alpha-1}$. *Hence, a fortiori,* $\forall\, \Delta \geq u_0$, $\rho(\Delta) \leq \Delta$ *(see also Fig. 3.5).* ♣

Remark 7 *If* $1 < \theta < \frac{\alpha+1}{\alpha-1}$, *by the above computations it follows that*

$$\begin{cases} \rho(u_0) = u_0 \\ \rho(\Delta) < \Delta \qquad\qquad \textit{if} \quad \Delta > u_0 \\ M(\Delta) > \frac{\Delta}{2}(\alpha - 1) \quad \textit{if} \quad \Delta \geq u_0\,. \end{cases} \tag{3.16}$$

It is immediate to check that these relations hold true also for $\alpha = 1$ *and* $\theta > 1$. *These properties will be useful in Chapter 4 when dealing with the control synthesis for practical stabilization.*
If instead $\theta > \frac{\alpha+1}{\alpha-1}$, *the set of the values of* Δ *such that* $Q_n(\Delta)$ *is controlled invariant becomes a non–connected union of intervals whose size decreases as* θ *increases and, for sufficiently large values of* θ, *it becomes empty.*

3.1.2 Hypercubes are minimal invariants

There are many possible shapes for invariant sets, the reasons for considering one class or another (e.g., ellipsoids, hypercubes or more general polytopes) can be varied. Three basic requirements one would aim at satisfying are: simplicity of description of the considered sets, simplicity of practical stability analysis and optimality. Since the goal of practical stabilization is to confine the trajectories of the system within *small* controlled invariant neighborhoods of the equilibrium, then optimality means that the considered family contains an invariant set which the trajectories can be made convergent to and that the size of such a set is *minimal* with respect to all controlled invariant sets. These requests are often trading off: e.g., ellipsoids can be easily described but they are not optimal (as it will be shown

in Section 3.1.4) ; polytopes instead are usually optimal but may be of arbitrarily complex description.
In this section, once explicitly remarked the simplicity of the controlled invariance analysis based on hypercubes, we analyze geometric properties holding for invariant sets of arbitrary shape and show the peculiarities exhibited by hypercubes in controller form coordinates. This analysis is helpful to properly define the concept of *size* for a set and is introductory to the statement of the minimality theorems for hypercubes. The bottom line is that the choice of considering hypercubes for the practical stabilization problem is motivated by the fact that they meet all the requirements of simplicity of description, simplicity of analysis and optimality.
Unless otherwise stated, throughout this section it is assumed that the system is represented in the controller form coordinates $\big($see assumption **A0** in equation (3.1)$\big)$. Notice that there is no loss of generality in changing coordinates in the state space and to compare different sets all in the same coordinates.

Invariance analysis for hypercubes is simple

The simplicity of description of hypercubes, as well as the resulting controlled invariance analysis (see Theorem 1) , is apparent. The following simple result is helpful to appreciate such a simplicity compared with other types of sets:

Lemma 7 *[43]* $\Omega \subseteq \mathbb{R}^n$ *is controlled invariant if and only if* $A\Omega \subseteq \bigcup_{u\in\mathcal{U}} (\Omega - Bu)$. □

Despite the simple formulation, the practical application of this invariance criterion is not straightforward when dealing with arbitrary sets Ω . In particular, to test the invariance of Ω , it is in general necessary to determine $A\Omega$. We have seen instead that for $\Omega = Q_n(\Delta)$ the analysis can be reduced to a 1–dimensional problem where invariant hypercubes are characterized by simple algebraic relations between Δ , α and the scalar functions $\rho(\Delta)$, $m(\Delta)$ and $M(\Delta)$. Furthermore, while Lemma 7 may give some insight on the geometric characteristics of controlled invariant sets, on the other hand it does not really answer the question of how to construct controlled invariant sets for a given system $\Sigma(A, B, \mathcal{U})$.

Geometric properties of invariant sets

The attention is now turned to the study of some geometric properties holding for arbitrarily shaped invariant sets and on how these results can be used for the practical stability analysis. First, it is useful to extend Definition 15 as follows:

Definition 15b *A static and time–invariant state feedback controller* $k : \mathbb{R}^n \to \mathcal{U}$ *is said to be* (X_0, Ω)–stabilizing *in* N steps *iff the closed loop system* $x^+ = Ax + Bk(x)$ *is* (X_0, Ω)–*stable and* $\forall\, x(0) \in X_0$ *, it holds that* $x(N) \in \Omega$.

Since we will widely exploit the properties of the canonical controller form, it is worth recalling that the control acts only on the n–th component while the others shift upwards.
Let $\Pr_{(i_1,\ldots,i_m)}\, x := (x_{i_1}, \ldots, x_{i_m})$:

Proposition 2 *Consider system* $\Sigma(A,B,\mathcal{U})$ *and assume* **A0**. *If* $\Omega \subseteq \mathbb{R}^n$ *is controlled invariant, then* $\mathrm{Pr}_{(2,\ldots,n)}\,\Omega \subseteq \mathrm{Pr}_{(1,\ldots,n-1)}\,\Omega$. *In particular,* $\mathrm{Pr}_n\,\Omega \subseteq \mathrm{Pr}_{n-1}\,\Omega \subseteq \cdots \subseteq \mathrm{Pr}_1\,\Omega$ *and* $\mathrm{diam}_n\,\Omega \leq \mathrm{diam}_{n-1}\,\Omega \leq \ldots \leq \mathrm{diam}_1\,\Omega$.

Proof. $\forall\, y = (y_2,\ldots,y_n) \in \mathrm{Pr}_{(2,\ldots,n)}\,\Omega$, $\exists\, x \in \Omega$ with $x = (x_1,y_2,\ldots,y_n)$. Let $u \in \mathcal{U}$ be such that $x^+ \in \Omega$: $x^+ = (y_2,\ldots,y_n,x_n^+)$, hence $y \in \mathrm{Pr}_{(1,\ldots,n-1)}\,\Omega$. ■

Given $\Omega \subseteq \mathbb{R}^n$, let $\mathcal{Z} := \mathrm{Pr}_n\,\Omega$ and

$$\Omega^* := \Omega \cap \mathcal{Z}^n\,. \tag{3.17}$$

The main property of Ω^* is exhibited by the following

Proposition 3 *Consider system* $\Sigma(A,B,\mathcal{U})$ *and assume* **A0**. *If* $\Omega \subseteq \mathbb{R}^n$ *is a controlled invariant neighborhood of the origin, then* Ω^* *is a controlled invariant neighborhood of the origin and* $\forall\,\phi : \mathbb{R}^n \to \mathcal{U}$ *rendering* Ω *positively invariant,* ϕ *is* (Ω,Ω^*)*–stabilizing in* $n-1$ *steps.*

Proof. Ω^* is a neighborhood of the origin because so are both Ω and $\mathcal{Z}^n$. Since Ω is positively invariant, to prove that ϕ is (Ω,Ω^*)–stabilizing in $n-1$ steps, we have to show that $\forall\, x(0) \in \Omega$, $\forall\, t \geq n-1$ and $\forall\, i = 1,\ldots,n$, $x_i(t) \in \mathcal{Z}$. Indeed, $\forall\, x \in \Omega$, $x_n^+ \in \mathcal{Z}$ by the definition of $\mathcal{Z}$. Since the system is in controller form, the thesis follows. ■

Corollary 1 *If* ϕ *is* (X_0,Ω)*–stabilizing, then* ϕ *is* (X_0,Ω^*)*–stabilizing.* □

Therefore, $\Omega \setminus \Omega^*$ is a redundant part of the invariant set Ω, meaning that the trajectories lie within $\Omega \setminus \Omega^*$ only for a transient time of at most $n-1$ steps. As the aim of practical stabilization is to confine the trajectories within small controlled invariant neighborhoods of the origin, in the analysis of stabilizing control laws it is then proper to replace the final set Ω by Ω^*, namely to "*cut off*" the redundant region. This procedure is effective because, by Proposition 2, $\mathrm{Pr}_i\,\Omega \subseteq \mathrm{Pr}_{i-1}\,\Omega$ $\forall\, i = 2,\ldots,n$ and $\Omega \setminus \Omega^* \neq \emptyset$ whenever one of the inclusions is strict. In general, $(\Omega^*)^* \subset \Omega^*$, namely the cut–off procedure can be iterated.
Notice that if $\Omega = Q_n(\Delta)$, then $\Omega^* = \Omega$, namely the hypercubes $Q_n(\Delta)$ are non–redundant. This is not the case for more commonly encountered types of invariant sets such as ellipsoids. Quantitative results on the effect of the cut–off procedure on ellipsoids will be given in Section 3.1.4.
Hypercubes are not the only example of invariant sets such that $\Omega^* = \Omega$, the same property holds if Ω is inscribed in a hypercube. This fact will be relevant in next section when discussing on the minimality of hypercubes and is related with the advisability of introducing two notions of minimality. Indeed, among these notions, the strongest one allows us to exclude the existence of invariant sets Ω inscribed in the smallest invariant hypercube.

Minimality properties of invariant hypercubes

In order to investigate the minimality properties of the smallest controlled invariant hypercube with respect to all controlled invariant sets, we need to introduce a suitable notion of size

for controlled invariant sets. We choose to study the minimality problem by comparing sets according to their extension in some vector norm $\|\cdot\|_*$. That is, for a neighborhood Ω of the origin, we consider

$$\|\Omega\|_* := \sup_{x\in\Omega} \|x\|_* .$$

Indeed, by achieving the convergence of the trajectories to within Ω, it is guaranteed that

$$\limsup_{t\to+\infty} \|x(t)\|_* \leq \|\Omega\|_* .$$

For comparison purposes, we will also consider the volume and the containment relation. Nevertheless, the volume only is not suitable in the practical stability framework because it does not provide any information about how far a trajectory can go away from the equilibrium. As for the containment relation, although it may appear to be a natural way of comparing invariant sets, this relation is not a total ordering and controlled invariance does not behave well as for intersection (i.e., if Ω_1 and Ω_2 are controlled invariant, then $\Omega_1 \cap \Omega_2$ is not necessarily controlled invariant), therefore the minimality problem formulated in terms of the containment relation is not well posed.
In what follows, sets are measured by considering their extension in the infinity norm in the controller form coordinates. More precisely, we consider the diameter of the sets along the n coordinate directions (i.e., $\operatorname{diam}_i \Omega$, $i = 1, \ldots, n$). Actually, according to Propositions 2 and 3, the relevant quantity is $\operatorname{diam}_n \Omega$, in fact longer extensions of Ω along the other directions can be cut off. Hence we give the following

Definition 17 *Consider a system $x^+ = Ax + Bu$ in the controller form coordinates and let Ω be a controlled invariant neighborhood of the origin: the quantity $\operatorname{diam}_n \Omega$ is referred to as the* magnitude *of Ω. We say that Ω is* minimal in magnitude *iff any bounded controlled invariant neighborhood of the origin Ω' has a magnitude greater than or equal to that of Ω.*

If the pair (A, B) is not in controller form, the magnitude of Ω can be easily computed through the formula $\operatorname{diam}\big(\Pr_B(\Omega)\big)/\|B\|_2^2$, where $\Pr_B(x) := x'B \in \mathbb{R}$.
Hence, the magnitude is the measure we use for comparing invariant sets. In some cases it is still possible to consider the containment relation and to study minimality properties which are stronger than minimality in magnitude: next, a result in this direction will be given in Theorem 3.
Obviously, if the open loop system $x^+(t) = Ax(t)$ is stable, then there exist arbitrarily small invariant neighborhoods of the origin. Therefore, we only consider the case of systems whose open loop dynamics is not stable in the Lyapunov sense: we identify these systems by saying that the matrix A is *unstable*.

Theorem 2 (Minimality in magnitude) *Consider system $\Sigma(A, B, \mathcal{U})$ and assume* **A0**. *Let $u_0 = \min_{u\in\mathcal{U}\setminus\{0\}} |u|$ be the resolution at 0 of $\mathcal{U}$ and Ω be a bounded controlled invariant neighborhood of the origin. If A is an unstable matrix, then $\operatorname{diam}_i \Omega \geq u_0 \ \forall\, i = 1, \ldots, n$. In particular, if $Q_n(u_0)$ is controlled invariant, then $Q_n(u_0)$ is minimal in magnitude.*

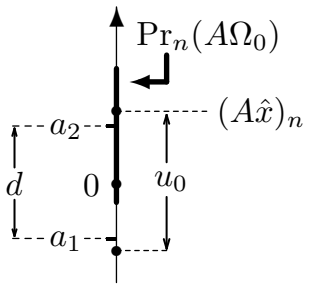

Figure 3.6: Visual help for the proof of Theorem 2 : the thicker segment represents $\mathrm{Pr}_n\,(A\Omega_0)$.

Proof. Thanks to Proposition 2 , it is sufficient to show that $\mathrm{diam}_n\,\Omega \geq u_0$ (i.e., that the magnitude of Ω is greater than or equal to u_0) . Let us assume by contradiction that $d := \mathrm{diam}_n\,\Omega < u_0$. Set $a_1 := \inf_{x\in\Omega} x_n$ and $a_2 := \sup_{x\in\Omega} x_n$, then $\mathrm{Pr}_n\,\Omega \subseteq [\,a_1\,,a_2\,]$, $a_2 - a_1 = d < u_0$ and $0 \in [\,a_1\,,a_2\,]$. Let Ω_0 be the path connected component of Ω containing 0 (see [115]) . As $\mathrm{Pr}_n \circ A$ is a continuous function, $\mathrm{Pr}_n\,(A\Omega_0)$ is an interval. Two cases can occur:

I) Suppose that $\mathrm{Pr}_n\,(A\Omega_0) \cap {}^c[\,a_1\,,a_2\,] \neq \emptyset$: since $\mathrm{Pr}_n\,(A\Omega_0)$ is an interval that intersects $[\,a_1\,,a_2\,]$ (in fact it contains 0) , then, with $\theta := u_0 - d$, there exists $\hat{x} \in \Omega_0$ such that $(A\hat{x})_n \in]\,a_1 - \theta\,, a_1\,[\,\cup\,]\,a_2\,, a_2 + \theta\,[$. In this case, by the definition of u_0 , it is easy to see that $\forall\, u \in \mathcal{U}$, $\hat{x}_n^+ \notin [\,a_1\,,a_2\,]$ (see Fig. 3.6) : this contradicts the controlled invariance of Ω as $\mathrm{Pr}_n\,\Omega \subseteq [\,a_1\,,a_2\,]$.

II) Suppose instead that $\mathrm{Pr}_n\,(A\Omega_0) \subseteq [\,a_1\,,a_2\,]$. We claim that $\exists\, x \in \Omega_0$ such that $Ax \notin \Omega$. The claim implies the thesis, in fact: for such an x , by the controlled invariance of Ω , $\exists\, u \in \mathcal{U}\setminus\{0\}$ such that $x^+ \in \Omega$, but $u \neq 0$ together with $(Ax)_n \in [a_1\,,\,a_2]$ and $a_2 - a_1 < u_0$ imply that $x_n^+ \notin [\,a_1\,,a_2\,]$ which contradicts the fact that $x^+ \in \Omega$.
Let us prove the claim: first, since Ω_0 is a bounded neighborhood of the origin, $A\Omega_0 \not\subseteq \Omega_0$. In fact, if the contrary held, then $\forall\, k \in \mathbb{N}$, $A^k\Omega_0 \subseteq \Omega_0$ which contradicts the fact that A is unstable. Since $A\Omega_0$ is path connected, if $A\Omega_0 \subseteq \Omega$, then $A\Omega_0$ would be contained in a path connected component of Ω . As $0 \in A\Omega_0 \cap \Omega_0$, then $A\Omega_0 \subseteq \Omega_0$ which is a contradiction. ■

Corollary 2 *If system $\Sigma(A,B,\mathcal{U})$ is reachable and A is unstable, a necessary condition for the (X_0,Ω)–stabilizability of the system is that the magnitude of Ω is greater than or equal to u_0 .* □

Clearly, for the (X_0,Ω)–stabilizability it is also necessary that Ω is reachable from X_0 . We will be back on this issue in next Chapters 4 , 5 and 6 , where the focus is on the control synthesis for practical stabilization. However, it is worth to mention that the cases in which $\mathcal{U}$ is uniformly or logarithmically quantized provide two classes of examples where the (X_0,Ω)–stabilizability holds with $\Omega = Q_n(u_0)$ (this will be shown in Examples 15 and 16 of Section 4.1) . Namely,

the lower bound for the magnitude of Ω necessary for the (X_0, Ω)–stabilizability is attained by a hypercube.

Here, we limit ourselves to notice that, if $\mathcal{U}$ is uniformly quantized with parameter u_0, the minimal invariant hypercube is $Q_n(u_0)$ (see Example 7 in Section 3.1.1) which, by Theorem 2, is minimal in magnitude. In the case of a logarithmically quantized control set with parameters (u_0, θ) and $1 < \theta \leq \frac{\alpha+1}{\alpha-1}$, the same property holds for $Q_n(u_0)$ (see Example 8). These fundamental classes of examples are not the only cases where the lower bound for the magnitude of Ω is attained by a hypercube. Indeed, the same properties hold for the particular system considered in Example 6 $\big($the $\big(X_0, Q_n(u_0)\big)$–stabilizability will be shown in Example 14 of Section 4.1$\big)$.
It has to be stressed that invariant neighborhoods Ω strictly contained in $Q_n(u_0)$ and with smaller volume can exist (see Example 9 below). Nevertheless, Theorem 2 states that, even if such an Ω exists, it spreads up to the border of $Q_n(u_0)$ in all the directions of the coordinate axes (i.e., $\forall\, i = 1, \ldots, n$, $\operatorname{diam}_i \Omega = u_0$; see Fig. 3.7) so that Ω and $Q_n(u_0)$ are equivalent as for their extension in the infinity norm. Therefore, even if the convergence of the trajectories to within such an Ω was proved, no improvement would be obtained in terms of the asymptotic behavior of the system, namely, it would be still guaranteed that $\limsup_{t\to+\infty} \|x(t)\|_\infty \leq \|\Omega\|_\infty = \|Q_n(u_0)\|_\infty$.

Example 9 *Let us consider the quantized input system*

$$\begin{cases} x^+ = Ax + Bu \\ x \in \mathbb{R}^n,\ u \in \mathbb{Z}, \end{cases}$$

where, as usual, the pair (A, B) is in controller form. Assume that A is an unstable matrix such that $0 < |\det A| < 1$.
By $\mathcal{U} = \mathbb{Z}$, it easily follows that the semi–open hypercube $Q_n^o(1) = \big[-\frac{1}{2}, \frac{1}{2}\big[^n$ is controlled invariant and that $\forall\, x \in Q_n^o(1)$, there exists a unique $u \in \mathbb{Z}$ such that $x^+ \in Q_n^o(1)$. It is hence univocally defined the map

$$\begin{array}{rccc} T: & Q_n^o(1) & \to & Q_n^o(1) \\ & x & \mapsto & x^+ \end{array} .$$

Since A is unstable, $Q_n^o(1)$ is minimal in magnitude by Theorem 2. Because $\det A \neq 0$, T is a local diffeomorphism at 0, therefore $\forall\, k \in \mathbb{N}$, the set $T^k\big(Q_n^o(1)\big)$ is a neighborhood of the origin. Moreover, since $T^{k+1}\big(Q_n^o(1)\big) \subseteq T^k\big(Q_n^o(1)\big)$, then $T^k\big(Q_n^o(1)\big)$ is controlled invariant and, being a subset of $Q_n^o(1)$, it is minimal in magnitude. Furthermore, we claim that

$$\forall\, k \in \mathbb{N}, \quad T^{k+1}\big(Q_n^o(1)\big) \subset T^k\big(Q_n^o(1)\big) \tag{3.18}$$

and, denoted by μ the Lebesgue measure,

$$\lim_{k\to+\infty} \mu\Big(T^k\big(Q_n^o(1)\big)\Big) = 0\,. \tag{3.19}$$

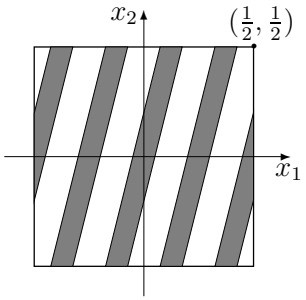

Figure 3.7: The non–connected shaded region represents the controlled invariant set $T\big(Q_n^o(1)\big)$ for the two dimensional system discussed in Example 9 and having $a_1 = 0.4$ and $a_2 = 4$.

Namely, $\{T^k\big(Q_n^o(1)\big)\}_{k\in\mathbb{N}}$ is a strictly decreasing sequence of controlled invariant neighborhoods of the origin made up of minimal in magnitude sets and containing elements of arbitrarily small volume. According to Theorem 2, all of these sets spread up to the border of $Q_n(1)$ in all the coordinate directions, thus having the same extension in the infinity norm as $Q_n(1)$. The typical structure of one of the sets of the sequence (in the two dimensional case) is represented by the shaded region in Fig. 3.7.
Before proving the claim, notice that, by definition, the set $T^k\big(Q_n^o(1)\big)$ is reachable in k steps by any point in $Q_n^o(1)$. In next Chapter 4, we will also see that, with a qdb–controller, $T^k\big(Q_n^o(1)\big)$ is reachable in $k+n$ steps by any point in $\mathbb{R}^n$ (see Example 15 in Section 4.1). Let us prove the claim. As for the inclusion (3.18)*, because $T^{k+1}\big(Q_n^o(1)\big) \subseteq T^k\big(Q_n^o(1)\big)$, we have only to show that indeed the inclusion is strict. It holds that*

$$\forall\, k \in \mathbb{N}, \quad \mu\Big(T^{k+1}\big(Q_n^o(1)\big)\Big) \leq \mu\Big(\big(A \circ T^k\big)\big(Q_n^o(1)\big)\Big), \tag{3.20}$$

in fact: $\forall\, u \in \mathbb{Z}$, let $\mathcal{S}_u := \big\{x \in \mathbb{R}^n \,|\, u - \frac{1}{2} \leq x_n < u + \frac{1}{2}\big\}$ and $\mathcal{R}_u := \mathcal{S}_u \cap \big(A \circ T^k\big)\big(Q_n^o(1)\big)$; then $\big(A \circ T^k\big)\big(Q_n^o(1)\big) = \bigcup_{u\in\mathbb{Z}} \mathcal{R}_u$ and $T^{k+1}\big(Q_n^o(1)\big) = \bigcup_{u\in\mathbb{Z}} (\mathcal{R}_u - Bu)$, inequality (3.20) *then easily follows. Since $\mu\Big(\big(A \circ T^k\big)\big(Q_n^o(1)\big)\Big) = |\det A| \cdot \mu\Big(T^k\big(Q_n^o(1)\big)\Big)$ and $|\det A| < 1$, then inequality* (3.20) *yields*

$$\forall\, k \in \mathbb{N}, \quad \mu\Big(T^{k+1}\big(Q_n^o(1)\big)\Big) \leq |\det A| \cdot \mu\Big(T^k\big(Q_n^o(1)\big)\Big) < \mu\Big(T^k\big(Q_n^o(1)\big)\Big).$$

This implies that $\forall\, k \in \mathbb{N}$, $T^{k+1}\big(Q_n^o(1)\big) \subset T^k\big(Q_n^o(1)\big)$ and $\mu\Big(T^k\big(Q_n^o(1)\big)\Big) \leq |\det A|^k \cdot \mu\big(Q_n^o(1)\big)$, thus the limit in equation (3.19) *holds.* ♣

Example 9 shows that a minimal in magnitude set can contain other minimal in magnitude sets having smaller volume (indeed, having an arbitrarily small volume). This gives the reasons for the need to introduce the concept of *strong* minimality which strengthens the minimality in magnitude by involving the containment relation.

Definition 18 *A controlled invariant neighborhood of the origin Ω is said to be* strongly minimal *iff it is minimal in magnitude and any neighborhood of the origin Ω' strictly contained in Ω is not controlled invariant.*

If the system is sufficiently unstable (in a sense specified below), then the strong minimality property holds for hypercubes. More precisely,

Theorem 3 (Strong minimality) *Consider system $\Sigma(A, B, \mathcal{U})$ and assume* **A0**. *Let $u_0 = \min_{u \in \mathcal{U} \setminus \{0\}} |u|$ be the resolution at 0 of $\mathcal{U}$. If*

$$|a_1| > 1 + \sum_{i=2}^{n} |a_i| \tag{3.21}$$

and $\Omega \subseteq Q_n^o(u_0)$ is a controlled invariant neighborhood of the origin, then $\Omega = Q_n^o(u_0)$. In particular, if $Q_n^o(u_0)$ is controlled invariant, then it is strongly minimal.

Proof. We show that if such an Ω exists, then it contains a subset whose uncontrolled evolution is confined within $Q_n^o(u_0)$ until it covers the whole semi–open hypercube. By definition of u_0, such an evolution is also the unique ensuring that the trajectories starting from this subset remain within $Q_n^o(u_0)$: since Ω is controlled invariant, this entails that $\Omega = Q_n^o(u_0)$.
In detail, the matrix A is invertible and

$$(A^{-1}x)_j = \begin{cases} \frac{x_n - \sum_{i=1}^{n-1} a_{i+1} x_i}{a_1} & \text{if } \ j = 1 \\ x_{j-1} & \text{otherwise}. \end{cases} \tag{3.22}$$

Let $\theta := \frac{(1 + \sum_{i=2}^{n} |a_i|)}{|a_1|}$. By the assumption in equation (3.21), $\theta < 1$, then

$$\forall\, x \in \mathbb{R}^n, \quad \left|(A^{-1}x)_1\right| \leq \frac{|x_n| + \sum_{i=1}^{n-1} |a_{i+1}||x_i|}{|a_1|} \leq \theta \cdot \|x\|_\infty < \|x\|_\infty . \tag{3.23}$$

Equations (3.22) and (3.23) imply that $A^{-1}Q_n^o(u_0) \subset Q_n^o(u_0)$, thus $\forall\, h \in \mathbb{N}$,

$$A^{-h}Q_n^o(u_0) \subseteq A^{-h+1}Q_n^o(u_0) \subseteq \cdots \subseteq A^{-1}Q_n^o(u_0) \subset Q_n^o(u_0), \tag{3.24}$$

and in particular $\|A^{-h}\|_\infty \leq 1$. Moreover, by the Hamilton–Cayley identity,

$$A^{-n} = \frac{1}{a_1}\Big(I_n - \sum_{i=2}^{n} a_i A^{-n-1+i} \Big),$$

therefore $\|A^{-n}\|_\infty \leq \theta$: this means that $A^{-n}Q_n(u_0) \subseteq Q_n(\theta\, u_0)$ and it immediately follows that $\forall\, k \in \mathbb{N}$, $A^{-nk}Q_n^o(u_0) \subseteq Q_n(\theta^k\, u_0)$.
Let Ω be a controlled invariant neighborhood of the origin: since $\lim_{k \to +\infty} \theta^k = 0$, $\exists\, \hat{k} \in \mathbb{N}$ such that $Q_n(\theta^{\hat{k}}\, u_0) \subseteq \Omega$, therefore $A^{-n\hat{k}}Q_n^o(u_0) \subseteq \Omega$. We claim that if $A^{-m}Q_n^o(u_0) \subseteq \Omega$ for

some $m \geq 1$, then $A^{-m+1}Q_n^o(u_0) \subseteq \Omega$. In our case, the hypothesis of the claim is satisfied $\forall\, m \geq n\hat{k}$ and the recursive application of the claim implies that $Q_n^o(u_0) \subseteq \Omega$, namely the thesis.
Let us prove the claim. First, we show that if $x \in \Omega$ and $Ax \in Q_n^o(u_0)$, then $Ax \in \Omega$. In fact, by the controlled invariance of Ω, $\exists\, u \in \mathcal{U}$ such that $x^+ \in \Omega \subseteq Q_n^o(u_0)$: such a control value must be $u = 0$ because for $u \neq 0$, $x^+ \notin Q_n^o(u_0)$. Indeed, $-\frac{u_0}{2} \leq (Ax)_n < \frac{u_0}{2}$ by assumption, hence

$$-\frac{u_0}{2} + u \leq x_n^+ = (Ax)_n + u < \frac{u_0}{2} + u\,, \tag{3.25}$$

and for $u \neq 0$ it holds that $|u| \geq u_0$ which, together with inequalities (3.25), yields either $x_n^+ \geq \frac{u_0}{2}$ or $x_n^+ < -\frac{u_0}{2}$. Now, consider $y \in A^{-m+1}Q_n^o(u_0)$ and let us show that $y \in \Omega$: since $-m+1 \leq 0$, then $y \in Q_n^o(u_0)$ $\big($see the inclusions in equation (3.24)$\big)$. Let $x := A^{-1}y \in A^{-m}Q_n^o(u_0)$: $x \in \Omega$ by assumption and $y = Ax \in Q_n^o(u_0)$, therefore $y \in \Omega$. ■

Thus, under the assumption in equation (3.21), $Q_n^o(u_0)$ is strongly minimal in both the cases of uniformly and logarithmically quantized controls $\big($in the latter case, if $1 < \theta \leq \frac{\alpha+1}{\alpha-1}\big)$.

Remark 8 *Assuming $|a_1| > 1+\sum_{i=2}^n |a_i|$, which by the way is a condition involving only the coefficients of the characteristic polynomial of A, is the same as asking that $A^{-1}Q_n^o(u_0) \subset Q_n^o(u_0)$, namely it is a stability requirement on the matrix A^{-1}, hence corresponding to an instability property of A.*

It can be shown that the condition ensuring the strong minimality of $Q_n^o(u_0)$ is only sufficient (see [99]), nevertheless the result is interesting because it shows that there are cases in which, among the minimal diameter sets (i.e., $\operatorname{diam}_i \Omega = u_0 \;\; \forall\, i = 1, \ldots, n$), the whole $Q_n^o(u_0)$ is actually the smallest controlled invariant set.

3.1.3 An extension to networked systems

In this section, we take advantage of some results presented in the previous sections to address an issue in the framework of the *control under communication constraints*. We consider the problem of controlling multiple scalar systems through a limited capacity shared channel (see Fig. 3.8). It is assumed that each system is affected by process noise and can be controlled by actuators with values in an assigned uniformly quantized set. The control objective is to bound the evolution of the systems in specified sets (controlled invariance). It is part of the problem to find an optimal allocation of the shared communication resource to the different control activities. This section provides fundamental conceptual tools to attack the design problem in the formal framework of an optimization problem.

Traditional control design is based on ideal assumptions concerning the type of information that can flow across the control loop. Unfortunately, real implementation platforms exhibit non–idealities that may substantially invalidate these assumptions. As a result, the system's closed loop performance can be severely affected. This problem shows up with particular strength when multiple control loops share a limited pool of computation and communication

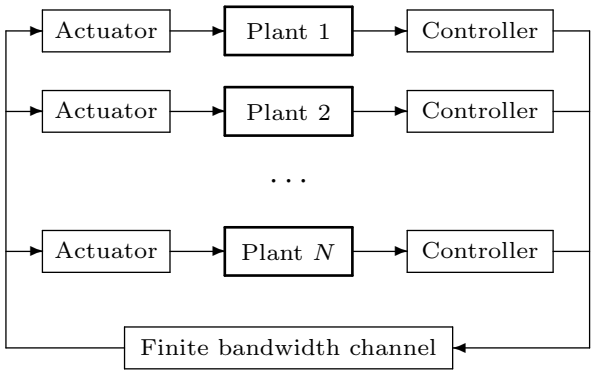

Figure 3.8: Pictorial representation of the problem analyzed in Section 3.1.3 .

resources. In this case, the designer is confronted with the challenging task of choosing at the same time the control law and the optimal allocation policy for the shared resources (control algorithm/system architecture co–design) . An intriguing general discussion for this class of problems can be found in [16] . Investigations in this field have been developed in several directions. A first prong of research activities has focused on the problem of resource sharing [95, 57, 22] . However, these works do not explicitly cope with quantization and bit–rate constraints that, on the contrary, play an important role in complex distributed systems. A remarkable thread of papers on the problem of stabilization under bit–rate constraints is included in the recent literature on quantized control systems (see [91, 106] and references therein) . In most of these works, the authors *design* quantization schemes instrumental to the goal of finding encoding/decoding policies that make for an optimal use of the channel. In this section, instead, we keep on studying the problem under the assumption that quantization is *assigned* and we focus on the analysis of the attainable control performance.

The general setting we refer to is depicted in Fig. 3.8 . This can be thought of as "smart–sensor" scenario, i.e., one where processing activities are located in the proximity of sensors and commands have to be sent to actuators by a channel.

For the sake of simplicity, the analysis is restricted to scalar systems under unform input quantization and ruled by control laws generated by periodic sampling. A limited bandwidth channel is shared between several independent control loops: in order to limit the "channel occupation", some of the levels provided by the quantized actuators can be left unused. Therefore, the subset $\mathcal{U}$ necessary to accomplish the control task is a design parameter along with the sampling period. Thus, the quantized actuators are regarded as given "hardware" components to build on the top of. In this framework, quantization has a twofold role: on the one hand it is a constraint imposed by the physics of the system (e.g., because actuation and/or sensing are inherently quantized) , on the other hand it is introduced on purpose in order to allow for communication over the finite capacity channel. The final goal is to produce automated procedures for the optimal allocation of the channel capacity among the different

loops and for the choice of the set of design parameters. It has to be stressed that, in this context, the presence of noise is of particular interest. Indeed, it emphasizes the importance of the sampling period especially for unstable systems where long sampling intervals determine performance degradation due to longer uncompensated actions of the noise term.

Notation: In this section, for $\Lambda > 0$, we let[1] $I(\Lambda) := \left[-\frac{\Lambda}{2} ; \frac{\Lambda}{2} \right]$.

Problem formulation

Consider a set of $N \geq 1$ continuous time scalar plants

$$\begin{cases} \dot{\tilde{x}}_i(\tau) = a_i \tilde{x}_i(\tau) + \tilde{u}_i(\tau) + \tilde{w}_i(\tau) \\ \tilde{x}_i(0) = \tilde{x}_i^0 \in \mathbb{R} \\ \tau \in \mathbb{R}^+ \\ i = 1, \ldots, N \,, \end{cases} \tag{3.26}$$

where the control function $\tilde{u}_i(\tau)$ takes values in a quantized set $\mathcal{U}_i \subseteq \epsilon_i \mathbb{Z}$ ($\epsilon_i > 0$) and $\tilde{w}_i(\tau) \in I(\mathrm{w}_i)$ ($\mathrm{w}_i \geq 0$) represents an exogenous noise term. It is supposed that $\tilde{w}_i(\tau)$ is an integrable function. Each scalar system is characterized by the triple $(a_i, \epsilon_i, \mathrm{w}_i)$ which are the *given* parameters and it is denoted by $\Sigma(a_i, \epsilon_i, \mathrm{w}_i)$.
We assume that the states $\tilde{x}_i$ ($i = 1, \ldots, N$) are sampled periodically at time 0, T_i, $2T_i, \ldots$ (the sampling intervals T_i are design parameters). Based on these samples, N individual control values are derived and transmitted over a *shared* communication channel to zero–order hold devices in the respective actuator nodes. The sampled–data control system corresponding to system (3.26) is

$$\begin{cases} x_i(t+1) = \alpha_i x_i(t) + \beta_i u_i(t) + w_i(t) \\ x_i(0) = \tilde{x}_i^0 \\ t \in \mathbb{N} \\ i = 1, \ldots, N \,, \end{cases} \tag{3.27}$$

where $x_i(t) = \tilde{x}_i(tT_i)$ and

$$\begin{cases} \alpha_i &= e^{a_i T_i} \\ \beta_i &= \int_0^{T_i} e^{a_i s} ds \\ w_i(t) &= \int_{tT_i}^{(t+1)T_i} e^{a_i \left((t+1)T_i - s \right)} \tilde{w}_i(s) ds \,. \end{cases}$$

Each T_i–sampled system is denoted by $\Sigma(a_i, \epsilon_i, \mathrm{w}_i, T_i)$. From the last equation it follows that the discrete time disturbance $w_i(t)$ takes values in $I(\beta \cdot \mathrm{w}_i)$.
In this setting, the discrete time control law is a function

$$\begin{array}{rccc} q_i : & \mathbb{R} & \to & \mathcal{U}_i \subseteq \epsilon_i \mathbb{Z} \\ & x_i(t) & \mapsto & u_i(t) \end{array} \tag{3.28}$$

[1] This notation is redundant because $I(\Lambda) = Q_1(\Lambda)$, but certainly it is more intuitive.

so that, with the zero–order hold, the continuous time control law is the piecewise constant function

$$\tilde{u}_i(\tau) = u_i(t)\,, \quad \tau \in [\,tT_i\,,\,(t+1)\,T_i\,[\,.$$

The control goal is to guarantee practical stability for the closed loop dynamics of each plant, more precisely we consider the controlled invariance problem. Namely, we are interested in finding neighborhoods of the equilibrium where the trajectories of each plant can be confined irrespective to any noise affecting the systems. In this framework, the controlled invariance notion introduced in Definition 14 of Section 2.3.1 is extended in the following way:

Definition 19 (Robust controlled invariance)
ı) For a continuous time system $\Sigma(a,\epsilon,\mathrm{w})$ *with control set* $\mathcal{U}\subseteq\epsilon\mathbb{Z}$, *the interval* $I(\Delta)$ *is said to be* (T,w)–controlled invariant *iff* $\forall\,\tilde{x}^0\in I(\Delta)$, $\exists\,u\in\mathcal{U}$ *such that for any integrable function* $\tilde{w}:[\,0\,,\,T\,]\to I(\mathrm{w})$ *the solution of*

$$\begin{cases}\dot{\tilde{x}}(\tau)=a\tilde{x}(\tau)+u+\tilde{w}(\tau)\\ \tilde{x}(0)=\tilde{x}^0\end{cases}$$

is such that, $\forall\,\tau\in[\,0\,,\,T\,]$, $\tilde{x}(\tau)\in I(\Delta)$.

ıı) For a discrete time system $\Sigma(a,\epsilon,\mathrm{w},T)$ *with control set* $\mathcal{U}\subseteq\epsilon\mathbb{Z}$, *the interval* $I(\Delta)$ *is said to be* w–controlled invariant *iff* $\forall\,x\in I(\Delta)$, $\exists\,u\in\mathcal{U}$ *such that* $\forall\,w\in I(\beta\cdot\mathrm{w})$, $x^+=\alpha x+\beta u+w\in I(\Delta)$.

The following proposition allows us to get rid of the distinction between continuous time and discrete time:

Proposition 4 *Consider system* $\Sigma(a,\epsilon,\mathrm{w},T)$*: if* $x^0\in I(\Delta)$ *and* $u\in\mathcal{U}$ *is such that,* $\forall\,w\in I(\beta\cdot\mathrm{w})$, $x^{0+}=\alpha x^0+\beta u+w\in I(\Delta)$, *then for any integrable function* $\tilde{w}:[\,0\,,\,T\,]\to I(\mathrm{w})$ *the solution* $\tilde{x}(\tau)$ *of*

$$\begin{cases}\dot{\tilde{x}}(\tau)=a\tilde{x}(\tau)+u+\tilde{w}(\tau)\\ \tilde{x}(0)=x^0\end{cases}\tag{3.29}$$

is such that, $\forall\,\tau\in[\,0\,,\,T\,]$, $\tilde{x}(\tau)\in I(\Delta)$.
In particular, if $I(\Delta)$ *is* w*–controlled invariant for the discrete time system* $\Sigma(a,\epsilon,\mathrm{w},T)$, *then it is* (T,w)*–controlled invariant for the continuous time system* $\Sigma(a,\epsilon,\mathrm{w})$.

Proof. See in Appendix A.2.2. ■

Conversely, it is obvious that if $I(\Delta)$ is (T,w)–controlled invariant for system $\Sigma(a,\epsilon,\mathrm{w})$, then $I(\Delta)$ is w–controlled invariant for system $\Sigma(a,\epsilon,\mathrm{w},T)$. Thus, it is sufficient to introduce and to check the properties based on invariance only for discrete time models.

The plants share a limited bandwidth channel with bit–rate $\mathcal{R}$.

Definition 20 (The channel) *By a limited bandwidth* channel of capacity $\mathcal{R}$ *we mean a device capable of transmitting* $\mathcal{R}$ *bits per unit of time without transmission error.*

In particular, the number of symbols σ that can be transmitted during the time interval T satisfies $\sigma \leq 2^{\mathcal{R}T}$. Since the number of bits to be transmitted at each sampling instant is integer, we require that $\sigma \leq 2^{\lfloor \mathcal{R}T \rfloor}$.
According to some optimality criterion, the total capacity $\mathcal{R}$ of the shared communication link is split among the N control loops. Hence, the bit–rates R_i's devoted to each control loop have to comply with the following inequality:

$$\sum_{i=1}^{N} R_i \leq \mathcal{R}\,.$$

It is supposed that the allocation R_i is time–invariant, in other words, the resource assignment is decided once and forever.
In accordance with the framework described so far, we introduce the following

Definition 21 *Consider system $\Sigma(a, \epsilon, \mathrm{w})$, suppose that a channel of capacity R is connecting the controller to the plant: the triple (R, T, Δ) is said to be* feasible *for the invariance problem iff there exists a control set $\mathcal{U} \subset \epsilon\mathbb{Z}$ rendering $I(\Delta)$ w–controlled invariant for system $\Sigma(a, \epsilon, \mathrm{w}, T)$ and satisfying $\#\mathcal{U} \leq 2^{\lfloor RT \rfloor}$.*
Consider a set of systems $\left\{\Sigma(a_i, \epsilon_i, \mathrm{w}_i)\right\}_{i=1,\ldots,N}$. Let $\vec{R} := (R_1, \ldots, R_N)$, $\vec{T} := (T_1, \ldots, T_N)$ and $\vec{\Delta} := (\Delta_1, \ldots, \Delta_N)$. The triple $(\vec{R}, \vec{T}, \vec{\Delta})$ is said to be feasible *iff $\forall\, i = 1, \ldots, N$, (R_i, T_i, Δ_i) is feasible for the invariance problem related to system $\Sigma(a_i, \epsilon_i, \mathrm{w_i})$.*

In this problem, the input quantization is *assigned* ($u_i \in \epsilon_i\mathbb{Z}$) but, in order to meet the communication constraint, the designer has the freedom to *choose* the input sets $\mathcal{U}_i \subset \epsilon_i\mathbb{Z}$ where the control laws are restricted to take values. Therefore, quantization is both a physical constraint *and* a means to enable communication over the shared channel.

We are ready for the explicit problem formulation. The presence of multiple plants opens up different design possibilities as to how the communication capacity of the link can be shared between the different control loops with the purpose of solving a controlled invariance problem. We propose the following approach:

• ***Problem formulation:*** For a given set of systems $\left\{\Sigma(a_i, \epsilon_i, \mathrm{w}_i)\right\}_{i=1,\ldots,N}$, consider a vector $\vec{\Delta}_0$ identifying a set of N intervals within which the trajectory of the N systems are desired to be confined. Let $f : (\mathbb{R}^+)^N \to \mathbb{R}^+$ be a cost function penalizing realizations $\vec{\Delta}$ differing from the desired target $\vec{\Delta}_0$, for instance:

$$f(\vec{\Delta}) := \frac{\|\vec{\Delta}_0 - \vec{\Delta}\|_\infty}{\|\vec{\Delta}_0\|_\infty}\,. \tag{3.30}$$

The design problem is formulated as follows: find

$$\begin{aligned} (\vec{R}^*, \vec{T}^*, \vec{\Delta}^*) = \quad & \underset{(\vec{R},\vec{T},\vec{\Delta})}{\operatorname{argmin}} \; f(\vec{\Delta}) \\ \text{subj. to:} \quad & \begin{cases} \vec{\Delta} \in \mathcal{D} \\ \sum_{i=1}^{N} R_i \leq \mathcal{R} \\ (\vec{R}, \vec{T}, \vec{\Delta}) \text{ feasible}\,, \end{cases} \end{aligned} \tag{3.31}$$

where $\mathcal{D} \subseteq (\mathbb{R}^+)^n$ is a specified domain within which $\vec{\Delta}$ is desired to lie. Thus, the constraint $\vec{\Delta} \in \mathcal{D}$ can be seen as a *performance* requirement and/or as a *safety* constraint.
For each of the N systems, the minimizing tern $(\vec{R}^*, \vec{T}^*, \vec{\Delta}^*)$ determines the bandwidth R_i^*'s to be assigned, the sampling periods T_i^*'s and the intervals $I(\Delta_i^*)$'s which can be made invariant. By definition, the feasibility of $(\vec{R}^*, \vec{T}^*, \vec{\Delta}^*)$ ensures the realizability of the solution, that is: the existence of N control laws of the type in equation (3.28), each one taking values in a finite set $\mathcal{U}_i \subset \epsilon_i \mathbb{Z}$ so that both the robust invariance of $I(\Delta_i^*)$ is ensured and the communication constraint imposed by the available bandwidth R_i is satisfied.

• ***Equivalent problem formulation and solution methodology:*** Consider

$$R_{\min}(\vec{\Delta}) := \sum_{i=1}^{N} R_{\min}^{(i)}(\Delta_i)\,,$$

where $R_{\min}^{(i)}(\Delta_i)$ is the smallest bit–rate R_i ensuring that there exists a choice of T_i such that (R_i, T_i, Δ_i) is feasible for the i–th system. It is immediate to check that the minimum of problem (3.31) is equal to

$$\text{subj. to:} \quad \begin{cases} \min\limits_{\vec{\Delta}} \; f(\vec{\Delta}) \\ \vec{\Delta} \in \mathcal{D} \\ R_{\min}(\vec{\Delta}) \leq \mathcal{R}\,. \end{cases} \tag{3.32}$$

For, the solution of problem (3.31) can be organized in the following five steps:

1. For each plant, determine the function $R_{\min}^{(i)}(\Delta)$ and consider problem (3.32);
2. Solve problem (3.32) and find a minimizing vector $\vec{\Delta}^*$;
3. Assign the bandwidth $R_{\min}^{(i)}(\Delta_i^*)$ to the i–th system;
4. Choose the sampling periods T_i's so that $\forall\, i = 1, \ldots, n$, the triple $\left(R_{\min}^{(i)}(\Delta_i^*), T_i, \Delta_i^*\right)$ is feasible;
5. Determine the corresponding control laws.

In order to compute the function $R_{\min}^{(i)}(\Delta)$, the characterization of the feasible triples is needed. Thus, once the first and the second steps of the above list are solved, the other steps directly follow. The determination of $R_{\min}^{(i)}(\Delta)$ is the only step involving issues from the control theory and quantization. This is the reason why, in this section, we go into the details of step 1 only. As far as the solution of problem (3.32) is concerned, we limit ourselves to mention that, once the function $R_{\min}^{(i)}(\Delta)$ is available, the problem can be algorithmically solved with a standard branch and bound scheme [70] and each step of the algorithm requires the (numerical) solution of a non–linear scalar equation. For the cost function in equation (3.30), the details of the algorithm can be found in Section 2.4 of [96].

Before going into the characterization of the feasible triples and into the computation of the function $R^{(i)}_{\min}(\Delta)$, let us illustrate a simple example which is helpful for the understanding of the problem setting.

Example 10 (Tracking of an unknown reference) *Consider N agents moving on a line. Denote by $\tilde{y}_i(\tau)$ the position of the i–th agent at time τ and assume that its dynamics is $\dot{\tilde{y}}_i(\tau) = \tilde{u}_i(\tau)$. Let $r(\tau)$ be an unknown reference to track and suppose that $|\dot{r}(\tau)| \leq \frac{\mathrm{w}}{2}$. A camera takes the measures of the displacements $\tilde{x}_i(\tau) := \tilde{y}_i(\tau) - r(\tau)$ of the agents and sends the quantized control values $\tilde{u}_i$'s to the actuators (i.e., to each agent) through a shared channel of capacity $\mathcal{R}$. The resulting dynamics of the displacement of the i–th agent is*

$$\dot{\tilde{x}}_i(\tau) = \tilde{u}_i(\tau) - \dot{r}(\tau)$$

so that the problem is modelled by system (3.26) with $a_i = 0 \; \forall\, i = 1, \ldots, N$. ♣

Single plant analysis: feasibility

This section is devoted to the characterization of the feasible triples (R, T, Δ) for a given system $\Sigma(a, \epsilon, \mathrm{w})$.

Given $\Delta > 0$ and $T > 0$, suppose that there exists a control set $\mathcal{U} \subset \epsilon\,\mathbb{Z}$ rendering $I(\Delta)$ w–controlled invariant for system $\Sigma(a, \epsilon, \mathrm{w}, T)$. Let $\ell(\Delta, T) \in \mathbb{N}$ be the minimum of the cardinality of the control sets $\mathcal{U} \subset \epsilon\,\mathbb{Z}$ rendering $I(\Delta)$ w–controlled invariant. By the definition of feasibility, it immediately follows that a triple (R, T, Δ) is feasible for system $\Sigma(a, \epsilon, \mathrm{w})$ if and only if

$$R \geq \frac{1}{T}\left\lceil \log_2 \ell(\Delta, T) \right\rceil . \tag{3.33}$$

This condition is an effective tool to solve the feasibility problem once the expression for the function $\ell(\Delta, T)$ is determined. Let us start with the characterization of the domain $\mathcal{D}(\ell)$ of the function $\ell(\Delta, T)$. Given $\Delta > 0$, assume that $\mathcal{U} = \epsilon\,\mathbb{Z}$ and let

$$\mathcal{T}(\Delta) := \left\{ T > 0 \,|\, I(\Delta) \text{ is w–controlled invariant for system } \Sigma(a, \epsilon, \mathrm{w}, T) \right\} .$$

The domain of $\ell(\Delta, T)$ is

$$\mathcal{D}(\ell) = \left\{ (\Delta, T) \,|\, T \in \mathcal{T}(\Delta) \right\} .$$

The following result is the main step to determine the set $\mathcal{T}(\Delta)$:

Proposition 5 (w–invariance conditions) *Consider system $\Sigma(a, \epsilon, \mathrm{w}, T)$: if $\mathrm{w} > 0$ and $\mathcal{U} = \epsilon\,\mathbb{Z}$, then*

i) if $a < 0$, $I(\Delta)$ is w–controlled invariant if and only if

$$\Delta \geq \min\left\{ \tfrac{\beta \mathrm{w}}{1-\alpha} \,;\, \beta\,(\epsilon + \mathrm{w}) \right\} = \min\left\{ \tfrac{\mathrm{w}}{|a|} \,;\, \tfrac{e^{aT}-1}{a}\,(\epsilon + \mathrm{w}) \right\} ;$$

ii) if $a \geq 0$, $I(\Delta)$ is w–controlled invariant if and only if

$$\Delta \geq \beta\,(\epsilon + \mathrm{w}) = \begin{cases} T \cdot (\epsilon + \mathrm{w}) & \text{if} \quad a = 0 \\ \frac{e^{aT}-1}{a}\,(\epsilon + \mathrm{w}) & \text{if} \quad a > 0 \,. \end{cases}$$

Proof. System $\Sigma(a,\epsilon,\mathrm{w},T)$ can be rewritten in the form

$$\begin{cases} x^+ = \alpha x + \hat{u} + w \\ \hat{u} \in \beta\epsilon\mathbb{Z} \\ w \in I(\beta\cdot\mathrm{w})\,. \end{cases}$$

Thus, $I(\Delta)$ is w–controlled invariant for system $\Sigma(a,\epsilon,\mathrm{w},T)$ if and only if $\forall\, x \in I(\Delta)$, $\exists\,\hat{u}\in\beta\epsilon\mathbb{Z}$ such that $\forall\, w \in I(\beta\cdot\mathrm{w})$ it holds that $|\alpha x+\hat{u}+w| \le \frac{\Delta}{2}$. This is equivalent to requiring that $\forall\, x\in I(\Delta)$, $\exists\,\hat{u}\in\beta\epsilon\mathbb{Z}$ such that $|\alpha x+\hat{u}| \le \frac{\Delta-\beta\mathrm{w}}{2}$. Therefore, with

$$\hat{\xi}(\Delta) := \max_{x\in I(\Delta)} \min_{\hat{u}\in\beta\epsilon\mathbb{Z}} |\alpha x+\hat{u}|\,,$$

it holds that $I(\Delta)$ is w–controlled invariant for system $\Sigma(a,\epsilon,\mathrm{w},T)$ if and only if $\hat{\xi}(\Delta) \le \frac{\Delta-\beta\mathrm{w}}{2}$, that is

$$\xi(\Delta) \le -\frac{\beta\mathrm{w}}{2}\,, \tag{3.34}$$

where $\xi(\Delta) := \hat{\xi}(\Delta) - \frac{\Delta}{2}$. It is straightforward to figure out that

$$\hat{\xi}(\Delta) = \begin{cases} \frac{\alpha}{2}\Delta & \text{if} \quad 0<\Delta\le\frac{\beta\epsilon}{\alpha} \\ \frac{\beta\epsilon}{2} & \text{if} \quad \Delta\ge\frac{\beta\epsilon}{\alpha}\,. \end{cases}$$

Thus, $\lim_{\Delta\to 0^+}\xi(\Delta)=0$, $\lim_{\Delta\to+\infty}\xi(\Delta)=-\infty$ and, for $\Delta\in\left]0\,,\frac{\beta\epsilon}{\alpha}\right]$, $\xi(\Delta)=\frac{\alpha-1}{2}\Delta$.

i) If $a<0$, then $0<\alpha<1$ and $\xi(\Delta)$ is a decreasing function. Therefore, there exists $\widetilde{\Delta}>0$ such that condition (3.34) is satisfied $\forall\,\Delta\ge\widetilde{\Delta}$. By the expression of $\xi(\Delta)$, it is a trivial computation to find that, if $-\frac{\beta\mathrm{w}}{2}\ge\xi\big(\frac{\beta\epsilon}{\alpha}\big)$, then $\widetilde{\Delta}=\frac{\beta\mathrm{w}}{1-\alpha}$; otherwise $\widetilde{\Delta}=\beta(\epsilon+\mathrm{w})$. Since $-\frac{\beta\mathrm{w}}{2}\ge\xi\big(\frac{\beta\epsilon}{\alpha}\big)=\frac{\beta\epsilon(\alpha-1)}{2\alpha}$ if and only if $\frac{\beta\mathrm{w}}{1-\alpha}\le\beta(\epsilon+\mathrm{w})$, then $\widetilde{\Delta}=\min\big\{\frac{\beta\mathrm{w}}{1-\alpha}\,;\,\beta(\epsilon+\mathrm{w})\big\}$.

ii) If $a\ge 0$, then $\alpha\ge 1$ and, for $\Delta\in\left]0\,,\frac{\beta\epsilon}{\alpha}\right]$, $\xi(\Delta)\ge 0$. Hence, in order that inequality (3.34) be satisfied, it is necessary that $\Delta>\frac{\beta\epsilon}{\alpha}$: with $\hat{\xi}(\Delta)=\frac{\beta\epsilon}{2}$, one finds that the invariance condition is $\Delta\ge\beta(\epsilon+\mathrm{w})$. ■

Remark 9 *Differently from the case considered in Section 3.1.1, because of the presence of the noise term, also for $a\le 0$ (namely, for $|\alpha|\le 1$) the invariance problem is meaningful. If $\mathrm{w}=0$, it is interesting to notice that, while for $a\ne 0$ the invariance conditions provided by Proposition 5 coincide with those we found in Section 3.1.1 for uniformly quantized control sets, this is not true for $a=0$. In fact, if $a=0$ and $\mathrm{w}=0$, then $u=0$ guarantees the invariance of any subset of $\mathbb{R}$; instead, with $\mathrm{w}=0$, the condition of Proposition 5.ii becomes $\Delta\ge\beta\epsilon$. This fact is not a contradiction (in fact, in Proposition 5 we assumed $\mathrm{w}>0$) and points out that, for a marginally stable system in the presence of input quantization, as $\mathrm{w}\ge 0$ varies, the size of the minimal invariant interval $I(\Delta)$ is* discontinuous *in $\mathrm{w}=0$.*

$\mathcal{T}(\Delta)$, and hence the domain of the function $\ell(\Delta,T)$, is determined by solving for T the invariance conditions provided by Proposition 5:

Corollary 3 *Consider system* $\Sigma(a,\epsilon,\mathrm{w},T)$, *assume* $\mathrm{w}>0$ *and that* $\mathcal{U}=\epsilon\,\mathbb{Z}$. *For* $\Delta>0$, *the following facts hold:*

i) if $a<0$ *and* $\Delta\geq\frac{\mathrm{w}}{|a|}$, $I(\Delta)$ *is* w*–controlled invariant* $\forall\,T>0$; *if* $\Delta<\frac{\mathrm{w}}{|a|}$, $I(\Delta)$ *is* w*–controlled invariant if and only if*

$$T\leq\min\left\{\,\tfrac{1}{a}\,\log\left(1-\tfrac{|a|\,\Delta}{\epsilon+\mathrm{w}}\right);\,\tfrac{1}{|a|}\,\log\left(1+\tfrac{\mathrm{w}}{\epsilon}\right)\,\right\};$$

ii) if $a=0$, $I(\Delta)$ *is* w*–controlled invariant if and only if*

$$T\leq\frac{\Delta}{\epsilon+\mathrm{w}};$$

iii) if $a>0$, $I(\Delta)$ *is* w*–controlled invariant if and only if*

$$T\leq\frac{1}{a}\,\log\left(1+\frac{a\,\Delta}{\epsilon+\mathrm{w}}\right).$$ □

Before we proceed to the calculation of $\ell(\Delta,T)$, we derive a lower bound for ℓ from which necessary conditions for the feasibility of a triple (R,T,Δ) are obtained. To this aim, let us introduce an invariance criterion holding for systems $\Sigma(a,\epsilon,\mathrm{w},T)$: for $\Delta>0$ and $u\in\epsilon\mathbb{Z}$, let

$$\begin{aligned}X_u:=&\ \left\{x\in\mathbb{R}\,|\,\forall\,w\in I(\beta\cdot\mathrm{w}),\ x^+=\alpha\,x+\beta\,u+w\in I(\Delta)\right\}=\\ =&\ \left\{x\in\mathbb{R}\,|\,-\tfrac{1}{\alpha}(\tfrac{\Delta}{2}+\beta\,u-\tfrac{\beta\,\mathrm{w}}{2})\leq x\leq\tfrac{1}{\alpha}(\tfrac{\Delta}{2}-\beta\,u-\tfrac{\beta\,\mathrm{w}}{2})\right\}.\end{aligned}\tag{3.35}$$

Consider

$$\begin{cases}\underline{x}_u:=-\frac{1}{\alpha}\big(\frac{\Delta}{2}+\beta\,u-\frac{\beta\,\mathrm{w}}{2}\big)\\ \overline{x}_u:=\frac{1}{\alpha}\big(\frac{\Delta}{2}-\beta\,u-\frac{\beta\,\mathrm{w}}{2}\big)\ :\end{cases}\tag{3.36}$$

if $X_u\neq\emptyset$, then $X_u=[\underline{x}_u\,,\,\overline{x}_u]$ with Lebesgue measure $\mu(X_u)=\frac{\Delta-\beta\,\mathrm{w}}{\alpha}$. It is straightforward to see that $I(\Delta)$ is w–controlled invariant if and only if[2]

$$\bigcup_{u\in\mathcal{U}}X_u\supseteq I(\Delta).\tag{3.37}$$

Hence, if $I(\Delta)$ is w–controlled invariant for system $\Sigma(a,\epsilon,\mathrm{w},T)$ with $\mathcal{U}\subset\epsilon\mathbb{Z}$, then $\Delta\leq\mu\Big(\bigcup_{u\in\mathcal{U}}X_u\Big)\leq\#\mathcal{U}\cdot\frac{\Delta-\beta\,\mathrm{w}}{\alpha}$. That is,

$$\#\mathcal{U}\geq\left\lceil\frac{\Delta\,\alpha}{\Delta-\beta\,\mathrm{w}}\right\rceil.\tag{3.38}$$

In particular,

$$\#\mathcal{U}\geq\lceil\alpha\rceil=\left\lceil e^{aT}\right\rceil.\tag{3.39}$$

[2]Both X_u and the invariance condition (3.37) can be formulated in the more general case $\mathcal{U}\subseteq\mathbb{R}$. Thus, condition (3.37) is an extension of Lemma 7 in Section 3.1.2 to systems affected by bounded noise.

Proposition 6 (Necessary conditions of feasibility) *Consider system* $\Sigma(a,\epsilon,\mathrm{w})$*: if the triple* (R,T,Δ) *is feasible, then*

i) $R \geq \frac{a}{\log 2} + \frac{1}{T}\log_2 \frac{\Delta}{\Delta-\beta(T)\cdot \mathrm{w}}$.

If moreover, $a>0$*, then:*

ii) $T \geq \frac{1}{R}$ *;*

iii) $\Delta \geq \frac{\epsilon+\mathrm{w}}{a}\left(e^{\,a/R}-1\right)$.

Proof. See in Appendix A.2.2 . ■

Remark 10 *When* $\mathrm{w}=0$ *and* $a>0$*, Proposition 6.i provides the well known bound*

$$R \geq \frac{a}{\log 2} \tag{3.40}$$

(see [5]). The same bound is approached when $\Delta \gg \beta(T)\cdot \mathrm{w}$.
The condition in Proposition 6.ii explicitly shows that the presence of a communication constraint induces a lower bound on the sampling period. Therefore, the possibility of approximating any continuous–time signal by switching very fast between discrete values is inhibited.

We pass now to the exact computation of the function $\ell(\Delta,T)$: we mainly address the case of an unstable plant $(a>0)$, which indeed is the most interesting as far as the design of the sampling interval T is concerned.

Proposition 7 (Computation of $\ell(\Delta,T)$ and of the corresponding $\mathcal{U}$) *Consider system* $\Sigma(a,\epsilon,\mathrm{w},T)$*, assume that* $\mathrm{w}>0$ *and* $a\geq 0$*. Let* $\Delta>0$ *be such that* $T\in\mathcal{T}(\Delta)$*, then*

$$\ell(\Delta,T) = 1 + \left\lceil \frac{2\left\lceil \frac{1}{2\epsilon}\left(\frac{\Delta(\alpha-1)}{\beta}+\mathrm{w}\right)\right\rceil}{\left\lfloor \frac{\Delta-\beta\mathrm{w}}{\beta\epsilon}\right\rfloor}\right\rceil , \tag{3.41}$$

where the dependence on T *is implicit in* α *and* β*. A control set* $\mathcal{U}\subset\epsilon\mathbb{Z}$ *of minimal cardinality* $\ell(\Delta,T)$ *between those ensuring the* w*–controlled invariance of* $I(\Delta)$ *is*

$$\mathcal{U} = \{u_1 < u_2 < \cdots < u_{\ell(\Delta,T)}\},$$

where

$$\begin{cases} u_1 := -\left\lceil \frac{1}{2\epsilon}\left(\frac{\Delta(\alpha-1)}{\beta}+\mathrm{w}\right)\right\rceil\cdot\epsilon \\ u_{k+1} := u_k + \left\lfloor \frac{\Delta-\beta\mathrm{w}}{\beta\epsilon}\right\rfloor\cdot\epsilon \qquad \big(\textit{for } k=1,\ldots,\ell(\Delta,T)-1\big). \end{cases} \tag{3.42}$$

Moreover, the invariance of $I(\Delta)$ *is ensured by any discrete time control law* $q:\mathbb{R}\to\mathcal{U}$ *such that*

$$\forall\, x\in I(\Delta),\quad q(x)\in\{u_k\in\mathcal{U}\,|\,x\in X_{u_k}\}. \tag{3.43}$$

Proof. See in Appendix A.2.2 . ■

For unstable plants (i.e., $a > 0$), let us make explicit the dependence of ℓ from the sampling period T:

$$\begin{cases} \ell(\Delta, T) = 1 + \left\lceil \dfrac{2\left\lceil \frac{a\Delta + \mathrm{w}}{2\,\epsilon} \right\rceil}{\left\lfloor \frac{a\Delta}{\epsilon\,(e^{aT}-1)} - \frac{\mathrm{w}}{\epsilon} \right\rfloor} \right\rceil \\ T \in \mathcal{T}(\Delta) \;\; \text{i.e.,} \;\; 0 < T \leq \frac{1}{a}\,\log\left(1 + \frac{a\,\Delta}{\epsilon + \mathrm{w}}\right). \end{cases} \tag{3.44}$$

Single plant analysis: determination of $R_{\min}(\Delta)$

We pass now to the determination of the smallest bit–rate R ensuring that, for a given $\Delta > 0$, there exists a choice of T such that the triple (R, T, Δ) is feasible. That is,

$$R_{\min}(\Delta) := \min\left\{R > 0 \,|\, \exists\, T \text{ such that } (R, T, \Delta) \text{ is feasible}\right\}.$$

For the sake of brevity, the analysis will henceforth be restricted to the case $a > 0$. Let

$$T_{\max}(\Delta) := \frac{1}{a}\,\log\left(1 + \frac{a\,\Delta}{\epsilon + \mathrm{w}}\right),$$

by condition (3.33) and Corollary 3.*iii*,

$$R_{\min}(\Delta) = \min_{T \in \left]\, 0\,,\, T_{\max}(\Delta)\,\right]} \frac{1}{T}\left\lceil \log_2 \ell(\Delta, T) \right\rceil. \tag{3.45}$$

Given $\Delta > 0$, the mapping $\ell(\Delta, T)$ is piecewise constant with T. Hence, the local minima of the "*channel occupation*" function $\frac{1}{T}\left\lceil \log_2 \ell(\Delta, T) \right\rceil$ are taken in correspondence of discontinuity points of $\ell(\Delta, T)$. The discontinuity points $T_1 < T_2 < \cdots < T_k$ can be determined using equation (3.44), thus the local minima of the channel occupation function can be listed. However, a closed formula for $R_{\min}(\Delta)$ is difficult to work out. We hence provide an expression which is a good estimate of $R_{\min}(\Delta)$.

Proposition 8 *Consider system $\Sigma(a, \epsilon, \mathrm{w})$ and assume that $a > 0$. For $\Delta > 0$, a sufficient condition on R in order that the triple (R, T, Δ) is feasible for some $T > 0$ is*

$$R \geq R^{\mathrm{suf}}_{\min}(\Delta) := \frac{a}{\log\left(1 + \frac{a\Delta}{2\,\epsilon\left\lceil \frac{a\Delta + \mathrm{w}}{2\,\epsilon} \right\rceil + \mathrm{w}}\right)}.$$

Proof. We show that, indeed, $R^{\mathrm{suf}}_{\min}(\Delta) = \frac{1}{T_1}\left\lceil \log_2 \ell(\Delta, T_1) \right\rceil$, where T_1 is the first discontinuity point of $\ell(\Delta, T)$. Let us compute T_1: for a given $\Delta > 0$, the argument of the floor in the denominator of equation (3.44) is a decreasing function of T, in particular $\ell(\Delta, T)$ is non–decreasing with T. Hence, to determine T_1, it is sufficient to find the largest T such that the denominator in equation (3.44) is greater than or equal to the numerator, that is to solve

$$\frac{a\,\Delta}{\epsilon\,(e^{aT} - 1)} - \frac{\mathrm{w}}{\epsilon} = 2\left\lceil \frac{a\,\Delta + \mathrm{w}}{2\,\epsilon} \right\rceil.$$

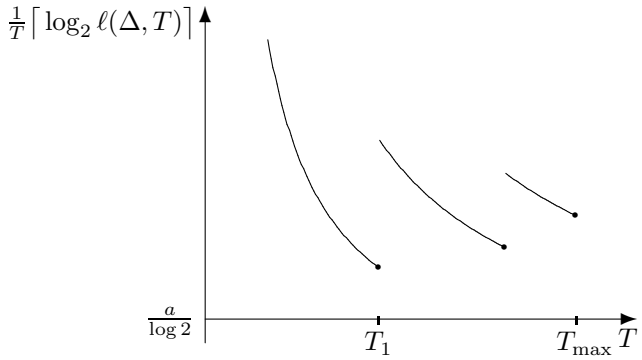

Figure 3.9: Graph of $\frac{1}{T}\lceil \log_2 \ell(\Delta,T)\rceil$ for $a=2\,,\ \epsilon=1\,,\ \mathrm{w}=1.8$ and $\Delta=2\,.$

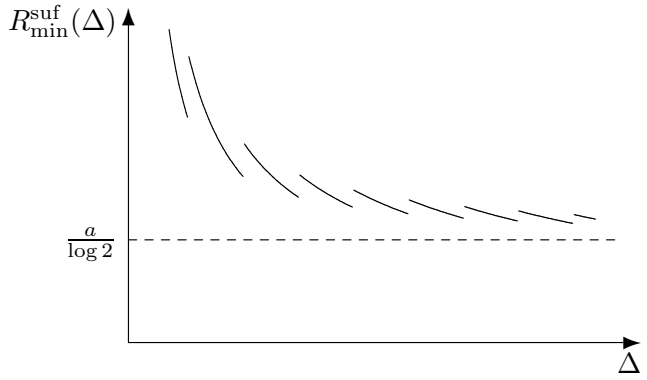

Figure 3.10: The graph of $R^{\mathrm{suf}}_{\min}(\Delta)$ for the system of Fig. 3.9.

Hence,

$$T_1 = \frac{1}{a}\log\left(1+\frac{a\Delta}{2\,\epsilon\left\lceil\frac{a\Delta+\mathrm{w}}{2\,\epsilon}\right\rceil+\mathrm{w}}\right).$$

Accordingly,

$$\forall\, T\in\,]\,0\,,\,T_1]\,,\quad \ell(\Delta,T)=2 \tag{3.46}$$

so that $\frac{1}{T_1}\lceil \log_2 \ell(\Delta,T_1)\rceil = \frac{1}{T_1}\,.$ ■

Remark 11 *The effect of the noise on the system is governed by the function $\beta(T)$ which grows exponentially: hence, as the sampling interval T increases, the system is more and more affected by the noise. It is then natural to expect that the local minimum in correspondence of the first discontinuity point T_1 is close to the actual minimum of the function (see Fig. 3.9). This fact can be explicitly verified for $a\,\Delta \gg \max\{\epsilon\,,\,\mathrm{w}\}$, in fact (see also Fig. 3.10):*

$$\lim_{\Delta\to+\infty} R^{\mathrm{suf}}_{\min}(\Delta) = \frac{a}{\log 2}\,,$$

namely, as $\Delta \to +\infty$, $R^{\mathrm{suf}}_{\mathrm{min}}(\Delta)$ *approaches the theoretical lower bound on* R *(see Proposition 6.ι and Remark 10).*

By equation (3.46), the feasible triple $\left(R^{\mathrm{suf}}_{\mathrm{min}}(\Delta), T_1, \Delta\right)$ leads to the implementation of a binary feedback law. It has been shown in [72] that the strategy of considering a binary control law and a short sampling interval T is the most robust with respect to uncertainties on the really available bandwidth.
For a given $\Delta > 0$, if R is not much larger than $R_{\mathrm{min}}(\Delta)$, the set of values of T such that (R, T, Δ) is feasible consists of disjoint intervals whose right extremes are discontinuity points of the function $\ell(\Delta, T)$ (see Fig. 3.9). A criterion to discriminate the allowed values for T, apart from robustness arguments, should take the entailed channel occupation $\frac{1}{T}\lceil \log_2 \ell(\Delta, T)\rceil$ into account.

3.1.4 Controlled invariant ellipsoids: multi–input

There is a strict relation between Lyapunov theory and invariance [11, 12]: for instance, under suitable assumptions on the considered dynamical system, from the knowledge of an invariant set it is possible to derive a Lyapunov function for the system. Conversely, if $V(x)$ is a Lyapunov function for a system $x^+ = f(x)$, a family of invariant sets is canonically associated to V. In fact, any sublevel set of V, namely, any set of the type $\{x \in \mathbb{R}^n \,|\, V(x) \leq v\}$, $v \geq 0$, is invariant. For a system $x^+ = f(x, u)$, control Lyapunov functions are still helpful to determine controlled invariant sets: in this case the controlled invariance analysis and the control synthesis problem are not entirely separated. A common technique to find controlled invariant sets consists of the synthesis of a control law $u(\cdot)$ so that a Lyapunov function V is available for the closed loop system; then, the sublevel sets of V form a family of controlled invariant sets for the open loop system. Once a controlled invariant set has been found in this way, the synthesis of a control law ensuring the invariance of such a set is not bound to the control law $u(\cdot)$ which, thus, may be viewed in general as an auxiliary tool.
Unfortunately, the relation between Lyapunov theory and controlled invariance becomes weaker in the presence of quantization. There are several mathematical reasons for that, most of them lie in the fact that major characteristics of the structure of the system are lost because of quantization. Indeed, not only quantization introduces nonlinearity in the system, but also convexity properties are lost (a quantized set is not convex). Even more substantially, since closed loop asymptotic stability cannot be achieved for quantized systems, then it is not possible to construct a control Lyapunov function. Therefore, classical techniques based on Lyapunov theory must be properly revised in order to deal with quantized systems. In this section we present results extending the Lyapunov based approach to the quantized input case. As it is clarified in the second part of this section, this kind of analysis is quite conservative form the point of view of the search for the invariant set of minimal size. On the other hand these results will turn out to be fruitful for control synthesis purposes in the framework of small–gain theory (see Section 5.3).
Consider a linear system $x^+ = Ax + Bu$, where the pair (A, B) is stabilizable and $u \in \mathcal{U} =$

$\mathbb{R}^m$. For any fixed $K \in \mathbb{R}^{m\times n}$ such that $A+BK$ is Schur, let $\mathbb{R}^{n\times n} \ni P > 0$ be a solution of the Lyapunov inequality

$$(A+BK)'P(A+BK) - P < 0\,.$$

The function $V(x) := x'Px$ is a quadratic Lyapunov function for the closed loop system $x^+ = (A+BK)x$, therefore any sublevel set of V is positively invariant for the closed loop dynamics. That is, $\forall\, r \geq 0$, the ellipsoid $\mathcal{E}_{P,r^2}$ is a controlled invariant set for the original system $x^+ = Ax + Bu$.
With the suitable corrections, this technique can be used to derive controlled invariant ellipsoids also when the control set is quantized. In fact, consider a stabilizing matrix K and $P > 0$ as above, and the quantized control law $u(x) = q_{\mathcal{U}}(Kx)$, for some input quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$. The corresponding closed loop system is

$$x^+ = (A+BK)x + Bq_e(Kx)\,, \tag{3.47}$$

where $q_e(Kx) = q_{\mathcal{U}}(Kx) - Kx$ is the quantization error. In this case, $V(x) = x'Px$ is no more a Lyapunov function for the system. Nevertheless, if x is such that the norm of the quantization error $q_e(Kx)$ is not too large with respect to the norm of x, we may still expect that $V(x^+) - V(x) \leq 0$. Therefore, for sufficiently large $r > 0$, the sublevel set $\mathcal{E}_{P,r^2}$ could be positively invariant for system (3.47), hence controlled invariant for the original system $x^+ = Ax + Bu$. Proposition 9 below makes precise this argument in the case where the quantization error is uniformly bounded. Actually, it is clear the intuition that, for this kind of argument, the meaningful quantity is the *relative* quantization error $\frac{\|q_e(Kx)\|_2}{\|Kx\|_2}$. However, the study in terms of the relative quantization error is more tightly intersecting control synthesis problems than the study in the case of bounded *absolute* quantization error. Therefore, we put off this more general case till Section 5.3, where the controlled invariance analysis under uniformly bounded relative quantization error follows from control synthesis results based on a small–gain theorem (see Remark 25 at the end of Section 5.3.2).

Proposition 9 (Uniformly bounded error) *Consider the system*

$$x^+(t) = Fx(t) + Be(t)\,, \tag{3.48}$$

assume that F is Schur and that $\forall\, t \geq 0$, $\|e(t)\|_2 \leq E_0$. For any $\mathbb{R}^{n\times n} \ni S > 0$, let P be the solution of the Lyapunov equation

$$F'PF - P = -S \tag{3.49}$$

and

$$\begin{cases} r_{\mathrm{i}}^2 := R^2\big(\lambda_{\max}(P-S) + \lambda_{\min}(S)\big)\,, \ \textit{where} \\ R = \frac{E_0}{\lambda_{\min}(S)}\,\alpha(P)\,, \ \textit{and} \\ \alpha(P) = \|F'PB\|_2 + \sqrt{\|F'PB\|_2^2 + \lambda_{\min}(S)\|B'PB\|_2}\,. \end{cases} \tag{3.50}$$

Then, $\forall\, r^2 \geq r_{\mathrm{i}}^2$, $\mathcal{E}_{P,r^2}$ is invariant.

Proof of Proposition 9. The initial part of the proof mimics arguments used in [17]. Let $V(x) := x'Px$, where $P > 0$ is the solution of the Lyapunov equation (3.49) for some $\mathbb{R}^{n\times n} \ni S > 0$. Then,

$$\begin{aligned}\Delta V(x) := \;& V(x^+) - V(x) = \\ =\;& -x'Sx + 2x'F'PBe + e'B'PBe \leq \\ \leq\;& -x'Sx + 2|x'F'PBe| + |e'B'PBe| \leq \\ \leq\;& -\lambda_{\min}(S)\|x\|_2^2 + 2E_0\|F'PB\|_2 \cdot \|x\|_2 + E_0^2\,\|B'PB\|_2 := f(\|x\|_2)\,,\end{aligned}$$

where the last expression is obtained thanks to the Cauchy–Schwarz inequality. Notice that $f(\|x\|_2)$ is a second order polynomial having roots of opposite sign. Let us determine $R > 0$ such that $f(R) = 0$: it holds that

$$\begin{cases} R = \frac{E_0}{\lambda_{\min}(S)}\,\alpha(P)\,, \quad \text{where} \\ \alpha(P) = \|F'PB\|_2 + \sqrt{\|F'PB\|_2^2 + \lambda_{\min}(S)\|B'PB\|_2}\,. \end{cases} \tag{3.51}$$

Hence, for x such that $\|x\|_2 > R$, $\Delta V(x) \leq f(\|x\|_2) < 0$. Namely, as long as the state lies outside the closed ball $\mathcal{B}_R$, V is decreasing along the trajectory.
Let us consider the behavior of the trajectories starting from $\mathcal{B}_R$. With

$$M_0^2 := \max_{x\in\mathcal{B}_R} V(x^+)\,,$$

it holds that $\forall\, r^2 \geq M_0^2$, the ellipsoid $\mathcal{E}_{P,r^2}$ is invariant. In fact,

$$\mathcal{E}_{P,r^2} = (\mathcal{E}_{P,r^2} \cap \mathcal{B}_R) \cup (\mathcal{E}_{P,r^2} \setminus \mathcal{B}_R)$$

and $x^+ \in \mathcal{E}_{P,r^2}$ if and only if $V(x^+) \leq r^2$: if $x \in \mathcal{E}_{P,r^2} \cap \mathcal{B}_R$, then $V(x^+) \leq M_0^2 \leq r^2$; if instead $x \in \mathcal{E}_{P,r^2} \setminus \mathcal{B}_R$, then $V(x^+) < V(x) \leq r^2$.
It is hence sufficient to show that r_{i}^2 defined in equation (3.50) is an upper bound for M_0^2:

$$\begin{aligned}V(x^+) =\;& \Delta V(x) + V(x) = \\ =\;& x'(P-S)x + 2x'F'PBe + e'B'PBe \leq \\ \leq\;& \lambda_{\max}(P-S)\|x\|_2^2 + 2E_0\|F'PB\|_2 \cdot \|x\|_2 + E_0^2\|B'PB\|_2 := g(\|x\|_2)\,.\end{aligned}$$

By equation (3.49), $\lambda_{\max}(P-S) = \lambda_{\max}(F'PF) \geq 0$, therefore

$$\max_{\|x\|_2\leq R} g(\|x\|_2) = g(R)$$

and

$$M_0^2 = \max_{x\in\mathcal{B}_R} V(x^+) \leq \max_{x\in\mathcal{B}_R} g(\|x\|_2) = g(R)\,.$$

Since $g(\|x\|_2) - f(\|x\|_2) = \lambda_{\max}(P-S)\|x\|_2^2 + \lambda_{\min}(S)\|x\|_2^2$ and $f(R) = 0$, we get

$$g(R) = R^2\big(\lambda_{\max}(P-S) + \lambda_{\min}(S)\big) = r_{\mathrm{i}}^2 \tag{3.52}$$

and this concludes the proof. ■

Remark 12 *The fact that practical stability properties hold for system* (3.48) *under bounded inputs is a straightforward consequence of the fact that a stable linear system is input–to–state stable [62]. The contribution of Proposition 9 consists in providing a quantitative invariance analysis for ellipsoids. Actually, Proposition 9 can be easily extended to include also the analysis of convergence properties: this aspect is postponed to Chapter 5 (see Proposition 14 in Section 5.3.1) which is specifically devoted to the convergence issue in practical stabilization.*

Let us show how Proposition 9 can be used to determine controlled invariant sets for a quantized input linear system. Consider a system

$$\begin{cases} x^+ = Ax + Bu \\ x \in \mathbb{R}^n, \quad u \in \mathcal{U} \subset \mathbb{R}^m, \end{cases} \tag{3.53}$$

where (A, B) is a stabilizable pair. Assume that the input quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ is such that the corresponding quantization error is uniformly bounded, namely

$$E_0 := \sup_{y \in \mathbb{R}^m} \|q_{\mathcal{U}}(y) - y\|_2 < +\infty \,. \tag{3.54}$$

Consider any $K \in \mathbb{R}^{m \times n}$ such that $F := A + BK$ is Schur. Let $u(x) := q_{\mathcal{U}}(Kx)$: the resulting closed loop dynamics is described by system (3.47). The assumption (3.54) ensures that $\forall\, K \in \mathbb{R}^{m \times n}$ and $\forall\, x \in \mathbb{R}^n$, $\|q_e(Kx)\|_2 \leq E_0$, therefore the closed loop system satisfies the hypotheses of Proposition 9. Accordingly, for any $\mathbb{R}^{n \times n} \ni S > 0$, let P be the solution of the Lyapunov equation (3.49) and r_{i}^2 be defined as in (3.50). Then, $\forall\, r^2 \geq r_{\mathrm{i}}^2$, the ellipsoid $\mathcal{E}_{P,r^2}$ is invariant for the closed loop system (3.47) and, in particular, it is controlled invariant for system (3.53).
Uniform quantizers (see Definition 7 in Section 2.1) are the typical examples where the corresponding quantization error is uniformly bounded and the proposed analysis technique can be applied.

Example 11 *Consider the quantized input system*

$$\begin{cases} x^+ = Ax + Bu = \begin{pmatrix} 2 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & -4 & 4 \end{pmatrix} x + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} u \\ u \in \mathcal{U} = \mathbb{Z}^2 \subset \mathbb{R}^2. \end{cases} \tag{3.55}$$

The pair (A, B) is reachable, thus stabilizable. However, denoted by $B^{[i]}$ the i–th column of B, both the pairs $(A, B^{[1]})$ and $(A, B^{[2]})$ are neither reachable nor stabilizable.
Following the theory developed above, let us provide controlled invariant ellipsoids.
Consider the quantized control law $u(x) := q_{\mathcal{U}}(Kx)$, where $q_{\mathcal{U}} : \mathbb{R}^2 \to \mathbb{Z}^2$ is a uniform quantizer. With the usual notation $F = A + BK$ and $q_e(Kx) = q_{\mathcal{U}}(Kx) - Kx \in \mathbb{R}^2$, the closed loop dynamics is

$$x^+ = Fx + Bq_e(Kx)\,, \tag{3.56}$$

where, according to Lemma 2, $\forall\, x \in \mathbb{R}^3,\ \|q_e(Kx)\|_2 \leq E_0 = \frac{\sqrt{2}}{2}$ *irrespective of* $K \in \mathbb{R}^{2\times 3}$. *Two choices for the matrix* K *are considered, in both cases* $u = Kx$ *is a deadbeat controller.*

Case 1: *with*

$$K = \begin{pmatrix} -2 & -1 & 1 \\ 0 & 4 & -4 \end{pmatrix},$$

the closed loop matrix

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

is such that $F^2 \neq 0$ *and* $F^3 = 0$. *With* $S = I$, *the Lyapunov equation* (3.49) *is solved by*

$$P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

According to equation (3.50), *elementary computations allow one to find* $r_{\mathrm{i}}^2 = \frac{27}{2}$. *Therefore,* $\forall\, r^2 \geq \frac{27}{2}$ *the ellipsoid*

$$\mathcal{E}_{P_1,r^2} = \left\{ x \in \mathbb{R}^3 \,|\, x_1^2 + 2x_2^2 + 3x_3^2 \leq r^2 \right\}$$

is invariant for system (3.56), *hence it is controlled invariant for system* (3.55).
The lengths of the semi–axes of the minimal invariant ellipsoid $\mathcal{E}_{P_1,r_{\mathrm{i}}^2}$ *are* $\mathrm{s}_j := \frac{r_{\mathrm{i}}}{\sqrt{\lambda_j(P)}}$, $j = 1, 2, 3$. *That is,*

$$\begin{cases} \mathrm{s}_1 = \frac{3}{2}\sqrt{6} \simeq 3.67 & \text{in the direction of } x_1 \\ \mathrm{s}_2 = \frac{3}{2}\sqrt{3} \simeq 2.6 & \text{in the direction of } x_2 \\ \mathrm{s}_3 = \frac{3}{2}\sqrt{2} \simeq 2.12 & \text{in the direction of } x_3\,. \end{cases}$$

Case 2: *if instead*

$$K = \begin{pmatrix} -2 & -2 & 1 \\ 0 & 4 & -4 \end{pmatrix},$$

the closed loop matrix

$$F = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

is such that $F^2 = 0$. *With* $S = I$, *the Lyapunov equation* (3.49) *is solved by*

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and it can be easily found that $r_{\mathrm{i}}^2 = 2$ *. Hence,* $\forall\, r^2 \geq 2$ *, the ellipsoid*

$$\mathcal{E}_{P_2,r^2} = \left\{ x \in \mathbb{R}^3 \,|\, x_1^2 + x_2^2 + 2x_3^2 \leq r^2 \right\}$$

is invariant for system (3.56) *, hence it is controlled invariant for system* (3.55) *.*
In this case, the lengths of the semi–axes of the minimal invariant ellipsoid $\mathcal{E}_{P_2,r_{\mathrm{i}}^2}$ *are:*

$$\begin{cases} \mathrm{s}_1 = \sqrt{2} \simeq 1.42 & \textit{in the direction of } x_1 \\ \mathrm{s}_2 = \sqrt{2} & \textit{in the direction of } x_2 \\ \mathrm{s}_3 = 1 & \textit{in the direction of } x_3\,. \end{cases}$$

In case 2, we found a minimal invariant ellipsoid which is smaller than in case 1. This is consistent with the fact that the considered control law is such that $F^2 = 0$ *while, in case 1 ,* $F^2 \neq 0$ *.* ♣

Single–input reachable systems: hypercubes vs ellipsoids

In this subsection, single–input reachable systems under uniformly quantized controls are considered and a quantitative comparison is provided between the result of the invariance analysis based on ellipsoids and on hypercubes. As explained at the beginning of Section 3.1.2 , with reference to the practical stabilization problem, a good mark for the considered family of sets is that it contain a *small* invariant set. Thus, the comparison involves different measures of the size of the minimal invariant set contained in each family. Both the volume, the containment relation and the diameter of the considered sets in specific directions are considered. Also the effect of the *cut–off* procedure (see equation (3.17)) is studied. It is shown that invariant ellipsoids are significantly more conservative than hypercubes and that also the cut–off procedure allows for significant improvements. The case of two dimensional systems is considered in full details. Some results are presented for the general case too and suggest that, as the state space dimension increases, ellipsoids are more and more conservative.

Let us consider a generic second order and single–input reachable system under uniform input quantization:

$$\begin{cases} x^+ = Ax + Bu = \begin{pmatrix} 0 & 1 \\ a_1 & a_2 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ u \in \mathbb{Z}\,. \end{cases} \tag{3.57}$$

By Example 7 in Section 3.1.1 , we know that $\forall\, \Delta \geq 1$, $Q_2(\Delta)$ is controlled invariant. By Proposition 1 , we also know that a qdb–controller makes these hypercubes (squares) positively invariant for the corresponding closed loop system. Let us analyze the invariance of ellipsoids when the system is controlled by a qdb–controller.
Let $q_{\mathcal{U}}$ be a nearest neighbor quantizer and let $u(x) = q_{\mathcal{U}}(Kx)$, where $K = (-a_1 \;\; -a_2)$. With $q_e(Kx) = q_{\mathcal{U}}(Kx) - Kx$, the closed loop dynamics is

$$x^+ = Fx + Bq_e(Kx) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} q_e(Kx) \tag{3.58}$$

and $\forall\, x \in \mathbb{R}^2\,,\ \big|q_e(Kx)\big| \le \frac{1}{2}$. Let us apply Proposition 9: for

$$\mathbb{R}^{2\times 2} \ni S = \begin{pmatrix} s_1 & s_3 \\ s_3 & s_2 \end{pmatrix} > 0 \tag{3.59}$$

(hence with $s_1 > 0\,,\ s_2 > 0$ and $s_1 s_2 - s_3^2 > 0$), it holds that

$$\lambda_{\min}(S) = \frac{s_1 + s_2 - \sqrt{(s_1 - s_2)^2 + 4s_3^2}}{2}\,. \tag{3.60}$$

Denote by $P(S)$ the solution of the Lyapunov equation (3.49), then

$$P(S) = \begin{pmatrix} s_1 & s_3 \\ s_3 & s_1 + s_2 \end{pmatrix}. \tag{3.61}$$

Therefore,

$$P(S) - S = \begin{pmatrix} 0 & 0 \\ 0 & s_1 \end{pmatrix}, \quad F'P(S)B = \begin{pmatrix} 0 \\ s_3 \end{pmatrix}, \quad B'P(S)B = s_1 + s_2\,,$$

consequently

$$R = \frac{1}{2\,\lambda_{\min}(S)}\Big(|s_3| + \sqrt{s_3^2 + \lambda_{\min}(S)(s_1 + s_2)}\,\Big) \tag{3.62a}$$

$$= \frac{1}{2\,\lambda_{\min}(S)}\Big(|s_3| + \sqrt{\lambda_{\min}^2(S) + s_1 s_2}\,\Big) \tag{3.62b}$$

and

$$r_{\rm i}^2(S) = R^2\big(s_1 + \lambda_{\min}(S)\big)\,. \tag{3.63}$$

Thus, for any given $\mathbb{R}^{2\times 2} \ni S > 0$ and $\forall\, r^2 \ge r_{\rm i}^2(S)$, the ellipsoid $\mathcal{E}_{P(S),r^2}$ is invariant for system (3.58), hence it is controlled invariant for system (3.57).

Let us compare the minimal invariant hypercube $Q_2(1)$ and the minimal invariant ellipsoid $\mathcal{E}_{P(S),r_{\rm i}^2}$ provided by Proposition 9 as the matrix $S > 0$ varies. The following result establishes that, $\forall\, S > 0$, the minimal invariant ellipsoid $\mathcal{E}_{P(S),r_{\rm i}^2}$ is larger than the hypercube $Q_2(1)$ both in terms of the volume and of the containment relation.

Proposition 10 *In the above setting it holds that:*

i) $\min\limits_{\mathbb{R}^{2\times 2}\ni S>0} \text{Area}\,(\mathcal{E}_{P(S),r_{\rm i}^2}) = \frac{\pi}{\sqrt{2}}$ *and such a minimum is achieved if and only if* $S = s_1 I$.

ii) $\forall\, S > 0\,,\ Q_2(1) \subset \mathcal{E}_{P(S),r_{\rm i}^2(S)}$.

Proof. See in Appendix A.2.3. ■

Thus, with the choice $S = I$ resulting in the controlled invariant ellipsoid of minimal area, one obtains $r_{\rm i}^2 = 1$ and

$$\mathcal{E}_{P(I),r_{\rm i}^2(I)} = \big\{x \in \mathbb{R}^2 \,|\, x_1^2 + 2x_2^2 \le 1\big\}\,.$$

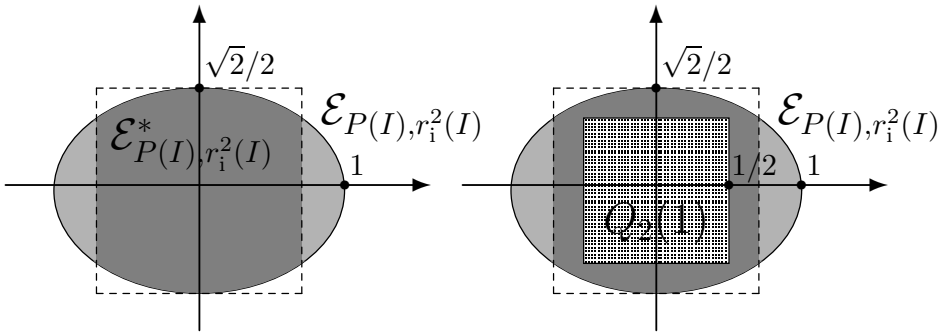

Figure 3.11: *Left :* The minimal invariant ellipse provided by Proposition 9 with the optimal choice of $S = I$ and, in darker grey, the smaller invariant set $\mathcal{E}^*_{P(I),r_{\rm i}^2(I)}$ obtained via the cut–off procedure. *Right:* The dotted region represents the minimal invariant hypercube.

The area reduction brought by considering the minimal invariant hypercube is really significant, in fact

$$\frac{\text{Area}\left(Q_2(1)\right)}{\text{Area}\left(\mathcal{E}_{P(I),r_{\rm i}^2(I)}\right)} = \frac{\sqrt{2}}{\pi} \simeq 0.45\,.$$

Let us quantify the effect of the cut–off procedure in terms of area reduction (see also Fig. 3.11) : it holds that $\text{Pr}_2\big(\mathcal{E}_{P(I),r_{\rm i}^2(I)}\big) = \left[-\frac{\sqrt{2}}{2}\,,\,\frac{\sqrt{2}}{2}\right]$ whereas $\text{Pr}_1\big(\mathcal{E}_{P(I),r_{\rm i}^2(I)}\big) = [-1\,,\,1]\,,$ therefore $\mathcal{E}^*_{P(I),r_{\rm i}^2(I)} = \mathcal{E}_{P(I),r_{\rm i}^2(I)} \cap Q_2(\sqrt{2})$ and[3]

$$\text{Area}\left(\mathcal{E}^*_{P(I),r_{\rm i}^2(I)}\right) = \frac{4}{\sqrt{2}}\int_0^{\sqrt{2}/2}\sqrt{1-x^2}dx = \frac{\pi/2+1}{\sqrt{2}}\,.$$

Hence,

$$\frac{\text{Area}\left(\mathcal{E}^*_{P(I),r_{\rm i}^2(I)}\right)}{\text{Area}\left(\mathcal{E}_{P(I),r_{\rm i}^2(I)}\right)} = \frac{1}{2} + \frac{1}{\pi} \simeq 0.82\,.$$

Namely, also taking the cut–off into account provide a tangible improvement in the invariance analysis.

We conclude the section with a brief study on the extension of the above comparison to n–th order single–input reachable systems.
Consider the n–th order version of the system in equation (3.57) . Also in this case, we know that $\forall\,\Delta \geq 1$, the hypercube $Q_n(\Delta)$ is controlled invariant. Let us analyze the invariance of ellipsoids when the system is controlled by a qdb–controller. For $S = I \in \mathbb{R}^{n\times n}$, the minimal invariant ellipsoid provided by Proposition 9 is

$$\mathcal{E}_{P(I),r_{\rm i}^2(I)} = \left\{x \in \mathbb{R}^n \mid \textstyle\sum_{j=1}^n j x_j^2 \leq \frac{n^2}{4}\right\},$$

[3]Indeed, $\int \sqrt{1-x^2}dx = \frac{1}{2}\left(x\sqrt{1-x^2} + \arcsin x\right)$.

and, for $j = 1, \ldots, n$, $\mathrm{Pr}_j\big(\mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}\big) = \big[-\frac{n}{2\sqrt{j}}, \frac{n}{2\sqrt{j}} \big]$. In particular,

$$\mathrm{diam}_n \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)} = \sqrt{n} \, .$$

Therefore, as the state space dimension n increases, while the magnitude (see Definition 17 in Section 3.1.2) of the minimal invariant hypercube remains constant equal to 1, the magnitude of $\mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}$ diverges.
A qualitative result on the analysis of the effect of the cut–off procedure can be obtained by considering the ratio between $\mathrm{diam}_1 \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}$ and $\mathrm{diam}_n \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}$. Indeed,

$$\mathcal{E}^*_{P(I),r_{\mathrm{i}}^2(I)} = \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)} \cap \Big(\mathrm{Pr}_n\big(\mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}\big) \Big)^n ,$$

so that $\mathrm{diam}_n \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}$ is the magnitude of the invariant set $\mathcal{E}^*_{P(I),r_{\mathrm{i}}^2(I)}$ and dictates the entity of the "cut", while, according to Proposition 2 in Section 3.1.2, the direction along the first coordinate is the one more affected by the cut–off procedure. Because

$$\frac{\mathrm{diam}_1 \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}}{\mathrm{diam}_n \, \mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}} = \sqrt{n} \, ,$$

also the effect of the cut–off procedure is more and more significant at the increasing of the state space dimension.
The same phenomenon is pointed out when analyzing the volume of the minimal invariant ellipsoid. In fact, such a volume, which is to be compared with the unitary volume of the minimal invariant hypercube $Q_n(1)$, is[4]

$$\mathrm{Volume}\,(\mathcal{E}_{P(I),r_{\mathrm{i}}^2(I)}) = \frac{(n\sqrt{\pi})^n}{2^n \, \Gamma\big(\frac{n}{2}+1\big)\sqrt{n!}} := \mathrm{V}(n)$$

and, using the Stirling's formula [50] to bound $n!$, it can be seen that $\lim_{n \to +\infty} \mathrm{V}(n) = +\infty$.

[4]Recall that the volume of the unit n–ball is $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$ and that $\Gamma(n+1) = n!$ (see the notation in Section 1.5).

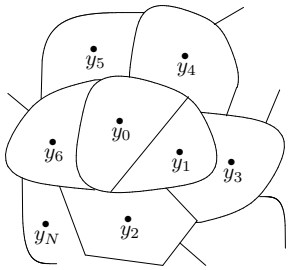

Figure 3.12: An example of state space partition induced by a state quantizer $q_{\mathcal{Y}}$.

3.2 Controlled invariance: quantized single–input and quantized measurements

In this section we study the $q_{\mathcal{Y}}$–controlled invariance problem for system (2.1) in the case where both the inputs and the measurements are quantized. We first consider the quantized state case, the theory is then extended to deal with quantized outputs. The latter case is handled by constructing a state–observer fed by the discrete output values of the system and returning a state space quantization. Because the state quantization obtained in this way is time–varying, the results from the quantized state case need to be further elaborated in order to be applied to the quantized output problem.
As in Section 3.1.1, throughout this section we assume that the pair (A, B) is reachable and that it is in the controller form (see assumption **A0** in equation (3.1)). Recall that

$$\alpha := \sum_{i=1}^{n} |a_i| .$$

3.2.1 Controlled invariant hypercubes: state quantization

In this section, we extend to the quantized state case the invariance analysis of hypercubes $Q_n(\Delta)$ developed in Section 3.1.1. According to the definition given in Section 2.2, this is the case in which, at least in a sufficiently large neighborhood of the equilibrium, the state space partition induced by $q_{\mathcal{Y}}$ is made of bounded sets, see Fig. 3.12. Namely, there exists a sufficiently large $r > 0$ such that $\forall\, y \in q_{\mathcal{Y}}(\mathcal{B}_r)$, $\mathcal{C}_y$ is bounded.
For a given $\Delta > 0$, consider the hypercube $Q_n(\Delta)$ and let

$$\mathcal{Y}(\Delta) := q_{\mathcal{Y}}\big(Q_n(\Delta)\big) \subseteq \mathcal{Y} \tag{3.64}$$

be the set of possible outputs when $x \in Q_n(\Delta)$. Consider also $\mathcal{C}_y^* := \mathcal{C}_y \cap Q_n(\Delta)$, where $\mathcal{C}_y = q_{\mathcal{Y}}^{-1}(y)$ (see equation (2.3)).
The $q_{\mathcal{Y}}$–controlled invariance of $Q_n(\Delta)$ is tantamount to requiring that $\forall\, y \in \mathcal{Y}(\Delta)$, $\exists\, u \in \mathcal{U}$ such that $A\mathcal{C}_y^* + Bu \subseteq Q_n(\Delta)$. By relation (3.4), this is equivalent to

$$\forall\, y \in \mathcal{Y}(\Delta)\,,\ \exists\, u \in \mathcal{U} \quad \text{such that} \quad \mathrm{Pr}_n(A\mathcal{C}_y^*) + u \subseteq \big[-\tfrac{\Delta}{2}\,,\,\tfrac{\Delta}{2}\big]\,. \tag{3.65}$$

As we did in Section 3.1.1 , we seek an algebraic relation to identify the $q_{\mathcal{Y}}$–controlled invariant hypercubes. To this aim, we have to introduce a quantity suited to describe the output quantizer $q_{\mathcal{Y}}$. For $\Delta > 0$ and $y \in \mathcal{Y}(\Delta)$, let $h^*(y) := \mathrm{diam}_n(A\mathcal{C}_y^*)$ and

$$H^*(\Delta) := \max_{y \in \mathcal{Y}(\Delta)} h^*(y) \, .$$

Moreover, let

$$\begin{cases} c^*_{\sup}(y) := \sup \{\mathrm{Pr}_n(A\mathcal{C}_y^*)\} \\ c^*_{\inf}(y) := \inf \{\mathrm{Pr}_n(A\mathcal{C}_y^*)\} \\ c^*_{\mathrm{mid}}(y) := \frac{c^*_{\sup}(y)+c^*_{\inf}(y)}{2} = c^*_{\inf}(y) + \frac{h^*(y)}{2} \ : \end{cases} \tag{3.66}$$

these quantities depend on Δ too.
Consider also

$$h(y) := \mathrm{diam}_n(A\mathcal{C}_y) \tag{3.67}$$

and

$$H(\Delta) := \sup_{y \in \mathcal{Y}(\Delta)} h(y) \, . \tag{3.68}$$

These quantities are more easily computable than $h^*(y)$ and $H^*(\Delta)$ as they avoid determining the intersection $\mathcal{C}_y^* = \mathcal{C}_y \cap Q_n(\Delta)$. But, as a consequence of the fact that $H(\Delta) \geq H^*(\Delta)$, they lead to a more conservative analysis. Notice that $H(\Delta)$ and $H^*(\Delta)$, which are defined in controller form coordinates, depend on A and are non–decreasing functions of Δ . Furthermore, since the state space partition induced by $q_{\mathcal{Y}}$ is locally finite, then $\forall\, \Delta_0 > 0$, $H(\Delta)$ takes only a finite number of values over the interval $[0, \Delta_0]$.

Theorem 4 ($q_{\mathcal{Y}}$–controlled invariant hypercubes) *Consider system* (2.1) *, assume* **A0** *and that* $\alpha = \|A\|_\infty > 1$ *. For* $\Delta > 0$ *, necessary conditions for the* $q_{\mathcal{Y}}$*–controlled invariance of* $Q_n(\Delta)$ *are:*

$$\begin{cases} m(\Delta) \leq -\dfrac{\Delta}{2}\big(\alpha - 1\big) & \text{(3.69a)} \\ M(\Delta) \geq \dfrac{\Delta}{2}\big(\alpha - 1\big) & \text{(3.69b)} \\ \rho(\Delta) \leq \Delta & \text{(3.69c)} \\ H^*(\Delta) \leq \Delta \, . & \text{(3.69d)} \end{cases}$$

If moreover $\rho(\Delta) + H^*(\Delta) \leq \Delta$ *, then* $Q_n(\Delta)$ *is* $q_{\mathcal{Y}}$*–controlled invariant.*

To prove Theorem 4 (and other results that will follow), we need two preliminary results. Next Proposition 11 is the counterpart of Proposition 1 in Section 3.1.1 :

Proposition 11 *Consider system* (2.1) *, assume* **A0** *and let* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *be a nearest neighbor quantizer. For* $\Delta > 0$ *,* $Q_n(\Delta)$ *is* $q_{\mathcal{Y}}$*–controlled invariant if and only if*

$$\forall\, y \in \mathcal{Y}(\Delta) \, , \quad \big|c^*_{\mathrm{mid}}(y) + q_{\mathcal{U}}\big(- c^*_{\mathrm{mid}}(y)\big)\big| + \frac{h^*(y)}{2} \leq \frac{\Delta}{2} \, .$$

Proof. By condition (3.65), $Q_n(\Delta)$ is $q_{\mathcal{Y}}$–controlled invariant if and only if

$$\begin{aligned}
&\forall\, y \in \mathcal{Y}(\Delta)\,,\ \exists\, u \in \mathcal{U} \quad \text{such that} \quad \mathrm{Pr}_n(A\mathcal{C}_y^*) + u \subseteq \big[\, -\tfrac{\Delta}{2}\,,\, \tfrac{\Delta}{2}\,\big] \quad \overset{\text{(a)}}{\Longleftrightarrow} \\
&\forall\, y \in \mathcal{Y}(\Delta)\,,\ \exists\, u \in \mathcal{U} \quad \text{such that} \quad \big|c^*_{\mathrm{mid}}(y) + u\big| + \tfrac{h^*(y)}{2} \leq \tfrac{\Delta}{2} \quad \overset{\text{(b)}}{\Longleftrightarrow} \\
&\forall\, y \in \mathcal{Y}(\Delta)\,, \quad \big|c^*_{\mathrm{mid}}(y) + q_{\mathcal{U}}\big(-c^*_{\mathrm{mid}}(y)\big)\big| + \tfrac{h^*(y)}{2} \leq \tfrac{\Delta}{2}\,,
\end{aligned}$$

where equivalence (a) is an easy consequence of the definitions of $c^*_{\mathrm{mid}}(y)$ and $h^*(y)$, and equivalence (b) holds because $\operatorname*{argmin}_{u\in\mathcal{U}} \big|c^*_{\mathrm{mid}}(y) + u\big| = q_{\mathcal{U}}\big(-c^*_{\mathrm{mid}}(y)\big)$. ■

Thus, nearest neighbor quantizers have a central role also for the $q_{\mathcal{Y}}$–invariance problem. We hence dwell on proving some basic properties of these quantizers. To this end, it is helpful to refer to the following notation: under the assumption that $\alpha \geq 1$, $\forall\, \Delta > 0$ such that $\rho(\Delta) < +\infty$, define the partition

$$\begin{aligned}
&\mathbb{R} = \mathcal{S}_{m(\Delta)} \cup \mathcal{N}_\Delta \cup \mathcal{S}_{M(\Delta)} \\
&\text{where} \\
&\begin{cases}
\mathcal{S}_{m(\Delta)} &:= \ \big]-\infty\,,\, m(\Delta) - \tfrac{\rho(\Delta)}{2}\big[\\
\mathcal{N}_\Delta &:= \ \big[m(\Delta) - \tfrac{\rho(\Delta)}{2}\,,\, M(\Delta) + \tfrac{\rho(\Delta)}{2}\big] \\
\mathcal{S}_{M(\Delta)} &:= \ \big]M(\Delta) + \tfrac{\rho(\Delta)}{2}\,,\, +\infty\big[\,.
\end{cases}
\end{aligned} \tag{3.70}$$

Let $\mathcal{S}_\Delta := \mathcal{S}_{m(\Delta)} \cup \mathcal{S}_{M(\Delta)}$. As it is better clarified in the following Lemma 8, the set $\mathcal{S}_\Delta$ represents the region where the nearest neighbor quantizers are saturating.

Lemma 8 (Properties of the nearest neighbor quantizers) *Consider system (2.1), assume* **A0** *and that $\alpha \geq 1$. Let $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ be a nearest neighbor quantizer. Consider $\Delta > 0$:*

ı) if inequalities (3.69a–b) hold, then $\forall\, z \in \mathrm{Pr}_n\big(AQ_n(\Delta)\big)$, $q_{\mathcal{U}}(z) \in \mathcal{U}(\Delta)$;

ıı) if $\rho(\Delta) < +\infty$ and $z \in \mathcal{N}_\Delta$, then $|q_{\mathcal{U}}(z) - z| \leq \frac{\rho(\Delta)}{2}$;

ıı) assume $\rho(\Delta) < +\infty$ and let z be such that $q_{\mathcal{U}}(z) \in \mathcal{U}(\Delta)$:

if $z \in \mathcal{S}_{M(\Delta)}$, then $q_{\mathcal{U}}(z) = M(\Delta)$ and $|q_{\mathcal{U}}(z) - z| = -\big(q_{\mathcal{U}}(z) - z\big) > \frac{\rho(\Delta)}{2}$;

if $z \in \mathcal{S}_{m(\Delta)}$, then $q_{\mathcal{U}}(z) = m(\Delta)$ and $|q_{\mathcal{U}}(z) - z| = q_{\mathcal{U}}(z) - z > \frac{\rho(\Delta)}{2}$.

Proof. See in Appendix A.2.4. ■

Proof of Theorem 4. First notice that, if $Q_n(\Delta)$ is $q_{\mathcal{Y}}$–controlled invariant for system (2.1), then a fortiori it is controlled invariant for the system $\Sigma(A, B, \mathcal{U})$ where the output map $q_{\mathcal{Y}}$ is replaced with the identity map. Therefore, the necessity of conditions (3.69a–b–c) holds by Theorem 1 in Section 3.1.1.

If $Q_n(\Delta)$ is $q_{\mathcal{Y}}$–controlled invariant, by Proposition 11 it holds that $\forall\, y \in \mathcal{Y}(\Delta)$, $h^*(y) \leq \Delta$: the necessity of (3.69d) hence follows.

Finally, let us show that the validity of inequalities (3.69a–b) together with $\rho(\Delta) + H^*(\Delta) \leq$

Δ is a sufficient condition for the $q_{\mathcal{Y}}$–controlled invariance of $Q_n(\Delta)$. We prove it by applying Proposition 11: first, since $\mathrm{Pr}_n\big(AQ_n(\Delta)\big) = \big[-\frac{\Delta}{2}\alpha\,,\,\frac{\Delta}{2}\alpha\big]$ (see equation (3.5)) and $\bigcup_{y\in\mathcal{Y}(\Delta)}\mathcal{C}_y^* = Q_n(\Delta)$, then $\forall\, y\in\mathcal{Y}(\Delta)$,

$$\begin{cases} c^*_{\mathrm{sup}}(y) \le \frac{\Delta}{2}\alpha \\ c^*_{\mathrm{inf}}(y) \ge -\frac{\Delta}{2}\alpha\,. \end{cases} \tag{3.71}$$

Consider a nearest neighbor quantizer $q_{\mathcal{U}}$. Let $y\in\mathcal{Y}(\Delta)$, if $-c^*_{\mathrm{mid}}(y)\in\mathcal{N}_\Delta$, then by Lemma 8.*iii*,

$$\big|c^*_{\mathrm{mid}}(y) + q_{\mathcal{U}}\big(-c^*_{\mathrm{mid}}(y)\big)\big| + \frac{h^*(y)}{2} \le \frac{\rho(\Delta)}{2} + \frac{h^*(y)}{2} \le \frac{\rho(\Delta)+H^*(\Delta)}{2} \le \frac{\Delta}{2}\,.$$

If instead $-c^*_{\mathrm{mid}}(y)\in\mathcal{S}_{M(\Delta)}$, then

$$\begin{aligned} \big|c^*_{\mathrm{mid}}(y) + q_{\mathcal{U}}\big(-c^*_{\mathrm{mid}}(y)\big)\big| + \tfrac{h^*(y)}{2} &\overset{(a)}{=} \big|c^*_{\mathrm{mid}}(y) + M(\Delta)\big| + \tfrac{h^*(y)}{2} = \\ &\overset{(b)}{=} -\big(c^*_{\mathrm{mid}}(y) + M(\Delta)\big) + \tfrac{h^*(y)}{2} = \\ &= -c^*_{\mathrm{inf}}(y) - M(\Delta) \le \\ &\overset{(c)}{\le} \tfrac{\Delta}{2}\alpha - \tfrac{\Delta}{2}(\alpha-1) = \tfrac{\Delta}{2}\,, \end{aligned}$$

where equalities (a) and (b) follows by Lemma 8.*iii* (which can be applied because, since $-c^*_{\mathrm{mid}}(y)\in\mathrm{Pr}_n\big(AQ_n(\Delta)\big)$, then by Lemma 8.*i*, $q_{\mathcal{U}}\big(-c^*_{\mathrm{mid}}(y)\big)\in\mathcal{U}(\Delta)\big)$, and inequality (c) follows by inequalities (3.71) and (3.69b).
The case $-c^*_{\mathrm{mid}}(y)\in\mathcal{S}_{m(\Delta)}$ is similar. ■

Since $H(\Delta)\ge H^*(\Delta)$, then an easily computable invariance condition is provided by the following corollary:

Corollary 4 *Under the same assumptions of Theorem 4, a sufficient condition for the $q_{\mathcal{Y}}$–controlled invariance of $Q_n(\Delta)$ is that inequalities (3.69a–b) hold together with $\rho(\Delta)+H(\Delta)\le\Delta$.* □

Remark 13 *The boundedness assumption of $\mathcal{C}_y$ has never been explicitly invoked and it is relevant only when dealing with the function $H(\Delta)$. Indeed, the results presented in this section hold true also in the quantized output case. Nevertheless, in this more general case, the hypotheses of Theorem 4 are typically not satisfied: in particular, condition* (3.69d) *is met only in special cases. As far as Corollary 4 is concerned, in the quantized output case $q_{\mathcal{Y}} = q_o\circ C$, $H(\Delta)\equiv+\infty$ unless*[5] $\mathrm{Ker}\,C = \mathrm{Ker}\,(a_1\ \cdots\ a_n)$.

[5] Where $\mathrm{Ker}\,C := \{x\in\mathbb{R}^n \,|\, Cx = 0\}$.

Example 12 (Finite control set and quantized state) *Let us consider a quantized state version of the system in Example 6 of Section 3.1.1, that is*

$$\begin{cases} x^+ = \begin{pmatrix} 0 & 1 \\ 5/4 & 1/4 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ y = q_{\mathcal{Y}}(x) = \begin{pmatrix} q_{\mathrm{cw}}(x_1) \\ q_{\mathrm{cw}}(x_2) \end{pmatrix}, \end{cases}$$

where, as in Example 6, $u \in \mathcal{U} = \{0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 8, \pm 12, \pm 16, \pm 24\}$, *and the quantizer* $q_{\mathrm{cw}} : \mathbb{R} \to \mathcal{Y}_1 \subset \mathbb{R}$ *is defined as follows:*

$$q_{\mathrm{cw}}(z) := \begin{cases} \lceil z \rceil - \frac{1}{2} & \text{if} \quad 0 \le z \le 4 \\ 2^{h-1} \cdot q_{\mathrm{cw}}\left(\frac{z}{2^{h-1}}\right) & \text{if} \quad 2^h < z \le 2^{h+1}, \quad \text{for} \quad h = 2, 3, 4 \\ 40 & \text{if} \quad z > 32 \\ -q_{\mathrm{cw}}(-z) & \text{if} \quad z < 0\,. \end{cases}$$

Let us determine the function $H(\Delta)$ *and the values of* Δ *such that the sufficient condition for the* $q_{\mathcal{Y}}$*–controlled invariance of* $Q_2(\Delta)$ *given in Corollary 4 is satisfied.*
To this end, we first briefly describe the state space partition induced by $q_{\mathcal{Y}}$ *(see Fig. 3.13). It is a componentwise quantization, we hence focus on the partition induced by* q_{cw}. *The interval* $[0, 4]$ *is divided into* 4 *intervals of equal length 1; each interval* $[4, 8]$, $[8, 16]$ *and* $[16, 32]$ *is divided into* 2 *intervals of equal length (2, 4 and 8, respectively). For* $z > 32$, *the quantizer is saturating. The partition induced by* q_{cw} *is a saturated version of a so called* floating–point *quantization, a type of quantization that will be studied in Chapter 7 and that extends the logarithmic quantization.*
Since $\mathrm{diam}_2\big(AQ_2(\lambda)\big) = \|A\|_\infty \cdot \lambda = \frac{3}{2}\lambda$, *by the properties of the quantizer* q_{cw} *it immediately follows that*

$$H(\Delta) = \begin{cases} 3/2 & \text{if} \quad 0 < \Delta \le 8 \\ 3 & \text{if} \quad 8 < \Delta \le 16 \\ 6 & \text{if} \quad 16 < \Delta \le 32 \\ 12 & \text{if} \quad 32 < \Delta \le 64 \\ +\infty & \text{if} \quad \Delta > 64\,. \end{cases}$$

Finally, let us determine the values of Δ *satisfying inequalities* (3.69a–b) *and* $\rho(\Delta) + H(\Delta) \le \Delta$. *The functions* $M(\Delta)$ *and* $m(\Delta)$ *depend only on the control set* $\mathcal{U}$ *and have been determined in Example 6. The function* $\rho(\Delta) + H(\Delta)$ *is obtained by combining* $H(\Delta)$ *with* $\rho(\Delta)$ *determined in equation* (3.14) *of Example 6 (see the plot reported in Fig. 3.14). Therefore, according to Corollary 4, a sufficient condition for the* $q_{\mathcal{Y}}$*–controlled invariance of* $Q_2(\Delta)$ *is* $\frac{5}{2} \le \Delta \le 64$. ♣

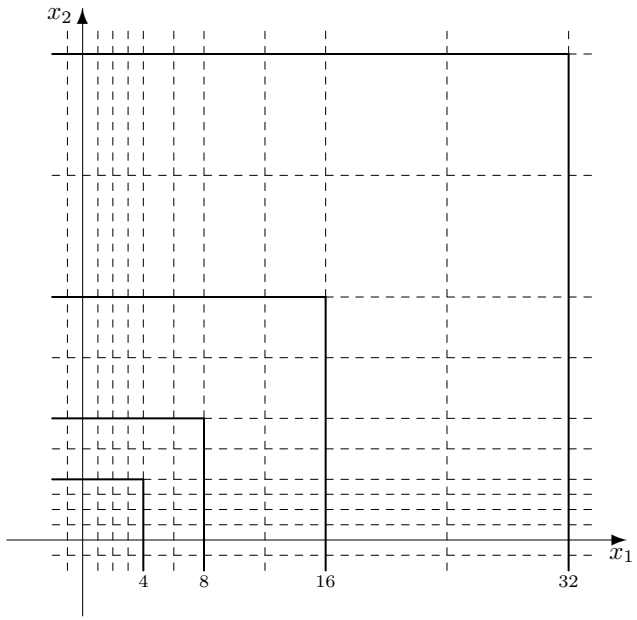

Figure 3.13: State space partition induced by the state quantizer q_y of Example 12 .

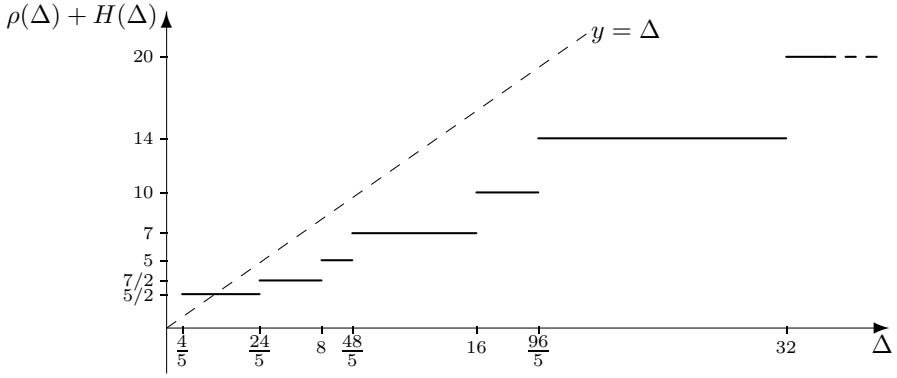

Figure 3.14: Plot of $\rho(\Delta) + H(\Delta)$ for the system in Example 12 .

3.2.2 Controlled invariant hypercubes: output quantization

The techniques we have introduced in the previous sections can be extended to deal with quantized outputs (see Section 2.2). We consider single–output systems: this is the case where $q_{\mathcal{Y}} = q_o \circ C$, with $C \in \mathbb{R}^{1\times n}$ and (A, C) is an observable pair. It is assumed that the map $q_o : \mathbb{R} \to \mathcal{Y}$ induces a locally finite partition of $\mathbb{R}$ and that $\forall\, y \in \mathcal{Y}$, $q_o^{-1}(y) \subseteq \mathbb{R}$ is a connected set $\big($thus, $q_o^{-1}(y) \subseteq \mathbb{R}$ is either an interval of finite length or a half–line, the latter case corresponding to the saturation of the output quantizer$\big)$.
As we noticed in Remark 13 of the previous section, the property of $q_{\mathcal{Y}}$–controlled invariance is too strong for a quantized output system. In this section, we hence consider a particular kind of invariance which consists of guaranteeing that, at time $t+1$, the state of the system can be confined within a set Ω assuming that the state was in Ω, not only at time t, but also for a sufficiently long past time–horizon; the only available information to select the control value is the corresponding sequence of past (quantized) inputs and outputs. This kind of invariance will be referred to as *dynamic $q_{\mathcal{Y}}$–controlled invariance* and will turn out to be useful in next Chapter 4 when considering the control synthesis for practical stabilization in the presence of output quantization.
First, a state–observer is constructed: in this framework, because of the output quantization, the observer is a machine fed by discrete values $y \in \mathcal{Y}$, and returning an estimate $\hat{x}(t)$ of the current state. More precisely, the machine also returns a set $\mathcal{C}_{\hat{x}(t)} \subset \mathbb{R}^n$ within which the current state $x(t)$ is guaranteed to lie. In accordance with the terminology of Section 2.2, the ensemble of the sets $\mathcal{C}_{\hat{x}(t)}$ that the observer can return at a given time t may be viewed as a state space quantization: that is, in a sufficiently large neighborhood of the origin, the sets $\mathcal{C}_{\hat{x}(t)}$'s are bounded. Therefore, in a second stage, the analysis of the dynamic $q_{\mathcal{Y}}$–controlled invariance of hypercubes can be performed by taking advantage of the theory developed for the quantized state case. However, those results need to be further elaborated because indeed the ensemble of the $\mathcal{C}_{\hat{x}(t)}$'s are a *time–varying* state space quantization. Let us go into the details.

Construction of the quantized state–observer

By suitably redefining q_o without modifying the induced output space partition (hence, without loss of generality), we can assume that $\mathcal{Y} \subset \mathbb{R}$ and, if $y \in \mathcal{Y}$ is such that the closure of $q_o^{-1}(y)$ is an interval of finite length λ_y, that y is the middle point of such an interval. Thus, with

$$\mathcal{Y}_\star := \{y \in \mathcal{Y} \,|\, q_o^{-1}(y) \text{ is an interval of finite length}\}\,, \tag{3.72}$$

it holds that

$$\forall\, y \in \mathcal{Y}_\star\,, \quad \overline{q_o^{-1}(y)} = [\, y - \tfrac{\lambda_y}{2}\,,\, y + \tfrac{\lambda_y}{2}\,]\,. \tag{3.73}$$

The function $\vec{q}_o : \mathbb{R}^n \to \mathcal{Y}^n$ defined by $\vec{q}_o(z) := \big(q_o(z_1), \dots, q_o(z_n)\big)$ induces a partition of $\mathbb{R}^n$ such that $\forall\, \vec{y} \in \mathcal{Y}_\star^n$, the closure of $\vec{q}_o^{-1}(\vec{y})$ is $\vec{y} + \mathcal{P}_{\vec{y}}$, where

$$\mathcal{P}_{\vec{y}} = \left[-\frac{\lambda_{\vec{y}_1}}{2}\,,\, \frac{\lambda_{\vec{y}_1}}{2} \right] \times \cdots \times \left[-\frac{\lambda_{\vec{y}_n}}{2}\,,\, \frac{\lambda_{\vec{y}_n}}{2} \right].$$

Let

$$S := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & \cdots & 0 \\ CAB & CB & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{n-2}B & CA^{n-3}B & \cdots & CB \end{pmatrix} \in \mathbb{R}^{n \times (n-1)}.$$

For $t \geq n-1$, denote by $\vec{u}\,(t)$ and $\vec{y}\,(t)$ the vectors collecting respectively the last $n-1$ inputs and the last n outputs of the system at time t, that is

$$\begin{cases} \vec{u}\,(t) := \big(u(t-n+1), \dots, u(t-1)\big)' \\ \vec{y}\,(t) := \big(y(t-n+1), \dots, y(t)\big)'. \end{cases}$$

Let $R := \big[A^{n-2}B \,|\, \cdots \,|AB\,|B\big] \in \mathbb{R}^{n \times (n-1)}$ and $\mathcal{O} \in \mathbb{R}^{n \times n}$ be the observability matrix (i.e., the matrix whose i–th row is CA^{i-1}): $\mathcal{O}$ is invertible because the pair (A, C) is observable. By standard theory on observability (see, for instance, [114]) it holds that

$$\vec{y}\,(t) = \vec{q}_o\left(\mathcal{O}\, x(t-n+1) + S\vec{u}\,(t)\right),$$

hence

$$x(t-n+1) \in \mathcal{O}^{-1}\Big(\vec{q}_o^{-1}\big(\vec{y}\,(t)\big) - S\vec{u}\,(t)\Big)$$

and

$$x(t) \in A^{n-1}\mathcal{O}^{-1}\Big(q_o^{-1}\big(\vec{y}\,(t)\big)\Big) - A^{n-1}\mathcal{O}^{-1}S\vec{u}\,(t) + R\vec{u}\,(t).$$

Accordingly, consider the map $\psi : \mathcal{Y}^n \times \mathcal{U}^{n-1} \to \mathbb{R}^n$ defined by

$$\psi(\,\vec{y}, \vec{u}\,) := A^{n-1}\mathcal{O}^{-1}\vec{y} + (R - A^{n-1}\mathcal{O}^{-1}S)\vec{u}\,.$$

For $t \geq n-1$, the *quantized state–observer* is defined by the following equations:

$$\begin{cases} \hat{x}(t) := \psi\big(\,\vec{y}\,(t)\,, \vec{u}\,(t)\big) \\ x(t) \in A^{n-1}\mathcal{O}^{-1}\Big(q_o^{-1}\big(\vec{y}\,(t)\big)\Big) + (R - A^{n-1}\mathcal{O}^{-1}S)\vec{u}(t)\,. \end{cases} \tag{3.74}$$

If $\vec{y}\,(t) \in \mathcal{Y}_\star^n$, since the closure of $\vec{q}_o^{-1}(\vec{y})$ is $\vec{y} + \mathcal{P}_{\vec{y}}$, then

$$x(t) \in \mathcal{C}_{\hat{x}(t)} := A^{n-1}\mathcal{O}^{-1}\big(\mathcal{P}_{\vec{y}(t)}\big) + A^{n-1}\mathcal{O}^{-1}\vec{y}\,(t) + \left(R - A^{n-1}\mathcal{O}^{-1}S\right)\vec{u}\,(t)\ : \tag{3.75}$$

$\mathcal{C}_{\hat{x}(t)}$ is a bounded parallelogram and $\hat{x}(t) = \psi\big(\,\vec{y}\,(t)\,, \vec{u}\,(t)\big)$ is the centroid of $\mathcal{C}_{\hat{x}(t)}$.

Analysis: invariance of hypercubes

As in equation (3.64), for $\Delta > 0$, let

$$\mathcal{Y}(\Delta) := \big(q_o \circ C\big)\big(Q_n(\Delta)\big)\,.$$

If $\Delta > 0$ is such that $\mathcal{Y}(\Delta) \subseteq \mathcal{Y}_\star$, let

$$\Lambda_\Delta := \max_{y \in \mathcal{Y}(\Delta)} \lambda_y\,. \tag{3.76}$$

We define

$$\widetilde{H}(\Delta) := \begin{cases} \operatorname{diam}_n\Big(A^n \mathcal{O}^{-1}\big(Q_n(\Lambda_\Delta)\big)\Big) & \text{if} \quad \mathcal{Y}(\Delta) \subseteq \mathcal{Y}_\star \\ +\infty & \text{otherwise}\,. \end{cases} \tag{3.77}$$

Similarly to the function $H(\Delta)$ defined in equation (3.68) of Section 3.2.1, the function $\widetilde{H}(\Delta)$ is non–decreasing with Δ and, $\forall\, \Delta_0 > 0$, $\widetilde{H}(\Delta)$ takes only a finite number of values over the interval $[0, \Delta_0]$ (the latter property is a consequence of the fact that the partition induced by $q_o : \mathbb{R} \to \mathcal{Y}$ is assumed to be locally finite).
The meaning of the quantity $\widetilde{H}(\Delta)$ and its analogy with $H(\Delta)$ become apparent in the proof of the following result:

Theorem 5 (Dynamic $q_\mathcal{Y}$–controlled invariant hypercubes) *Consider system (2.1), assume* **A0** *and that $\alpha = \|A\|_\infty \geq 1$. Suppose that the system is quantized single–output (see Section 2.2) with $q_\mathcal{Y} = q_o \circ C$, where $C \in \mathbb{R}^{1\times n}$ and (A, C) is an observable pair. Let $\Delta > 0$ be such that*

$$\begin{cases} m(\Delta) \leq -\dfrac{\Delta}{2}\big(\alpha - 1\big) & \text{(3.78a)} \\ M(\Delta) \geq \dfrac{\Delta}{2}\big(\alpha - 1\big) & \text{(3.78b)} \\ \rho(\Delta) + \widetilde{H}(\Delta) \leq \Delta & \text{(3.78c)} \end{cases}$$

and $\mathcal{U} = \mathcal{U}(\Delta)$. For some $t \geq n-1$, assume that $\forall\, \tau = t-n+1, \ldots, t$, $x(\tau) \in Q_n(\Delta)$. Let $\hat{x}(t)$ be the issue of the quantized state–observer defined in equation (3.74) and $q_\mathcal{U} : \mathbb{R} \to \mathcal{U}$ be a nearest neighbor quantizer. Then, $u = q_\mathcal{U}\Big(-\big(A\hat{x}(t)\big)_n\Big)$ is such that

$$x(t+1) = Ax(t) + Bu \in Q_n(\Delta)\,.$$

Proof. The proof is given below after a preliminary result. ■

Remark 14 *The assumption $\mathcal{U} = \mathcal{U}(\Delta)$ is needed to ensure that $u = q_\mathcal{U}\Big(-\big(A\hat{x}(t)\big)_n\Big) \in \mathcal{U}(\Delta)$. Indeed, by Lemma 6 in Section 3.1.1, this is a necessary condition in order that $x(t+1) \in Q_n(\Delta)$. Notice, however, that this is a mild assumption that can be always satisfied simply by saturating the controller (i.e., by neglecting the control values out of $\mathcal{U}(\Delta)$).*

Remark 15 *Differently from the analogous invariance theorems presented in the previous sections (i.e., Theorem 1 in Section 3.1.1 and Theorem 4 in Section 3.2.1), in Theorem 5 we assumed* $\alpha \geq 1$. *While it is still true that the invariance analysis for* $\alpha = 1$ *is trivial (in this case, in fact,* $u = 0$ *ensures the invariance of any hypercube), on the other hand, since Theorem 5 consists of an invariance analysis under a given control law, then also the case* $\alpha = 1$ *is significant.*

The main tool to prove Theorem 5, as well as other results in next Chapter 4, is the following:

Lemma 9 (Main tool) *Consider system* (2.1), *assume* **A0** *and that* $\alpha = \|A\|_\infty \geq 1$. *Let* $\Delta > 0$ *be such that* $\rho(\Delta) < +\infty$ *and inequalities (3.78a–b) hold. Consider a qdb–controller* $k : \mathbb{R}^n \to \mathcal{U}$. *Given* $x \in Q_n(\Delta)$ *and* $\hat{x} \in \mathbb{R}^n$, *let* $\mathscr{H} \geq 0$ *be such that* $\left|\big(A(x - \hat{x})\big)_n\right| \leq \frac{\mathscr{H}}{2}$. *If* $k(\hat{x}) \in \mathcal{U}(\Delta)$, *then* $x^+ = Ax + Bk(\hat{x})$ *is such that*

$$|x_n^+| \leq \max\left\{\frac{\rho(\Delta) + \mathscr{H}}{2}\,,\, \|x\|_\infty - \varphi(\Delta)\right\}, \tag{3.79}$$

where

$$\varphi(\Delta) := \min\left\{\, M(\Delta) - \tfrac{\Delta}{2}(\alpha - 1)\,,\, -\tfrac{\Delta}{2}(\alpha - 1) - m(\Delta)\,\right\}. \tag{3.80}$$

Proof. By definition of k, $x_n^+ = (Ax)_n + q_{\mathcal{U}}\big(-(A\hat{x})_n\big)$, where $q_{\mathcal{U}}$ is a nearest neighbor quantizer. Notice that, by Lemma 8.$\imath$ in Section 3.2.1,

$$q_{\mathcal{U}}\big(-(Ax)_n\big) \in \mathcal{U}(\Delta). \tag{3.81}$$

With reference to the partition $\mathbb{R} = \mathcal{S}_{m(\Delta)} \cup \mathcal{N}_\Delta \cup \mathcal{S}_{M(\Delta)}$ defined in equation (3.70) of Section 3.2.1, three cases can occur:

I) Suppose that $-(A\hat{x})_n \in \mathcal{N}_\Delta$, then

$$\begin{aligned} |x_n^+| = \;& \left|\big(A(x - \hat{x})\big)_n + (A\hat{x})_n + q_{\mathcal{U}}\big(-(A\hat{x})_n\big)\right| \quad \leq \\ \leq \;& \left|\big(A(x - \hat{x})\big)_n\right| + \left|(A\hat{x})_n + q_{\mathcal{U}}\big(-(A\hat{x})_n\big)\right| \quad \leq \\ \leq \;& \tfrac{\mathscr{H}}{2} + \tfrac{\rho(\Delta)}{2}\,, \end{aligned}$$

where the last inequality follows by the hypothesis on $\mathscr{H}$ and by Lemma 8.$\imath\imath$.

II) Suppose that $-(A\hat{x})_n \in \mathcal{S}_\Delta$ and x is such that $-(Ax)_n \in \mathcal{N}_\Delta$. If $-(A\hat{x})_n \in \mathcal{S}_{m(\Delta)}$, then $k(\hat{x}) = m(\Delta)$ thanks to Lemma 8.$\imath\imath\imath$ which can be applied because, by assumption, $k(\hat{x}) \in \mathcal{U}(\Delta)$. Hence,

$$x_n^+ = (Ax)_n + m(\Delta) \overset{\text{(a)}}{\leq} (Ax)_n + q_{\mathcal{U}}\big(-(Ax)_n\big) \overset{\text{(b)}}{\leq} \tfrac{\rho(\Delta)}{2}\,,$$

where inequality (a) holds because $m(\Delta) = \min \mathcal{U}(\Delta) \ni q_{\mathcal{U}}\big(-(Ax)_n\big)$ $\big($see equation (3.81)$\big)$, and inequality (b) follows by Lemma 8.$\imath\imath$. Moreover, by Lemma 8.$\imath\imath\imath$, $(A\hat{x})_n + q_{\mathcal{U}}\big(-(A\hat{x})_n\big) > \frac{\rho(\Delta)}{2}$, and, by assumption, $\big(A(x - \hat{x})\big)_n \geq -\frac{\mathscr{H}}{2}$, therefore

$$x_n^+ = \big(A(x - \hat{x})\big)_n + (A\hat{x})_n + q_{\mathcal{U}}\big(-(A\hat{x})_n\big) > -\tfrac{\mathscr{H}}{2} + \tfrac{\rho(\Delta)}{2} > -\tfrac{\mathscr{H} + \rho(\Delta)}{2}\,.$$

To sum up, $|x_n^+| \leq \frac{\rho(\Delta)+\mathscr{H}}{2}$.
The case $-(A\hat{x})_n \in \mathcal{S}_{M(\Delta)}$ is similar.
III) Suppose that $-(A\hat{x})_n \in \mathcal{S}_\Delta$ and $-(Ax)_n \in \mathcal{S}_\Delta$. If $-(A\hat{x})_n \in \mathcal{S}_{m(\Delta)}$, we know by part II that $k(\hat{x}) = m(\Delta)$ and $x_n^+ > -\frac{\mathscr{H}+\rho(\Delta)}{2}$. Assume that $-(Ax)_n \in \mathcal{S}_{M(\Delta)}$, since $q_{\mathcal{U}}\big(-(Ax)_n\big) \in \mathcal{U}(\Delta)$, then

$$x_n^+ = (Ax)_n + m(\Delta) < (Ax)_n + M(\Delta) \overset{\text{(c)}}{=} (Ax)_n + q_{\mathcal{U}}\big(-(Ax)_n\big) \overset{\text{(d)}}{<} -\tfrac{\rho(\Delta)}{2},$$

where both equality (c) and inequality (d) hold by Lemma 8.*iii*. Hence, $|x_n^+| < \frac{\mathscr{H}+\rho(\Delta)}{2}$.
If instead $-(Ax)_n \in \mathcal{S}_{m(\Delta)}$, then $|x_n^+| \leq \|x\|_\infty - \varphi(\Delta)$. In fact: in this case $k(x) = k(\hat{x}) = m(\Delta)$ and, thanks to inequalities (3.78a–b), we can write $m(\Delta) = -\frac{\Delta}{2}(\alpha-1) - \varphi(\Delta) - \theta$, with $\theta \geq 0$. Again by Lemma 8.*iii*, $x_n^+ = (Ax)_n + m(\Delta) > \frac{\rho(\Delta)}{2} > 0$, hence

$$\begin{aligned} |x_n^+| = \;& (Ax)_n + m(\Delta) \leq \textstyle\sum_{i=1}^n |a_i|\,|x_i| + m(\Delta) \leq \alpha \cdot \|x\|_\infty + m(\Delta) = \\ = \;& \alpha \cdot \|x\|_\infty - \tfrac{\Delta}{2}(\alpha-1) - \varphi(\Delta) - \theta \leq \|x\|_\infty - \varphi(\Delta), \end{aligned}$$

where the last inequality holds because $\big(\|x\|_\infty - \frac{\Delta}{2}\big)(\alpha-1) - \theta \leq 0$.
The case $-(A\hat{x})_n \in \mathcal{S}_{M(\Delta)}$ is similar. ■

Proof of Theorem 5. By the controller form of the system, it is sufficient to show that $x^+(t)$ is such that $|x_n^+| \leq \frac{\Delta}{2}$. Since $\rho(\Delta) + \widetilde{H}(\Delta) \leq \Delta$ and, by inequalities (3.78a–b), $\varphi(\Delta) \geq 0$, then $|x_n^+| \leq \frac{\Delta}{2}$ thanks to inequality (3.79) of Lemma 9 with $\mathscr{H} = \widetilde{H}(\Delta)$. Hence, it is sufficient to show that the hypotheses of Lemma 9 are satisfied: since $\mathcal{U} = \mathcal{U}(\Delta)$, then $q_{\mathcal{U}}\Big(-\big(A\hat{x}(t)\big)_n\Big) = k\big(\hat{x}(t)\big) \in \mathcal{U}(\Delta)$; it remains only to show that

$$\Big|\Big(A\big(x(t) - \hat{x}(t)\big)\Big)_n\Big| \leq \frac{\widetilde{H}(\Delta)}{2}. \tag{3.82}$$

First notice that, by inequality (3.78c), it holds that $\widetilde{H}(\Delta) < +\infty$ and therefore Δ is such that $\mathcal{Y}(\Delta) \subseteq \mathcal{Y}_\star$. Moreover, since $\forall\, \tau = t-n+1, \ldots, t$, $x(\tau) \in Q_n(\Delta)$, then $\vec{y}(t) \in \mathcal{Y}_\star^n$. Hence, $\hat{x}(t)$ is the centroid of the parallelogram $\mathcal{C}_{\hat{x}(t)}$ defined in equation (3.75) and containing $x(t)$. Therefore, to prove that inequality (3.82) holds, it is sufficient to show that $\operatorname{diam}_n(A\mathcal{C}_{\hat{x}(t)}) \leq \widetilde{H}(\Delta)$. Indeed, according to equation (3.75), $\mathcal{C}_{\hat{x}(t)}$ is a translation of the set $A^{n-1}\mathcal{O}^{-1}\big(\mathcal{P}_{\vec{y}(t)}\big)$ and $\mathcal{P}_{\vec{y}(t)} \subseteq Q_n(\Lambda_\Delta)$, then

$$\operatorname{diam}_n(A\mathcal{C}_{\hat{x}(t)}) = \operatorname{diam}_n\big(A^n\mathcal{O}^{-1}(\mathcal{P}_{\vec{y}(t)})\big) \leq \operatorname{diam}_n\Big(A^n\mathcal{O}^{-1}\big(Q_n(\Lambda_\Delta)\big)\Big) = \widetilde{H}(\Delta).$$

■

Example 13 (Finite control set and quantized output) *Let us consider a quantized output version of the system in Example 6 of Section 3.1.1, that is*

$$\begin{cases} x^+ = \begin{pmatrix} 0 & 1 \\ 5/4 & 1/4 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ y = q_{\mathcal{Y}}(x) = q_o(Cx), \end{cases}$$

where, as in Example 6, $u \in \mathcal{U} = \{0\,,\pm1\,,\pm2\,,\pm3\,,\pm4\,,\pm6\,,\pm8\,,\pm12\,,\pm16\,,\pm24\}$. *As far as the output map* q_y *is concerned, suppose that* $C = (\ 3/2 \quad 1/3\ )$ *and that the extremes of the intervals forming the output space partition induced by* q_o *are* $\{\pm\frac{3}{2}\,,\pm\frac{9}{2}\,,\pm\frac{15}{2}\,,\pm\frac{25}{2}\,,\pm\frac{39}{2}\}$. *Let us determine the equation of the quantized state–observer, the function* $\widetilde{H}(\Delta)$ *and the values of* Δ *satisfying inequalities* (3.78).

According to equations (3.72) *and* (3.73), *let* $\mathcal{Y} = \mathcal{Y}_\star \cup \{\pm y_{\rm s}\} = \{0\,,\pm3\,,\pm6\,,\pm10\,,\pm16\,,\pm y_{\rm s}\}$ $\big($*where* $\mathcal{Y}_\star$ *collects the middle points of the output quantization intervals and, for* $|Cx| > \frac{39}{2}$, q_o *takes the saturation values* $\pm y_{\rm s} \in \mathbb{R}$*: e.g.,* $y_{\rm s} = 20\big)$. *With this definition of* $\mathcal{Y}$, *the equation of the quantized state–observer is*

$$\hat{x}(t) = \psi\big(\,\vec{y}\,(t)\,,\vec{u}\,(t)\big) = A\mathcal{O}^{-1}\vec{y}\,(t) + (R - A\mathcal{O}^{-1}S)\vec{u}\,(t)\,,$$

where, since

$$\mathcal{O}^{-1} = \frac{1}{161}\begin{pmatrix} 114 & -24 \\ -30 & 108 \end{pmatrix}, \quad R = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \textit{and} \quad S = \begin{pmatrix} 0 \\ 1/3 \end{pmatrix},$$

then

$$\hat{x}(t) = \frac{1}{161}\left[\begin{pmatrix} -30 & 108 \\ 135 & -3 \end{pmatrix}\begin{pmatrix} y(t-1) \\ y(t) \end{pmatrix} + \begin{pmatrix} -36 \\ 162 \end{pmatrix} u(t-1)\right].$$

In order to determine the function $\widetilde{H}(\Delta)$, *we have to compute* Λ_Δ $\big($*see equation* (3.76)$\big)$. *First,* $C\big(Q_2(\Delta)\big) = \big[-\|C\|_\infty\tfrac{\Delta}{2}\,,\|C\|_\infty\tfrac{\Delta}{2}\big] = \big[-\frac{11}{12}\Delta\,,\frac{11}{12}\Delta\big]$. *Hence,*

$$\mathcal{Y}(\Delta) = \begin{cases} \{0\} & \textit{if } 0 < \Delta < \frac{18}{11} \quad \big(\textit{as it results by } \frac{11}{12}\Delta < \frac{3}{2}\big) \\ \{0\,,\pm3\} & \textit{if } \frac{18}{11} < \Delta < \frac{54}{11} \quad \big(\textit{as it results by } \frac{11}{12}\Delta < \frac{9}{2}\big) \\ \textit{and so on}\,. \end{cases}$$

For $\Delta = 18/11$, $\Delta = 54/11$, *and so on, the set* $\mathcal{Y}(\Delta)$ *depends on the value taken by the output quantizer at the extremes of the intervals forming the output space partition: let us assume that the intervals where* $\mathcal{Y}(\Delta)$ *is constant are those reported above but closed on the right (in this example, the final result does not change if other cases are considered).*

For $y \in \mathcal{Y}_\star$, *the values of* λ_y *are:*

$$\lambda_0 = \lambda_{\pm3} = \lambda_{\pm6} = 3\,, \quad \lambda_{\pm10} = 5 \quad \textit{and} \quad \lambda_{\pm16} = 7\,.$$

Therefore, with the above assumption on $\mathcal{Y}(\Delta)$,

$$\Lambda_\Delta = \begin{cases} 3 & \textit{if } \Delta \le \frac{90}{11} \quad \big(\textit{as it results by } \frac{11}{12}\Delta \le \frac{15}{2}\big) \\ 5 & \textit{if } \frac{90}{11} < \Delta \le \frac{150}{11} \quad \big(\textit{as it results by } \frac{11}{12}\Delta \le \frac{25}{2}\big) \\ 7 & \textit{if } \frac{150}{11} < \Delta \le \frac{234}{11} \quad \big(\textit{as it results by } \frac{11}{12}\Delta \le \frac{39}{2}\big). \end{cases}$$

For Δ *such that* $\mathcal{Y}(\Delta) \subseteq \mathcal{Y}_\star$ $\big($*i.e., for* $\Delta \le \frac{234}{11}\big)$, *we have*

$$\widetilde{H}(\Delta) = \operatorname{diam}_2\Big(A^2\mathcal{O}^{-1}\big(Q_2(\Lambda_\Delta)\big)\Big).$$

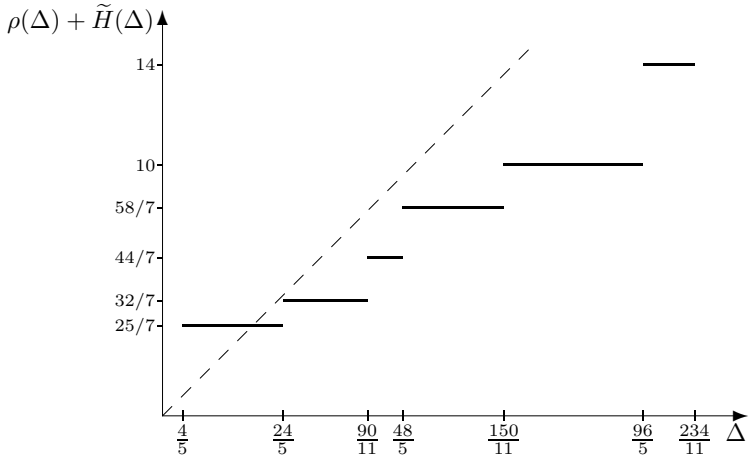

Figure 3.15: Plot of $\rho(\Delta)+\widetilde{H}(\Delta)$ for the system in Example 13 .

Since

$$A^2\mathcal{O}^{-1}=\frac{1}{4\cdot 161}\begin{pmatrix}540 & -12\\ -15 & 537\end{pmatrix},$$

then $\operatorname{diam}_2\Big(A^2\mathcal{O}^{-1}\big(Q_2(\Lambda_\Delta)\big)\Big)=\frac{15+537}{4\cdot 161}\Lambda_\Delta=\frac{6}{7}\Lambda_\Delta$ *so that*

$$\widetilde{H}(\Delta)=\begin{cases}\frac{18}{7} & \text{if}\quad 0<\Delta\le\frac{90}{11}\\ \frac{30}{7} & \text{if}\quad \frac{90}{11}<\Delta\le\frac{150}{11}\\ 6 & \text{if}\quad \frac{150}{11}<\Delta\le\frac{234}{11}\\ +\infty & \text{if}\quad \Delta>\frac{234}{11}\,.\end{cases}$$

Finally, let us determine the values of Δ *satisfying inequalities* (3.78) . *As for* $M(\Delta)$ *and* $m(\Delta)$, *see Example 6 . The function* $\rho(\Delta)+\widetilde{H}(\Delta)$ *is obtained by combining* $\widetilde{H}(\Delta)$ *with* $\rho(\Delta)$ *determined in equation* (3.14) *of Example 6 (see the plot reported in Fig. 3.15) . Therefore,* Δ *satisfies inequalities* (3.78) *if and only if* $\frac{25}{7}\le\Delta\le\frac{234}{11}$. ♣

Chapter 4

The qdb–controller

This chapter and next Chapters 5 and 6 are devoted to the control synthesis for practical stabilization of system (2.1) .

For single–input reachable systems, we have introduced in previous Chapter 3 a nice and easy technique for the invariance analysis of hypercubic sets. In the case of quantized input systems, we have also proved optimality properties of invariant hypercubes: that is, the smallest final set Ω so that (X_0, Ω)–stabilization can be achieved is a hypercube. We have also seen that the controlled invariance of hypercubes is directly related to their positive invariance under a qdb–controller. In this chapter, we go further into the study of single–input reachable systems: we consider the practical stabilization by means of a qdb–controller and we analyze the (X_0, X_1, Ω)–stability properties of the corresponding closed loop dynamics in terms of hypercubic sets X_0 , X_1 and Ω . According to the terminology introduced in Section 2.2 , all the three cases of quantized input, quantized state and quantized outputs are considered: see next Sections 4.1 , 4.2.1 and 4.2.2 , respectively.

Throughout this chapter, it is assumed that system (2.1) is represented in the controller form coordinates $\big($see assumption **A0** in equation (3.1) of Section 3.1.1$\big)$. Recall that

$$\alpha := \sum_{i=1}^{n} |a_i| \, .$$

Differently from the invariance problem, which is meaningful only for $\alpha > 1$, the practical stabilization problem is significant also when $\alpha = 1$ because convergence properties (from X_0 to Ω) are to be ensured. The case $\alpha < 1$ is trivial because, by Lemma 28 in Appendix A.5.1 , the matrix A is Schur and $u(t) \equiv 0$ guarantees asymptotic stability.

4.1 Practical stabilization: quantized single–input

In this section, we consider system (2.1) in the quantized input case $\Sigma(A, B, \mathcal{U})$ and a qdb–controller $\big($see equation (3.2)$\big)$. The main result on the practical stability properties of the corresponding closed loop dynamics is the following:

Theorem 6 ((X_0, Ω)–stabilization: quantized single–input reachable systems) *Consider system* $\Sigma(A, B, \mathcal{U})$ *, assume* **A0** *and that* $\alpha = \|A\|_\infty \geq 1$ *. Let* $k : \mathbb{R}^n \to \mathcal{U}$ *be a qdb–controller. If* $\Delta_0 > 0$ *is such that*

$$\begin{cases} m(\Delta_0) < -\dfrac{\Delta_0}{2}\big(\alpha - 1\big) & \text{(4.1a)} \\ M(\Delta_0) > \dfrac{\Delta_0}{2}\big(\alpha - 1\big) & \text{(4.1b)} \\ \rho(\Delta_0) < \Delta_0\,, & \text{(4.1c)} \end{cases}$$

then it is well defined

$$\Delta_{\inf} := \max\big\{\Delta < \Delta_0 \,|\, \rho(\Delta) = \Delta\big\} \tag{4.2}$$

and the control law $u(x) = k(x)$ *is* $\big(Q_n(\Delta_0), Q_n(\Delta_{\inf})\big)$*–stabilizing.*

Proof. The proof is based on the following idea: it is shown that the qdb–controller is $\big(Q_n(\Delta_0), Q_n(\Delta_1)\big)$–stabilizing, with $\Delta_1 := \rho(\Delta_0)$. The argument is then iterated until $\Delta_{\hat{m}}$ is found such that $\rho(\Delta_{\hat{m}}) = \Delta_{\hat{m}}$ (it holds that $\Delta_{\hat{m}} = \Delta_{\inf}$) . The details of the proof of the theorem are given after the proof of some preliminary results needed to show these facts. ■

The main tool to prove Theorem 6 is the following result:

Lemma 10 (Main tool) *Consider system* (2.1) *, assume* **A0** *and that* $\alpha = \|A\|_\infty \geq 1$ *. Let* $\Delta > 0$ *be such that* $\rho(\Delta) < +\infty$ *and inequalities* (4.1a–b) *hold. Consider a qdb–controller* $k : \mathbb{R}^n \to \mathcal{U}$ *.*

ı) If $x \in Q_n(\Delta)$ *, then* $x^+ = Ax + Bk(x)$ *is such that*

$$|x_n^+| \leq \max\left\{\frac{\rho(\Delta)}{2}\,,\, \|x\|_\infty - \varphi(\Delta)\right\}\,, \tag{4.3}$$

where $\varphi(\Delta)$ *is defined in equation* (3.80) *of Section 3.2.2 .*

ıı) Let $\Delta' := \rho(\Delta)$ *: if* $\rho(\Delta) < \Delta$ *, then the qdb–controller is* $\big(Q_n(\Delta), Q_n(\Delta')\big)$*–stabilizing.*

Proof. *ı)* It is a particular case of Lemma 9 in Section 3.2.2 : in fact, inequalities (4.1a–b) imply inequalities (3.78a–b) , then it is sufficient to consider $\hat{x} = x$ and $\mathscr{H} = 0$; the only apparent discrepancy is the supplementary assumption $k(\hat{x}) \in \mathcal{U}(\Delta)$ in Lemma 9 but, as we notice in equation (3.81) , for $\hat{x} = x$ this hypothesis is satisfied.
Since this statement is a simplified version of Lemma 9 , it is useful to provide also an explicit proof of it. This can be found in Appendix A.3.1 .

ıı) It is a consequence of part *ı* and of the controller form of the system. In fact: $\varphi(\Delta) > 0$ by inequalities (4.1a-b) , therefore inequality (4.3) implies that $\forall\, \gamma \in [\Delta', \Delta]$, $Q_n(\gamma)$ is positively invariant. Moreover, because $x^+ = (x_2, \ldots, x_n, x_n^+)$, inequality (4.3) also implies that

$$\forall\, x(0) \in Q_n(\Delta)\,, \quad \|x(n)\|_\infty \leq \max\left\{\tfrac{\Delta'}{2} = \tfrac{\rho(\Delta)}{2}\,, \|x(0)\|_\infty - \varphi(\Delta)\right\} :$$

since $\varphi(\Delta) > 0$, the iteration of this argument yields the $\big(Q_n(\Delta), Q_n(\Delta')\big)$–stability. ■

Two more technical results are needed. These statements turn out to be useful also to prove other results that will follow in Section 4.2 and in Section 6.2.

Lemma 11 *Let $a < \Delta_0$ and $\phi : [a, \Delta_0] \to [a, \Delta_0[$ be a non–decreasing function. If ϕ is right continuous or ϕ takes a finite number of values for $\Delta \in [a, \Delta_0]$, then the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ defined by $\Delta_{k+1} := \phi(\Delta_k)$ is non–increasing and*

$$\Delta_{\inf} := \lim_{k\to+\infty} \Delta_k = \max\big\{\Delta < \Delta_0 \,|\, \phi(\Delta) = \Delta\big\}. \tag{4.4}$$

Moreover, $\forall\, \Delta \in]\Delta_{\inf}, \Delta_0]$ it holds that $\phi(\Delta) < \Delta$.

Proof. See in Appendix A.3.1. ■

Lemma 12 *Consider system (2.1), assume* **A0** *and that $\alpha \geq 1$. If $\Delta > 0$ satisfies inequalities (4.1a–b), then any $\Delta' \in [\rho(\Delta)\,,\, \Delta[$ satisfies inequalities (4.1a–b).*

Proof. See in Appendix A.3.1. ■

Proof of Theorem 6. For $\Delta \geq \bar{\Delta}$, the function ρ is right continuous, non–decreasing and $\rho : [\bar{\Delta}, +\infty[\to [\bar{\Delta}, +\infty[$ (see the discussion after equation (3.8) in Section 3.1.1). Therefore, thanks to inequality (4.1c), the restriction of ρ to the interval $[\bar{\Delta}, \Delta_0]$ satisfies the assumptions of Lemma 11. Hence, the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ defined by $\Delta_{k+1} := \rho(\Delta_k)$ is non–increasing and $\lim_{k\to+\infty} \Delta_k = \Delta_{\inf}$, with $\Delta_{\inf}$ as defined in equation (4.2). Actually, since for $\Delta \in [\bar{\Delta}\,, \Delta_0]$, $\rho(\Delta)$ takes only a finite number of values (see Remark 5), then $\exists\, \hat{m} \in \mathbb{N}$, $\hat{m} \geq 1$, such that $\Delta_{\hat{m}} = \Delta_{\inf}$. Thus, it is defined a finite and decreasing sequence

$$\Delta_0 > \Delta_1 > \cdots > \Delta_{\hat{m}-1} > \Delta_{\hat{m}} = \Delta_{\inf}. \tag{4.5}$$

The thesis of the theorem is achieved by showing that $\forall\, k = 0, \ldots, \hat{m} - 1$, the qdb–controller is $\big(Q_n(\Delta_k), Q_n(\Delta_{k+1})\big)$–stabilizing. To this end, we apply Lemma 10.*u*: it is sufficient to show that $\forall\, k = 0, \ldots, \hat{m} - 1$, Δ_k satisfies inequalities (4.1). Indeed, $\rho(\Delta_k) < \Delta_k$ by construction of the sequence (4.5); inequalities (4.1a–b) are satisfied by Δ_k as it follows by recursive application of Lemma 12. ■

Theorem 6 will be extended to more general controllers in Theorem 13 of Section 6.2.2.

Example 14 (Finite control set) *Let us consider the system in Example 6 of Section 3.1.1 and let $k : \mathbb{R}^2 \to \mathcal{U}$ be a qdb–controller. It holds that $\forall\, \Delta_0 \in]1, 96[$, the control law $u(x) = k(x)$ is*

$$\big(Q_2(\Delta_0)\,, Q_2(1)\big)\text{–stabilizing}.$$

In fact, by the computations done in Example 6, it follows that Δ satisfies inequalities (4.1) if and only if $1 < \Delta < 96$, and $\rho(1) = 1$: the result holds by Theorem 6. ♣

Example 15 (Uniformly quantized controls) *Similarly to Example 7, let us consider a system* $\Sigma(A,B,\mathcal{U})$ *in the controller form coordinates with* $\alpha = \|A\|_\infty \geq 1$ *and* $\mathcal{U}$ *is a uniformly quantized set with parameter* u_0. *Let* $k : \mathbb{R}^n \to \mathcal{U}$ *be a qdb–controller. It holds that* $\forall\, \Delta_0 > u_0$, *the control law* $u(x) = k(x)$ *is*

$$\big(Q_n(\Delta_0)\,, Q_n(u_0)\big)\text{–stabilizing in } n \text{ steps}$$

(see Definition 15b in Section 3.1.2).
In fact, by the computations done in Example 7, it follows that Δ *satisfies inequalities* (4.1) *if and only if* $\Delta > u_0$, *and* $\rho(u_0) = u_0$ *(the check that this is true also for* $\alpha = 1$ *is straightforward). Hence, the* $\big(Q_n(\Delta_0)\,, Q_n(u_0)\big)$*–stability of the closed loop dynamics holds by Theorem 6. To see that the convergence to within* $Q_n(u_0)$ *is guaranteed in* n *steps, it is sufficient to notice that the system is in controller form and that* $\forall\, x \in \mathbb{R}^n$, $|x_n^+| = \big|q_e\big(-(Ax)_n\big)\big| \leq \frac{u_0}{2}$ *(see equation* (3.3) *in Section 3.1.1 and Lemma 2 in Section 2.1).* ♣

Example 16 (Logarithmically quantized controls) *Similarly to Example 8, let us consider a system* $\Sigma(A,B,\mathcal{U})$ *in the controller form coordinates with* $\alpha = \|A\|_\infty \geq 1$, $\mathcal{U}$ *is a logarithmically quantized set with parameters* (u_0, θ) *and* $1 < \theta < \frac{\alpha+1}{\alpha-1}$ *(if* $\alpha = 1$, *simply assume that* $\theta > 1$*). Let* $k : \mathbb{R}^n \to \mathcal{U}$ *be a qdb–controller. It holds that* $\forall\, \Delta_0 > u_0$, *the control law* $u(x) = k(x)$ *is*

$$\big(Q_n(\Delta_0)\,, Q_n(u_0)\big)\text{–stabilizing}.$$

In fact, by equation (3.16) *in Remark 7 of Section 3.1.1,* Δ *satisfies inequalities* (4.1) *if and only if* $\Delta > u_0$, *and* $\rho(u_0) = u_0$*: the result holds by Theorem 6.* ♣

4.2 Practical stabilization: quantized single–input and quantized measurements

In this section, the practical stabilization technique presented in the previous section, as well as the practical stability analysis of the corresponding closed loop dynamics, is extended to the case where both the inputs and the measurements are quantized.
The qdb–controller k is a state feedback law: therefore, first an estimate $\hat{x}$ of the current state x has to be found by processing the quantized measurements, then a control law of the type $u = k(\hat{x})$ can be considered. In the quantized output case, we have seen in Section 3.2.2 how to construct a state observer to obtain $\hat{x}$. In the quantized state case, instead, the problem is easier: indeed, we shall see that it is reasonable to take the output $y \in \mathcal{Y}$ as an estimation of the current state.

4.2.1 Practical stabilization: state quantization

Let us consider system (2.1) in the quantized state case. According to the presentation in Section 2.2, we assume that, at least up to sufficiently large values of Δ,

$$H(\Delta) < +\infty$$

(see equation (3.68) in Section 3.2.1).
Let $\Delta > 0$ be such that $H(\Delta) < +\infty$ and, for $y \in \mathcal{Y}(\Delta)$, consider $\mathcal{C}_y$. Similarly to the definition of $c^*_{\text{mid}}(y)$ in equation (3.66), let

$$c_{\text{mid}}(y) := \frac{\sup\{\text{Pr}_n(A\mathcal{C}_y)\} + \inf\{\text{Pr}_n(A\mathcal{C}_y)\}}{2}$$

be the middle point of $\text{Pr}_n(A\mathcal{C}_y)$. By suitably redefining $q_{\mathcal{Y}}$ without modifying the induced state space partition (hence, without loss of generality), we can assume that

$$\begin{cases} \mathcal{Y} \subset \mathbb{R}^n \\ \forall\, y \in \mathcal{Y} \text{ such that } \mathcal{C}_y \text{ is bounded}, \ y \text{ is such that } (Ay)_n = c_{\text{mid}}(y). \end{cases} \tag{4.6}$$

We take $y = q_{\mathcal{Y}}(x)$ as an estimate of the current state x. The control action is then selected through a qdb–controller, as if the state was y. The main result on the practical stability properties of the resulting closed loop dynamics is the following:

Theorem 7 ((X_0, Ω)–stabilization: quantized state and quantized single–input reachable systems) *Consider system (2.1), assume* **A0** *and that $\alpha = \|A\|_\infty \geq 1$. Let $k : \mathbb{R}^n \to \mathcal{U}$ be a qdb–controller. If $\Delta_0 > 0$ is such that*

$$\begin{cases} m(\Delta_0) < -\dfrac{\Delta_0}{2}(\alpha - 1) & \text{(4.7a)} \\ M(\Delta_0) > \dfrac{\Delta_0}{2}(\alpha - 1) & \text{(4.7b)} \\ \rho(\Delta_0) + H(\Delta_0) < \Delta_0 & \text{(4.7c)} \end{cases}$$

and[1] *$\mathcal{U} = \mathcal{U}(\Delta_0)$, then it is well defined*

$$\Delta_{\text{inf}} := \max\{\Delta < \Delta_0 \,|\, \rho(\Delta) + H(\Delta) = \Delta\} \tag{4.8}$$

and the control law $u(x) = (k \circ q_{\mathcal{Y}})(x)$ is $\big(Q_n(\Delta_0), Q_n(\Delta_{\text{inf}})\big)$–stabilizing.

Proof. The proof is based on the same we idea used to prove Theorem 6 in previous Section 4.1. Also in this case, a preliminary result is needed. Hence, the details of the proof of the theorem are reported after the proof of next Lemma 13. ■

The main tool to prove Theorem 7 is the following result:

Lemma 13 (Main tool) *Consider system (2.1), assume* **A0** *and that $\alpha = \|A\|_\infty \geq 1$. Let $\Delta > 0$ be such that inequalities (4.7) hold. Consider $\Delta' := \rho(\Delta) + H(\Delta)$ and let $k : \mathbb{R}^n \to \mathcal{U}$ be a qdb–controller: if $(k \circ q_{\mathcal{Y}})\big(Q_n(\Delta)\big) \subseteq \mathcal{U}(\Delta)$, then the controller $(k \circ q_{\mathcal{Y}})$ is $\big(Q_n(\Delta), Q_n(\Delta')\big)$–stabilizing and $(k \circ q_{\mathcal{Y}})\big(Q_n(\Delta')\big) \subseteq \mathcal{U}(\Delta')$.*

[1] As far as the assumption $\mathcal{U} = \mathcal{U}(\Delta_0)$ is concerned, see the discussion in Remark 14 of Section 3.2.2.

Proof. First notice that, $\forall\, x \in Q_n(\Delta)$, the hypotheses of Lemma 9 in Section 3.2.2 are satisfied with $\hat{x} := q_y(x)$ and $\mathscr{H} = H(\Delta)$. In fact: inequalities (4.7) imply inequalities (3.78); the assumption $(k \circ q_y)\big(Q_n(\Delta)\big) \subseteq \mathcal{U}(\Delta)$ ensures that $k(\hat{x}) \in \mathcal{U}(\Delta)$; finally,

$$\big|\big(A(x-\hat{x})\big)_n\big| \overset{\text{(a)}}{=} \big|(Ax)_n - c_{\text{mid}}(\hat{x})\big| \overset{\text{(b)}}{\leq} \tfrac{h(\hat{x})}{2} \overset{\text{(c)}}{\leq} \tfrac{H(\Delta)}{2}\,,$$

where equality (a) follows by the assumption in equation (4.6), inequalities (b) and (c) hold by definition of $h(\hat{x})$ $\big($see equation (3.67)$\big)$ and $H(\Delta)$ $\big($see equation (3.68)$\big)$, respectively. Thus, $x^+ = Ax + B(k \circ q_y)(x)$ is such that inequality (3.79) holds with $\mathscr{H} = H(\Delta)$. Moreover, thanks to inequalities (4.7a–b), $\varphi(\Delta) > 0$ $\big($see equation (3.80)$\big)$. Hence, inequality (3.79) implies that $\forall\, \gamma \in [\Delta', \Delta]$, $Q_n(\gamma)$ is positively invariant and, by Lemma 6 in Section 3.1.1, it holds that $(k \circ q_y)\big(Q_n(\gamma)\big) \subseteq \mathcal{U}(\gamma)$. Moreover, because $x^+ = (x_2, \ldots, x_n, x_n^+)$, inequality (3.79) also implies that

$$\forall\, x(0) \in Q_n(\Delta)\,, \quad \|x(n)\|_\infty \leq \max\left\{\tfrac{\Delta'}{2} = \tfrac{\rho(\Delta)+H(\Delta)}{2}\,, \|x(0)\|_\infty - \varphi(\Delta)\right\} :$$

the iteration of this argument yields the $\big(Q_n(\Delta), Q_n(\Delta')\big)$–stability because $\varphi(\Delta) > 0$. ∎

Proof of Theorem 7. For $\Delta \in [\bar{\Delta}, \Delta_0]$, let $\phi(\Delta) := \rho(\Delta) + H(\Delta)$. It holds that: the function ϕ is non–decreasing because so are both ρ and H; $\phi(\bar{\Delta}) \geq \bar{\Delta}$ because $\phi(\bar{\Delta}) > \rho(\bar{\Delta}) \geq \bar{\Delta}$; $\phi(\Delta_0) < \Delta_0$ thanks to inequality (4.7c); and, for $\Delta \in [\bar{\Delta}, \Delta_0]$, it takes only a finite number of values because this property holds for both ρ (see Remark 5 in Section 3.1.1) and H $\big($see the discussion on the properties of $H(\Delta)$ after equation (3.68)$\big)$. Hence, by Lemma 11, the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ defined by $\Delta_{k+1} := \rho(\Delta_k) + H(\Delta_k)$ is non–increasing and $\lim_{k\to+\infty} \Delta_k = \Delta_{\text{inf}}$, with Δ_{inf} as defined in equation (4.8). Moreover, $\exists\, \hat{m} \in \mathbb{N}$, $\hat{m} \geq 1$, such that $\Delta_{\hat{m}} = \Delta_{\text{inf}}$. Thus, it is defined a finite and decreasing sequence

$$\Delta_0 > \Delta_1 > \cdots > \Delta_{\hat{m}-1} > \Delta_{\hat{m}} = \Delta_{\text{inf}}\,. \tag{4.9}$$

The thesis of the theorem is achieved by showing that $\forall\, k = 0, \ldots, \hat{m}-1$, the control law $u(x) = (k \circ q_y)(x)$ is $\big(Q_n(\Delta_k), Q_n(\Delta_{k+1})\big)$–stabilizing. To this end, we apply Lemma 13: it is sufficient to show that $\forall\, k = 0, \ldots, \hat{m}-1$, Δ_k satisfies inequalities (4.7) and $(k \circ q_y)\big(Q_n(\Delta_k)\big) \subseteq \mathcal{U}(\Delta_k)$. Indeed, $\rho(\Delta_k) < \Delta_k$ by construction of the sequence (4.9); inequalities (4.7a–b) are satisfied by Δ_k as it follows by recursive application of Lemma 12; finally, $(k \circ q_y)\big(Q_n(\Delta_k)\big) \subseteq \mathcal{U}(\Delta_k)$: this holds for $k = 0$ as we assumed $\mathcal{U} = \mathcal{U}(\Delta_0)$, while for $k = 1, \ldots, \hat{m}-1$ it follows by recursive application of Lemma 13. ∎

Example 17 (Finite control set and quantized state) *Let us consider the system in Example 12 of Section 3.2.1. It holds that* $\forall\, \Delta_0 \in \left]\frac{5}{2}, 64\right]$, *the control law* $u(x) = (k \circ q_y)(x)$, *where* $k : \mathbb{R}^2 \to \mathcal{U}(\Delta_0)$ *is a qdb–controller with saturated inputs (see Remark 14 in Section 3.2.2), is*

$$\Big(Q_2(\Delta_0)\,, Q_2\big(\tfrac{5}{2}\big)\Big)\text{–stabilizing}\,.$$

In fact, by the computations done in Example 12, it follows that Δ *satisfies inequalities* (4.7) *if and only if* $\frac{5}{2} < \Delta \leq 64$, *and* $\rho\big(\frac{5}{2}\big) + H\big(\frac{5}{2}\big) = \frac{5}{2}$: *the result holds by Theorem 7.* ♣

4.2.2 Practical stabilization: output quantization

Let us consider system (2.1) in the quantized single–output case. The framework and the terminology is the same as in Section 3.2.2 . We define a dynamic qdb–controller: it is a controller taking the more general form in equation (2.9) and consisting of a qdb–controller k but the control action is selected as if the current state was $\hat{x}(t)$, the state estimation resulting from the quantized state–observer (see equation (3.74)). We then analyze the practical stability properties of the resulting closed loop dynamics. Since the output $\hat{x}(t) = \psi\big(\vec{y}\,(t)\,,\vec{u}\,(t)\big)$ of the quantized state–observer is available only after a transient (i.e., for $t \geq n-1$) , then the controller needs to be initialized for $t \leq n-2$: this is the reason why, in this case, only (X_0, X_1, Ω)–stability can be guaranteed rather than (X_0, Ω)–stability.
Let the *dynamic qdb–controller* be defined as follows: denote by $k : \mathbb{R}^n \to \mathcal{U}$ a qdb–controller and let

$$u(t) := \begin{cases} 0 & \text{if } \ t \leq n-2 \\ (k \circ \psi)\big(\vec{y}\,(t), \vec{u}\,(t)\big) & \text{if } \ t \geq n-1\,. \end{cases} \tag{4.10}$$

This controller can be modelled in the form of equation (2.9) with $\mathcal{W} := \mathcal{Y}^n \times \mathcal{U}^{n-1}$ and suitably defined maps γ and $\bar{k}$: see the details in Appendix A.3.2 .
The corresponding practical stability result is the following:

Theorem 8 ((X_0, X_1, Ω)–stabilization: quantized single–output and quantized single–input reachable systems) *Consider system* (2.1) *, assume* **A0** *and that* $\alpha = \|A\|_\infty \geq 1$ *. Suppose that the system is quantized single–output (see Section 2.2) with* $q_y = q_o \circ C$ *, where* $C \in \mathbb{R}^{1\times n}$ *and* (A, C) *is an observable pair. Let* $\Delta_1 > 0$ *be such that*

$$\begin{cases} m(\Delta_1) < -\dfrac{\Delta_1}{2}\big(\alpha - 1\big) & \text{(4.11a)} \\ M(\Delta_1) > \dfrac{\Delta_1}{2}\big(\alpha - 1\big) & \text{(4.11b)} \\ \rho(\Delta_1) + \widetilde{H}(\Delta_1) < \Delta_1 & \text{(4.11c)} \end{cases}$$

and[2] $\mathcal{U} = \mathcal{U}(\Delta_1)$ *. Consider* $\Delta_0 := \frac{\Delta_1}{\|A^{n-1}\|_\infty}$ *, then it is well defined*

$$\Delta_{\inf} := \max\big\{\Delta < \Delta_1 \,|\, \rho(\Delta) + \widetilde{H}(\Delta) = \Delta\big\} \tag{4.12}$$

and the dynamic qdb–controller (4.10) *is* $\big(Q_n(\Delta_0), Q_n(\Delta_1), Q_n(\Delta_{\inf})\big)$*–stabilizing.*

Proof. For $\Delta \in [\bar{\Delta}, \Delta_1]$, let $\phi(\Delta) := \rho(\Delta) + \widetilde{H}(\Delta)$. It holds that: the function ϕ is non–decreasing because so are both ρ and $\widetilde{H}$; $\phi(\bar{\Delta}) \geq \bar{\Delta}$ because $\phi(\bar{\Delta}) > \rho(\bar{\Delta}) \geq \bar{\Delta}$; $\phi(\Delta_1) < \Delta_1$ thanks to inequality (4.11c) ; and, for $\Delta \in [\bar{\Delta}, \Delta_1]$, it takes only a finite number of values because this property holds for both ρ (see Remark 5 in Section 3.1.1) and $\widetilde{H}$ (see the discussion on the properties of $\widetilde{H}(\Delta)$ after equation (3.77)). Hence, by Lemma 11 , the sequence $\{\Delta_h\}_{h\in\mathbb{N}\setminus\{0\}}$ defined by $\Delta_{h+1} := \rho(\Delta_h) + \widetilde{H}(\Delta_h)$ is non–increasing

[2] As far as the assumption $\mathcal{U} = \mathcal{U}(\Delta_1)$ is concerned, see the discussion in Remark 14 of Section 3.2.2 .

and $\lim_{h\to+\infty}\Delta_h = \Delta_{\inf}$, with $\Delta_{\inf}$ as defined in equation (4.12). Moreover, $\exists\,\hat{m}\in\mathbb{N}$, $\hat{m}\geq 2$, such that $\Delta_{\hat{m}} = \Delta_{\inf}$. Thus, it is defined a finite and decreasing sequence

$$\Delta_1 > \Delta_2 > \cdots > \Delta_{\hat{m}-1} > \Delta_{\hat{m}} = \Delta_{\inf}\,. \tag{4.13}$$

Let us analyze the closed–loop dynamics under the dynamic qdb–controller (4.10). For $t\leq n-1$, by the controller form of A, it holds that

$$\forall\, x(0)\in Q_n(\Delta_0) \text{ and } \forall\, t\leq n-1\,,\quad x(t)\in Q_n(\Delta_0\|A^{n-1}\|_\infty) = Q_n(\Delta_1)\,. \tag{4.14}$$

For notational convenience, we let $t_1 := 0$. The hypotheses of Theorem 5 in Section 3.2.2 are satisfied with $\Delta=\Delta_1$ and $t = t_1+n-1$: this follows by the property (4.14) and by the assumption $\mathcal{U}=\mathcal{U}(\Delta_1)$. Therefore, since for $t\geq n-1$, $u(t) = k\big(\hat{x}(t)\big) = q_{\mathcal{U}}\Big(-\big(A\hat{x}(t)\big)_n\Big)$, then it holds that

$$\forall\, t\geq n\,,\quad x(t)\in Q_n(\Delta_1)\,.$$

Furthermore, the application of Lemma 9 as in the proof of Theorem 5, guarantees that

$$\forall\, t\geq n\,,\quad |x_n(t)|\leq \max\Big\{\tfrac{\Delta_2}{2} = \tfrac{\widetilde{H}(\Delta_1)+\rho(\Delta_1)}{2}\,, \|x(t-1)\|_\infty - \varphi(\Delta_1)\Big\}\,, \tag{4.15}$$

where $\varphi(\Delta_1)$ is defined in equation (3.80). Since the system is in controller form and, by inequalities (4.11a–b), $\varphi(\Delta_1)>0$, then inequality (4.15) implies that

$$\exists\, t_2>0 \text{ such that } \forall\, t\geq t_2\,,\quad x(t)\in Q_n(\Delta_2)\,. \tag{4.16}$$

If $\hat{m}>2$ $\big($see the sequence (4.13)$\big)$, then we claim that at time t_2+n-1 we are in the position to repeat the above arguments and to prove that $\exists\, t_3>0$ such that $\forall\, t\geq t_3$, $x(t)\in Q_n(\Delta_3)$. To this end, it is sufficient to show that the hypotheses of Theorem 5 are satisfied with $\Delta=\Delta_2$ and $t=t_2+n-1$, and that the strict inequalities (4.11a–b), which ensure that $\varphi(\Delta_2)>0$ and guarantee the convergence to within $Q_n(\Delta_3)$, hold true. Let us do this check. By the property (4.16), two facts follow: first, $\forall\,\tau = t_2,\ldots,t_2+n-1$, $x(\tau)\in Q_n(\Delta_2)$; secondly, thanks to Lemma 6, $\forall\, t\geq t_2$, $k\big(\hat{x}(t)\big)\in\mathcal{U}(\Delta_2)$. The latter property means that there is no contradiction in arguing as if $\mathcal{U}=\mathcal{U}(\Delta_2)$. Finally, inequalities (4.11a–b) are satisfied by Δ_2 thanks to Lemma 12.
The arguments above can be repeated until $t_{\hat{m}}$ is found such that $\forall\, t\geq t_{\hat{m}}$, $x(t)\in Q_n(\Delta_{\hat{m}})$. This proves the $\big(Q_n(\Delta_0), Q_n(\Delta_1), Q_n(\Delta_{\inf})\big)$–stability of the closed loop dynamics. ■

Example 18 (Finite control set and quantized output) *Let us consider the system in Example 13 of Section 3.2.2. It holds that* $\forall\,\Delta_1\in\big]\frac{25}{7},\frac{234}{11}\big]$, *the dynamic qdb–controller* (4.10) *with saturated inputs* $\mathcal{U}=\mathcal{U}(\Delta_1)$ *(see Remark 14 in Section 3.2.2), is*

$$\Big(Q_2\big(\tfrac{\Delta_1}{\alpha}\big)\,, Q_2(\Delta_1)\,, Q_2\big(\tfrac{5}{2}\big)\Big)\text{–}\textit{stabilizing}\,,$$

where $\alpha = \|A\|_\infty = \frac{3}{2}$.
In fact, by the computations done in Example 13, it follows that Δ *satisfies inequalities* (4.11)

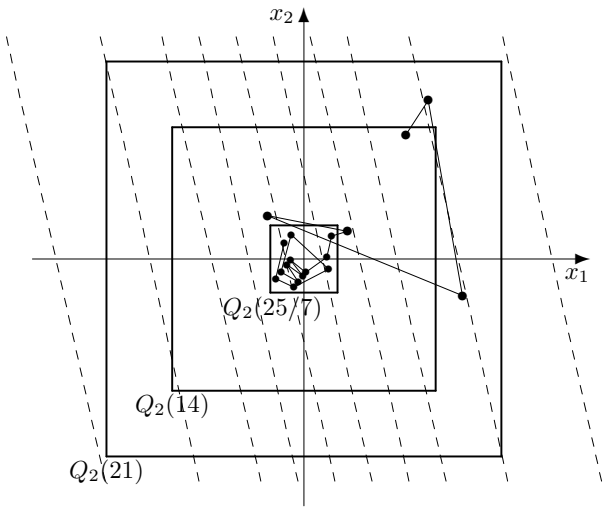

Figure 4.1: A trajectory generated by the dynamic qdb–controller for the system in Example 18 with $\Delta_1 = 21$ ($\Delta_0 = 14$) and $x(0) = (5.42\ 6.60)$. Broken lines identify the state space partition induced by $q_{\mathcal{Y}}$.

if and only if $\frac{25}{7} < \Delta \leq \frac{234}{11}$, *and* $\rho\left(\frac{25}{7}\right) + \widetilde{H}\left(\frac{25}{7}\right) = \frac{25}{7}$ *: the result holds by Theorem 8. An example of trajectory, corresponding to the case of the dynamic qdb–controller with saturated inputs* $\mathcal{U} = \mathcal{U}(\Delta_1)$, $\Delta_1 = 21$, *is reported in Fig. 4.1.* ♣

Remark 16 (On the design of the I/O quantization) *In this thesis, our first concern is the stabilizability analysis under* assigned *quantization. However, the presented results can be applied also to* design *quantization so as to guarantee that desired stability properties hold true. To ensure invariance, it is sufficient to design the control set* $\mathcal{U}$ *and the output map* $q_{\mathcal{Y}}$ *so that the functions* $\rho(\Delta) + H(\Delta)$, $m(\Delta)$ *and* $M(\Delta)$ *satisfy the hypotheses of Corollary 4 in Section 3.2.1 (or of Theorem 5 in Section 3.2.2 if quantization is on the output); if also convergence properties are desired, then it is sufficient to satisfy the hypotheses of Theorem 7 in Section 4.2.1 (or of Theorem 8). This can be done by elementary computations.*
For instance, when $q_{\mathcal{Y}}(x) = x$ *(i.e.,* $H(\Delta) \equiv 0$*), a control set ensuring the invariance of* $Q_n(\Delta)$ *can be constructed according to the conditions provided by Theorem 1 in Section 3.1.1. Among these sets, one of minimal cardinality is a saturated version of a uniformly quantized control set with parameter* Δ *and it is made up of approximately* $\|A\|_\infty$ *elements. In [126, 89, 80], it has been proved that the minimal number of control symbols necessary and sufficient for stabilization is*

$$\#\mathcal{U} \simeq \Pi_{\lambda_{\mathrm{u}}(A) \in \mathcal{S}_{\mathrm{u}}(A)} |\lambda_{\mathrm{u}}(A)| \,,$$

where $\mathcal{S}_{\mathrm{u}}(A) := \big\{ \lambda(A) \,|\, |\lambda(A)| > 1 \big\}$. *Compared with this bound, our result can be conservative. For instance, if* A *is antistable* $\big($*that is,* $\mathcal{S}(A) = \mathcal{S}_{\mathrm{u}}(A)\big)$*, then*

$$\Pi_{\lambda_{\mathrm{u}}(A) \in \mathcal{S}_{\mathrm{u}}(A)} |\lambda_{\mathrm{u}}(A)| = |\alpha_1| \leq \|A\|_\infty = \textstyle\sum_{i=1}^{n} |\alpha_i| \,.$$

On the other hand, our controller guarantees better performance in terms of convergence rate because it prevents the trajectories from making large excursions while converging towards the equilibrium. Indeed, our theory can be usefully combined with so–called "zooming" [17] or "nesting" [45] techniques: examples have been provided in [45] showing that a nesting version of our controller approaches optimal theoretical bounds relating the number of control values and the speed of convergence $\big($*both considered as functions of the contraction* $\mathtt{C} := \mathit{Volume}\,(X_0)/\,\mathit{Volume}\,(\Omega)\,\big)$.

Chapter 5

The small–gain approach in H_∞: quantized multi–input

5.1 Introduction to the small–gain approach

In this chapter and in next Chapter 6, we develop a more systematic approach to the problem of the synthesis of practically stabilizing controllers. The proposed methodology is based on the so called small–gain theorems, a class of results on the stability analysis of nonlinear feedback systems (see Fig. 5.1). In this framework, the system is viewed as a "black box". That is, the behavior of the system is described by an input/output operator that specifies the output of the system as a function of the input only, while the internal dynamics is not explicitly involved. Although the norm of this operator provides a rough description of the properties of the system, yet this information can be enough for control synthesis purposes. In fact, if the system under consideration can be decomposed into the feedback interconnection of subsystems and the norms of the operators associated to these subsystems satisfy a suitable relation, the so called small–gain condition, then stability of the overall dynamics can be asserted. The stabilization procedure based on small–gain theorems exactly consists in finding a controller so that the norms of the operators associated to the subsystems satisfy the desired relation. It is hence an abstract methodology, meaning that the control synthesis

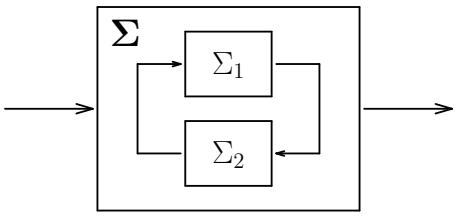

Figure 5.1: Factorization of an overall dynamics Σ into the feedback interconnection of two subsystems Σ_1 and Σ_2.

for stabilization is converted into the search for suitable operators within a functional space. In both Chapters 5 and 6 , we consider system (2.1) in the quantized input case $\Sigma(A, B, \mathcal{U})$, while full state is assumed to be available. A natural, but not necessarily successful, approach to the control synthesis consists of considering controllers resulting from the quantization of a control law that guarantee asymptotic stability in the ideal conditions of absence of quantization. By operating in this way, the outcome is a closed loop dynamics that can be described as the feedback interconnection of the ideal closed loop dynamics with a nonlinearity representing the quantization error (see Example 3 in Section 2.3) . Stabilization can be achieved if the ideal controller has robustness properties with respect to the quantization error. In general, the inference of stability properties from the analysis of the subsystems constituting a feedback loop is not a trivial issue. For instance, it is well known that the interconnection of stable systems may return an overall unstable dynamics. When nonlinear terms are included into the loop, there is not a complete and exhaustive theory for the stability analysis. There are instead various techniques and, although some of them are of quite general applicability, they typically provide only sufficient conditions for stability. In this respect, classical methods are those based on the absolute stability criteria [67] and include small–gain theorems. As it will be shown into details in next Section 5.2 , the gain of a system is a generalization of the notion of norm for the input/output operator representing the system. Namely, the gain is a measure of the amplification (or attenuation) of the norm of a signal as it passes through the system. The small–gain condition ensures that the norm of the signals circulating in the feedback loop remains bounded and, under suitable assumptions, this is also sufficient to prove stability of the internal dynamics of the system. Nevertheless, the outcome of a classical small–gain theorem is typically Lyapunov stability. Since open loop unstable discrete–time quantized systems are not stabilizable in the classical sense, this approach cannot be directly applied to this kind of systems. As shown in Example 2 of Section 2.3 , the reason why asymptotic stability is not achievable in the presence of quantization lies in the loss of control resolution as the trajectories approach the origin, which is indeed caused by the truncation of the quantizer. In particular, it can be seen that the gain associated to the quantization error is dictated by the behavior of the corresponding operator exactly in the proximity of 0 . Moreover, the value of such a gain is too large so that there is no way to satisfy the classical small–gain condition (see, for instance, Remark 21 in Section 5.3.2) . Consequently, a relaxed notion of gain for the quantization error is needed that does not take into account for the input/output relation when the input is small. Accordingly, we introduce *generalized* notions of gain for a static nonlinear map which essentially consist of the classical gain but the behavior of the nonlinearity in a neighborhood of the origin is overlooked. This allows us to prove generalized versions of the small–gain theorem whose outcome is practical stability. Indeed, a small–gain condition given in terms of the gain of the ideal system (i.e., in the absence of quantization) and of the generalized gain of the quantization error, together with a bound on the quantization error in a neighborhood of the origin, guarantees that the closed loop system is practically stable. A quantitative analysis of the size of the final invariant set within which trajectories are confined is included. This result is the basis to

carry out a systematic procedure for the solution of the stabilization problem in the presence of input quantization. In fact, the synthesis of practically stabilizing controllers is obtained by direct application of the theoretical tools that allow one to analyze the generalized gain of a nonlinear operator and to synthesize a controller that guarantees that the gain of the ideal system is below a threshold ensuring that the small–gain condition is met. Moreover, a family of quantized controllers can be obtained simply by tuning the design parameters as, for instance, the desired gain of the ideal system or, if also the quantized control set can be chosen, the generalized gain of the quantization error. In this way, it is possible to study more general problems than mere practical stabilization, where also requests on the closed loop performance can be enforced (such as a minimal guaranteed speed of convergence or desired practical stability properties expressed in terms of the size of the set within which trajectories can be confined, see Example 24 at the end of Section 5.3.3) .

Since the gain of a system depends on the particular choice of the norm used to measure signals, the control synthesis technique and the practical stability analysis based on a small–gain theorem changes according to the variation of the considered norm. In this thesis we focus on two cases: the ℓ_2–norm and the ℓ_∞–norm. The former case is extensively faced in Section 5.3 . When the ℓ_2–norm is considered, the input/output operator associated to the system is an element of the so called Hardy's functional space H_∞ and the control synthesis for practical stabilization can be transformed into a particular control problem in H_∞ (see [132, 117]) . The corresponding practical stability analysis is based on Lyapunov arguments for quadratic functions and on invariant ellipsoids. The other case, when the ℓ_∞–norm is considered, is studied in next Chapter 6 : here, the corresponding input/output operator is naturally associated with an element of the functional space ℓ_1 , hence the practical stabilization problem can be turned into a control problem in ℓ_1 (see [128, 34]) . The corresponding practical stability analysis provides hypercubes within which the trajectories are proved to be confined.

The choice of studying the practical stabilization in the ℓ_1 functional space is natural when quantized control sets, instead of the most often encountered case of *generalized* quantized sets (see Definition 3 in Section 2.1) , are considered. In fact, as a consequence of the fact that the quantized control values are isolated and do not have accumulation points, the quantization error is a persistent (i.e., non–vanishing) disturbance, hence to be treated as a signal in ℓ_∞ . Accordingly, the steady–state analysis based on results from the control in ℓ_1 appears to be less conservative than the one based on H_∞ theory. This fact has a counterpart in the minimality properties of hypercubes proved in Section 3.1.2 . Nevertheless, the literature on the control in ℓ_1 is not as exhaustive as that for the H_∞ control and, somehow, the same holds for the theory we develop here. In fact, some results presented in the framework of ℓ_1 theory are less general than their analogous in the H_∞ case. Anyhow, the complementation of the H_∞ approach with the ℓ_1 theory is shown to bring significant contributions to the stabilization problem. In this respect, our main result is a generalized small–gain theorem for practical stability analysis (see Theorem 12 in Section 6.2.1) . The combination of this theorem with the practical stability analysis relying on H_∞ theory leads to a mixed H_∞/ℓ_1

analysis tool whose potency is pointed out through some numerical examples. Moreover, also a solution to the stabilization problem is given in terms of a mixed H_∞/ℓ_1 control synthesis problem whose study is one of the most interesting open issues for future investigations. These results has allowed us to provide a first extension to multi–input systems of the practical stabilization technique presented in Chapter 4 and based on the analysis of invariant hypercubes.

Next sections are organized as follows: in Section 5.2, the main definitions on the norms of signals and systems are briefly reviewed and the notation for the various operators and functional spaces is fixed. In Section 5.3, a generalized small–gain theorem is proved for operators belonging to the H_∞ functional space and the application to the practical stabilization of quantized input linear systems are illustrated. Thorough explanations are provided for the implementation of the proposed technique and several numerical examples are reported. Chapter 6 follows the same line but in the case of the ℓ_1 functional space, it includes results on the mixed H_∞/ℓ_1 analysis and an example of the mixed H_∞/ℓ_1 control synthesis.

5.2 Signals and systems

In this section, we briefly review the definitions and some classical results on the norms of signals and systems. The presentation is restricted to the notions needed for the understanding of the subsequent sections. For a comprehensive presentation we refer to [67] (which, although devoted to continuous–time systems, has partially inspired the organization of this section) or to any good textbook on systems theory.

By a *signal* we mean a function $\vec{v} : \mathbb{N} \to \mathbb{R}^{h\times k}$. For $t \in \mathbb{N}$, let $v(t)$ be the t–th element of the sequence defining $\vec{v}$, hence we may write $\vec{v} = \{v(t)\}_{t\in\mathbb{N}}$. The signal $\vec{v}$ is said to be *positive* iff $\forall\, i = 1,\dots,h$, $\forall\, j = 1,\dots,k$ and $\forall\, t \in \mathbb{N}$, $v_{i,j}(t) \geq 0$. The *null* signal $\vec{v} \equiv 0$ is denoted by $\vec{0}$. The shift operator σ is defined by

$$\sigma\vec{v} := \{v(t+1)\}_{t\in\mathbb{N}}\,, \tag{5.1}$$

its k–th iteration is $\sigma^k\vec{v} = \{v(t+k)\}_{t\in\mathbb{N}}$.
For any given norm $\|\cdot\|_*$ on $\mathbb{R}^{h\times k}$ and for $p \in [1,\infty]$, let

$$\ell_p(\mathbb{R}^{h\times k}) := \begin{cases} \{\vec{v} : \mathbb{N} \to \mathbb{R}^{h\times k} \mid \sum_{t=0}^{+\infty} \|v(t)\|_*^p < +\infty\} & \text{if} \quad p \in [1,\infty[\\ \left\{\vec{v} : \mathbb{N} \to \mathbb{R}^{h\times k} \mid \sup\limits_{t\in\mathbb{N}} \|v(t)\|_* < +\infty\right\} & \text{if} \quad p = \infty\,. \end{cases}$$

In both cases, $\ell_p(\mathbb{R}^{h\times k})$ is a normed space with

$$\|\vec{v}\|_p := \begin{cases} \left(\sum_{t=0}^{+\infty} \|v(t)\|_*^p\right)^{1/p} & \text{if} \quad p \in [1,\infty[\\ \sup\limits_{t\in\mathbb{N}} \|v(t)\|_* & \text{if} \quad p = \infty\,. \end{cases} \tag{5.2}$$

While the value of $\|\vec{v}\|_p$ depends on the particular choice of the norm $\|\cdot\|_*$ on $\mathbb{R}^{h\times k}$, on the other hand, by the equivalence of the norms defined on finite dimensional vector spaces, the set $\ell_p(\mathbb{R}^{h\times k})$ does not. The choice of the norm $\|\cdot\|_*$ can be hence specified from time to time. From now on, we assume that a vector norm $\|\cdot\|_*$ has been fixed on $\mathbb{R}^l$ for any $l\in\mathbb{N}$, and, $\forall\, M\in\mathbb{R}^{l_1\times l_2}$, we consider the induced operator norm

$$\|M\|_* := \sup_{x\in\mathbb{R}^{l_2}\setminus\{0\}} \frac{\|Mx\|_*}{\|x\|_*}\,.$$

In this section, we deal with the following three cases of signal norms: $p=1$, $p=2$ or $p=\infty$. As for the choice of the norm $\|\cdot\|_*$, when $p=1$ we assume that

$$\|v(t)\|_* := \|v(t)\|_\infty = \max_{i=1,\dots,h} \sum_{j=1}^{k} |v_{i,j}(t)|\,; \tag{5.3}$$

when $p=2$ we assume that

$$\|v(t)\|_* := \|v(t)\|_2 = \sqrt{\rho\big(v(t)'\cdot v(t)\big)}\,;$$

when $p=\infty$ we assume that

$$\|v(t)\|_* := \|v(t)\|_\infty = \max_{i=1,\dots,h} \sum_{j=1}^{k} |v_{i,j}(t)|\,.$$

Let us consider an input/output dynamical system of the type

$$\begin{cases} x(t+1) = f\big(x(t)\,,u(t)\big) \\ y(t) = h\big(x(t)\big) \\ x(0)\in\mathbb{R}^n \\ u\in\mathbb{R}^m,\ y\in\mathbb{R}^q, \end{cases} \tag{5.4}$$

assume that $(x=0, u=0)$ is an equilibrium pair, namely $f(0,0)=0$, and that $h(0)=0$. Such a system is denoted by $\Sigma_{x(0)}$ (where the subscript aims at stressing the dependence from the initial condition). For a given input signal $\vec{u}=\{u(t)\}_{t\in\mathbb{N}}$, it is univocally identified the corresponding output signal $\vec{y}(\vec{u})=\{y(t)\}_{t\in\mathbb{N}}$, where

$$\begin{cases} y(0) := h\big(x(0)\big) \\ y(t) := h\Big(f\big(x(t-1)\,,u(t-1)\big)\Big) \quad \text{for} \quad t>0\,. \end{cases}$$

There is hence an input/output relation that may be represented as an operator $\varsigma_{x(0)}$ between suitable signal sets $\ell^{(u)}$ and $\ell^{(y)}$:

$$\begin{array}{rccc} \varsigma_{x(0)}\ : & \ell^{(u)} & \to & \ell^{(y)} \\ & \vec{u} & \mapsto & \vec{y}(\vec{u})\,. \end{array} \tag{5.5}$$

When such a relation is "well behaved", the system is said to be *input/output* stable. More precisely:

Definition 22 *Let* $p \in [1\,,\,\infty]$*, system* $\Sigma_{x(0)}$ *(or, equivalently, the input/output operator* $\varsigma_{x(0)}$*) is said to be* ℓ_p–stable *iff* $\forall\,\vec{u} \in \ell_p(\mathbb{R}^m)$*,* $\vec{y}(\vec{u}) \in \ell_p(\mathbb{R}^q)$*.*
If moreover, $\exists\,\gamma \geq 0$ *and* $\beta \geq 0$ *such that*

$$\forall\,\vec{u} \in \ell_p(\mathbb{R}^m)\,, \quad \|\vec{y}(\vec{u})\|_p \leq \gamma \|\vec{u}\|_p + \beta\,, \tag{5.6}$$

then the system is said to be finite–gain ℓ_p*–stable.*
If there exists $\min\,\{\gamma \geq 0\,|\,\exists\,\beta \geq 0$ *such that relation* (5.6) *is satisfied*$\}\,:=\,\gamma_s$*, then* γ_s *is said to be* the ℓ_p–gain *of the system.*

The ℓ_p–stability is concerned with the input/output relation only and, in general, it does not say anything on the properties of the internal dynamics $\vec{x} = \{x(t)\}_{t\in\mathbb{N}}$. The notion of gain is a generalization of the concept of norm of an operator to the case where $\varsigma_{x(0)}(\vec{0}) \neq \vec{0}$: the affine term β in inequality (5.6) is such that $\|\vec{y}(\vec{0})\|_p \leq \beta$ and it is called *bias* term.

Linear systems are a particular class of system (5.4): let us discuss input/output stability for these special models. In this case, it turns out to be convenient to denote by e the input of the system. For a linear system of the type

$$\begin{cases} x(t+1) = Fx(t) + Be(t) \\ y(t) = Cx(t) \\ x(0) \in \mathbb{R}^n \\ e \in \mathbb{R}^m,\ y \in \mathbb{R}^q, \end{cases} \tag{5.7}$$

it holds that $y(t) = y_{\rm a}(t) + y_{\rm f}(t)$, where $y_{\rm a}(t) := CF^t x(0)$ is the *autonomous* output of the system and $y_{\rm f}(t) := \sum_{\tau=0}^{t-1} CF^{t-\tau-1}Be(\tau)$ is the *forced* output $\big($which is not depending on $x(0)\big)$. A sufficient condition in order that system $\Sigma_{x(0)}$ is ℓ_p–stable $\forall\,x(0) \in \mathbb{R}^m$ and $\forall\,p \in [1\,,\,\infty]$ is that F is a Schur matrix. Such a condition is not necessary in general, nevertheless this is the only case we are interested in for our analysis.
Let us analyze the output signals $\vec{y}_{\rm a}$ and $\vec{y}_{\rm f}$ under the assumption that F is a Schur matrix. As far as the autonomous output is concerned, it holds that $\lim_{t\to+\infty} y_{\rm a}(t) = 0$ with an exponential decaying rate. More precisely, two constants $b \geq 1$ and $0 \leq r < 1$ can be computed such that

$$\forall\,t \geq 0\,, \quad \|F^t\|_* \leq b\,r^t\,. \tag{5.8}$$

Also,

$$\inf\,\{r \geq 0\,|\,\exists\,b \geq 1 \text{ such that relation (5.8) holds}\} = \rho(F)\,.$$

Therefore, $\exists\,\tilde{\beta} \geq 0$ such that $\|y_{\rm a}(t)\|_* = \|CF^t x(0)\|_* \leq \tilde{\beta}\,r^t$ $\big($e.g., $\tilde{\beta} = \|C\|_* \cdot \|x(0)\|_* \cdot b\big)$ and $\|\vec{y}_{\rm a}\|_p \leq \beta$, where

$$\beta := \begin{cases} \frac{\tilde{\beta}}{(1-r^p)^{1/p}} & \text{if } p \in [\,1, \infty\,[\\ \tilde{\beta} & \text{if } p = \infty\,. \end{cases} \tag{5.9}$$

The forced output $y_{\mathrm{f}}(t) = \sum_{\tau=0}^{t-1} CF^{t-\tau-1}Be(\tau)$ depends linearly on the input and can be rewritten as a convolution: $y_{\mathrm{f}}(t) = \sum_{\tau=0}^{t-1} g(t-\tau)e(\tau) := (\vec{g} * \vec{e})(t)$, where $\vec{g}$ is the *impulse response* of the system, namely

$$g(t) = \begin{cases} 0 & \text{if } t = 0 \\ CF^{t-1}B & \text{if } t \geq 1\,. \end{cases} \tag{5.10}$$

Let us denote by $\mathcal{G}$ the linear operator representing the dependence of the forced output from the input:

$$\begin{array}{rccl} \mathcal{G}\,: & \ell_p(\mathbb{R}^m) & \to & \ell_p(\mathbb{R}^q) \\ & \vec{e} & \mapsto & \vec{y}_{\mathrm{f}}(\vec{e}) = \vec{g} * \vec{e}\,. \end{array}$$

By definition,

$$\|\mathcal{G}\|_p := \sup_{\vec{e} \in \ell_p(\mathbb{R}^m)\setminus\{\vec{0}\}} \frac{\|\vec{y}_{\mathrm{f}}(\vec{e})\|_p}{\|\vec{e}\|_p}\,. \tag{5.11}$$

It holds that $\forall\, p \in [1, \infty]$, $\|\mathcal{G}\|_p < +\infty$ (i.e., $\mathcal{G}$ is continuous) and $\|\mathcal{G}\|_p$ is the ℓ_p–gain of system (5.7) (see Chapter 5 of [67] for more details). As the operator $\mathcal{G}$ acts between the input and the output signal spaces, then it is invariant under change of coordinates in the state space (and so it is its norm).
To sum up, consider system (5.7) and suppose that F is Schur. Then, for $x(0) \in \mathbb{R}^m$ and $p \in [1\,,\, \infty]$, it holds that

$$\forall\, \vec{e} \in \ell_p(\mathbb{R}^m)\,, \quad \|\vec{y}(\vec{e})\|_p \leq \|\mathcal{G}\|_p \|\vec{e}\|_p + \beta\,,$$

where the bias term β is provided in (5.9) and is the only quantity depending on the initial condition $x(0)$.
It is useful to fix some terminology:

Definition 23 *Let system* (5.7) *be given and assume that* F *is a Schur matrix. The linear operator* $\mathcal{G} : \ell_p(\mathbb{R}^m) \to \ell_p(\mathbb{R}^q)$ *is called the* input/output *operator associated to the system (with* p *to be specified from time to time according to the problem under consideration). Consider the auxiliary system* $\Sigma(F, B, I)$*: the input/output operator associated to this system is denoted by* $\mathcal{G}^{(I)}$ *and is called the* input/state *operator associated to system* (5.7)*; the corresponding impulse response is denoted by* $\vec{g}^{\,(I)}$.

For a linear system of the type in equation (5.7), it is possible to get rid of the dependence on the initial condition $x(0)$ by defining a notion of input/output stability that takes into account the forced output only. When the ℓ_∞–norm is considered we have the following

Definition 24 *Consider system* (5.7) *and let* $\vec{g} : \mathbb{N} \to \mathbb{R}^{q\times m}$ *be its impulse response. If the linear relation* $\vec{y}_{\mathrm{f}} = \vec{g} * \vec{e}$ *defines a bounded operator*

$$\ell_\infty(\mathbb{R}^m) \ni \vec{e} \overset{\mathcal{G}}{\longmapsto} \vec{y}_{\mathrm{f}} \in \ell_\infty(\mathbb{R}^q)\,,$$

then the system is said to be BIBO–stable.

With reference to the Kalman decomposition of a linear system (see, e.g., [114]), it is well known that the impulse response $\vec{g}$, and hence the input/output operator $\mathcal{G}$, only depends on the restriction of the system to the so called "reachable and observable" part.

Definition 25 *Given system* (5.7), *the eigenvalues of the reachable and observable part of the system are called the* poles *of the system.*
Let $\{\lambda_1, \ldots, \lambda_{n_p}\} \subset \mathbb{C}$ *be the set of the poles of system* (5.7) *and,* $\forall\, k = 1, \ldots, n_p$, *let* $m_k \in \mathbb{N}$ *be the algebraic multiplicity of the* k*–th pole. The polynomial* $d(z) := \Pi_{k=1}^{n_p}(z-\lambda_k)^{m_k}$ *is called the* polynomial of the poles *of the system.*

Lemma 14 *For a given system* (5.7), *the following properties are equivalent:*

1- *the system is BIBO–stable;*
2- *the poles of the system belong to the interior of the unit ball of* $\mathbb{C}$*;*
3- $\vec{g} \in \ell_1(\mathbb{R}^{q\times m})$.

In particular, if system (5.7) *is reachable and observable, then it is BIBO–stable if and only if* F *is a Schur matrix.*

Proof. See, e.g., [114]. ■

As a consequence of the fact that the notion of ℓ_∞–stability involves the full output signal $\vec{y} = \vec{y}_\mathrm{a} + \vec{y}_\mathrm{f}$, there are BIBO–stable systems that are not ℓ_∞–stable[1]. Hence, the study of $\|\mathcal{G}\|_\infty$ is meaningful in the more general framework of BIBO–stable systems. On the other hand, when dealing with ℓ_∞–stability we always assume that F is a Schur matrix: in this case we know by the above discussion that $\forall\, x(0) \in \mathbb{R}^n$, system (5.7) is ℓ_∞–stable (in particular, it is BIBO–stable) and the ℓ_∞–gain of the system coincides with $\|\mathcal{G}\|_\infty$.
The control problem which consists of finding a stabilizing controller for system (5.7), so that the ℓ_∞–gain of the closed loop dynamics is below a desired threshold, is referred to as the control problem in ℓ_1. This terminology is motivated by property 3 of Lemma 14 and by the fact that there is a relation between $\|\mathcal{G}\|_\infty$ and the ℓ_1–norm of $\vec{g}$ (this relation will be illustrated in Corollary 6 of Section 6.1.1).
In order to evaluate the ℓ_p–gain of a linear system, it is often useful to resort to its representation in the *frequency domain*. Hence, let us briefly recall the basic facts about the monolateral $\mathcal{Z}$–transformation of a signal.
Given a signal $\vec{v} : \mathbb{N} \to \mathbb{R}^{q\times m}$, the (monolateral) $\mathcal{Z}$–transformation of $\vec{v}$ is defined as the formal series

$$\mathcal{Z}[\vec{v}](z) := \sum_{t=0}^{+\infty} v(t) z^{-t} \tag{5.12}$$

and it is usually denoted by $V(z)$. The basic properties of the $\mathcal{Z}$–transformation are linearity $\big($i.e., $\mathcal{Z}[c_1\vec{v}_1 + c_2\vec{v}_2](z) = c_1\mathcal{Z}[\vec{v}_1](z) + c_2\mathcal{Z}[\vec{v}_2](z)\big)$, the functorial property with respect to

[1] It is sufficient to consider a BIBO–stable system such that the pair (F, C) is observable but there exists an eigenvalue of the non–reachable part of the system which is outside the unit ball: if $x(0)$ is an eigenvector corresponding to such an eigenvalue, then the autonomous output signal is divergent.

the convolution of signals $\big($i.e., $\mathcal{Z}[\vec{v}_1 * \vec{v}_2](z) = \mathcal{Z}[\vec{v}_1](z) \cdot \mathcal{Z}[\vec{v}_2](z)\big)$ and injectivity.
For a given linear system (5.7), let $\vec{g}$ be its impulse response and $G(z)$ be the $\mathcal{Z}$–transformation of $\vec{g}$. Because $\mathcal{Z}[\vec{y}_{\rm f}](z) = G(z) \cdot \mathcal{Z}[\vec{e}](z)$, $G(z)$ is called the *transfer matrix* of the system. When considered as a function of the complex variable z, it is well known that

$$G(z) = C(zI - F)^{-1}B\,.$$

In particular, $G(z)$ is a strictly proper rational matrix $\big($i.e., $\forall\, i = 1, \ldots, q$ and $\forall\, j = 1, \ldots, m$, $G_{i,j}(z) = n_{i,j}(z)/m_{i,j}(z)$, where $n_{i,j}(z)$ and $m_{i,j}(z)$ are polynomials such that $\deg(n_{i,j}) < \deg(m_{i,j})\big)$. Moreover, if F is Schur, then $G \in H_\infty$, where

$$H_\infty := \big\{V : \mathbb{C} \to \mathbb{C}^{q\times m} \,|\, \text{for } z \in \mathbb{C}\,,\ |z| > 1\,,\ V \text{ is analytical and bounded}\big\}\,.$$

H_∞ is a Banach space (called *Hardy* space) with the norm defined by

$$\|V\|_\infty := \sup_{z\in\mathbb{C}\,,\ |z|>1} \|V(z)\|_2\,.$$

By the "maximum modulus principle" [21], it follows that

$$\|G\|_\infty = \max_{\theta\in[0,2\pi]} \|G(e^{i\theta})\|_2\,.$$

Remark 17 *The norm $\|\mathcal{G}\|_\infty$ should not be confused with the norm $\|G\|_\infty$: the former is the ℓ_∞–gain of the system and it is concerned with the input/output operator defined on $\ell_\infty(\mathbb{R}^m)$ (hence, in the time domain); the latter is the H_∞–norm of the transfer matrix (hence, in the frequency domain) and, actually, it is the ℓ_2–gain of the system (see Proposition 13 in Section 5.3).*

Finally, let us briefly review some basic facts on the relation between state space and input/output descriptions of linear systems $\big($i.e., representation as in equation (5.7) or through the transfer matrix $G(z)\big)$. A more detailed and thorough presentation can be found in [114]. For any given strictly proper rational matrix $G(z)$, it is well known by realization theory that there exists a finite dimensional linear system of the type in equation (5.7) whose transfer matrix is $G(z)$. Given two linear systems $\Sigma(F_1, B_1, C_1)$ and $\Sigma(F_2, B_2, C_2)$ sharing the same transfer matrix $G(z)$, the reachable and observable part of the two systems coincide but for a linear change of coordinates. In view of these facts, the notion of pole of a system given in Definition 25 applies without ambiguity both to a linear system as in equation (5.7) or to a rational transfer matrix $G(z)$.

Let us introduce the nonlinear input/output relation that will be of primary importance for our study. A *static nonlinearity* is a non–dynamical input/output relation of the type $y = h(u)$ for some $h : \mathbb{R}^m \to \mathbb{R}^q$. We can associate to such a h an input/output operator as in equation (5.5), where $\vec{y}(\vec{u}) := \big\{h\big(u(t)\big)\big\}_{t\in\mathbb{N}}$, and to study the ℓ_p–stability of this operator according to Definition 22. With slight abuse of terminology, the ℓ_p–stability properties of this operator are referred to as the ℓ_p–stability properties of h. If $\exists\, \gamma_* \geq 0$ and $\beta_* \geq 0$ such

that $\forall\, u \in \mathbb{R}^m$, $\|h(u)\|_* \leq \gamma_* \|u\|_* + \beta_*$, then h is finite–gain ℓ_∞–stable and relation (5.6) is satisfied with $\gamma = \gamma_*$ and $\beta = \beta_*$. If $h(0) \neq 0$ then $\forall\, p \in [1\,,\, \infty[$, h is not ℓ_p–stable. If instead $\beta_* = 0$, then $\forall\, p \in [1\,,\, \infty]$ and $\forall\, \vec{u} \in \ell_p(\mathbb{R}^m)$, $\|\vec{y}(\vec{u})\|_p \leq \gamma_* \|\vec{u}\|_p$ (in particular, γ_* does not depend on p, it only depends on the vector norm $\|\cdot\|_*$). This motivates the following

Definition 26 *Let $p \in [1, \infty]$, a function $h : \mathbb{R}^m \to \mathbb{R}^q$ is said to be* unbiased *with* classical[2] *ℓ_p–gain γ iff $\forall\, u \in \mathbb{R}^m$, $\|h(u)\|_p \leq \gamma \|u\|_p$.*

We conclude this section with the statement of one classical version of the small–gain theorem for the ℓ_p–stability analysis of discrete–time linear systems under nonlinear static output feedback:

Proposition 12 (Small–gain for $\boldsymbol{\ell_p}$–stability) *Consider the linear system (5.7) and assume that F is a Schur matrix. Consider on $\mathbb{R}^m$ and $\mathbb{R}^q$ the norm $\|\cdot\|_{\hat{p}}$, for some $\hat{p} \in [1\,,\, \infty]$, and assume that the vector norm $\|\cdot\|_*$ appearing in equation (5.2) to define the ℓ_p–norm of the input and output signals $\vec{e}$ and $\vec{y}$ is $\|\cdot\|_{\hat{p}}$. Denote by γ_s the corresponding ℓ_p–gain of the system. Let $\varphi : \mathbb{R}^q \to \mathbb{R}^m$ be a static nonlinearity which is unbiased with classical $\ell_{\hat{p}}$–gain γ. If $\gamma_s \cdot \gamma < 1$, then the system*

$$\begin{cases} x(t+1) = Fx(t) + B\varphi\big(Cx(t)\big) + Be \\ y(t) = Cx(t) \end{cases}$$

is ℓ_p–stable.

Proof. See, e.g., [69]. It is a consequence of the fact that the small–gain condition implies that the map $\vec{\varphi} \circ \mathcal{G} : \ell_p(\mathbb{R}^m) \to \ell_p(\mathbb{R}^m)$, where $\vec{\varphi}(\vec{y}) := \big\{\varphi\big(y(t)\big)\big\}_{t \in \mathbb{N}}$, is a contraction with $\vec{0}$ being its unique fixed point. ■

In this thesis we deal with generalized versions of this proposition suitable to study practical stability, hence including also the analysis of the internal dynamics. We consider the two cases $p = 2$ (in Section 5.3) and $p = \infty$ (in Section 6.2). In both cases, a reference to the computation (or the estimation) of $\|\mathcal{G}\|_p$ is included.

5.3 ℓ_2/ℓ_2 small–gain for practical stability of multi–input systems

Before stating the generalized small–gain theorem for practical stability, let us recall the basic facts on the ℓ_2–gain of a linear system and let us introduce a generalized notion of gain for a static nonlinearity.

[2] In opposition to the *generalized* notions of gain that are defined later on.

Proposition 13 (ℓ_2–gain and bounded real lemma) *Consider the linear system* (5.7) *and assume that* F *is a Schur matrix. Then the* ℓ_2*–gain of the system is equal to* $\|G\|_\infty$. *For* $\gamma > 0$, $\|G\|_\infty < \gamma$ *if and only if there exists a unique* $P \geq 0$ *such that*

$$\begin{cases} P = F'PF + F'PB(\gamma^2 I - B'PB)^{-1}B'PF + C'C \\ \gamma^2 I - B'PB > 0 \\ F + B(\gamma^2 I - B'PB)^{-1}B'PF \quad \textit{is Schur}. \end{cases} \tag{5.13}$$

Moreover, if the pair (F, C) *is observable, then* $P > 0$.

Proof. See [33] and [129]. ■

Motivated by the discussion in Section 5.1, let us define a generalized notion of gain in ℓ_2.

Definition 27 *Let* $\varrho_0 > 0$ *and* $\gamma_e \geq 0$. *A map*

$$\begin{array}{rccc} \varphi : & \mathbb{R}^p & \to & \mathbb{R}^m \\ & y & \mapsto & \varphi(y) \end{array}$$

is said to have ϱ_0–external gain γ_e *iff* $\|y\|_2 > \varrho_0 \;\Rightarrow\; \|\varphi(y)\|_2 \leq \gamma_e \|y\|_2$.

For a fixed value of $\varrho_0 > 0$, if $\tilde{\gamma}_e$ is a ϱ_0–external gain of φ, then also any $\gamma_e > \tilde{\gamma}_e$ is a ϱ_0–external gain of φ. The smallest feasible value for γ_e to be a ϱ_0–external gain is

$$\underline{\gamma_e}(\varrho_0) = \sup_{\|y\|_2 > \varrho_0} \frac{\|\varphi(y)\|_2}{\|y\|_2}.$$

Sometimes we will refer to *the* ϱ_0–external gain of φ: with this terminology we mean $\underline{\gamma_e}(\varrho_0)$. As $\underline{\gamma_e}$ is a non increasing function, if $\underline{\gamma_e}(\varrho_0) < +\infty$, a smaller value for the external gain may be obtained by increasing ϱ_0. However, there are important cases where such a decreasing property does not hold (we will see examples later on in the class of logarithmic quantizers), we hence give the following

Definition 28 *A map* $\varphi : \mathbb{R}^p \to \mathbb{R}^m$ *is said to be* standard *with* natural *external gain* γ_e *iff* $\exists\, \bar{\varrho}_0 > 0$ *such that*

$$\forall\, \varrho_0 > \bar{\varrho}_0, \qquad \sup_{\|y\|_2 > \varrho_0} \frac{\|\varphi(y)\|_2}{\|y\|_2} = \gamma_e.$$

Notice that, although a natural value of the external gain is associated to a standard non linearity φ, decreasing ϱ_0 below a certain threshold makes in general the ϱ_0–external gain increase (see Examples 19 and 21 in Section 5.3.2).

5.3.1 Practical stability analysis in H_∞

The main result in this section provides a sufficient condition for practical stability in terms of a generalized small–gain theorem, it also includes a quantitative analysis of practical stability.

Theorem 9 (Small–gain in H_∞: (X_0, Ω)–stability analysis) *Let us consider a linear system*

$$\begin{cases} x(t+1) = Fx(t) + Be(t) \\ y(t) = Cx(t) \\ x \in \mathbb{R}^n,\ e \in \mathbb{R}^m,\ y \in \mathbb{R}^p, \end{cases} \tag{5.14}$$

where F is a Schur matrix. Let $G(z)$ be the transfer matrix of the system and $\gamma_s := \|G\|_\infty$ be the ℓ_2–gain of the system. For a given $q_e : \mathbb{R}^p \to \mathbb{R}^m$, consider the control law

$$e(t) = q_e\big(y(t)\big) \ :$$

the corresponding closed–loop dynamics is

$$x(t+1) = Fx(t) + Bq_e\big(Cx(t)\big)\,. \tag{5.15}$$

Suppose that $\exists\, \varrho_0 > 0$, $\gamma_e \geq 0$ and $E_0 \geq 0$ such that the following conditions hold:

a) *q_e has ϱ_0–external gain γ_e;*
b) *if $\|y\|_2 \leq \varrho_0$, then $\|q_e(y)\|_2 \leq E_0$;*
c) *$\gamma_s \cdot \gamma_e < 1$.*

Then a matrix $\mathbb{R}^{n\times n} \ni P > 0$ and a constant $r_{\rm i}^2 > 0$ can be explicitly determined such that the following properties hold for system (5.15):

ι) $\forall\, r^2 \geq r_{\rm i}^2$, the ellipsoid $\mathcal{E}_{P,r^2}$ is positively invariant;
ιι) $\forall\, r_1^2 \geq r_2^2 > r_{\rm i}^2$ the system is $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\big)$–stable.

Proof. Let $\gamma > \gamma_s$ be such that $\gamma\cdot\gamma_e < 1$. For $\mathbb{R}^{n\times n} \ni Q > 0$, let $G_Q(z) := Q^{\frac{1}{2}}(zI - F)^{-1}B$. Consider $Q > 0$ such that

$$\gamma_s + \|G_Q\|_\infty < \gamma \tag{5.16}$$

$\big($this can be achieved, for instance, with $Q = \lambda I$ and sufficiently small $\lambda > 0\big)$. With

$$\widetilde{C} := \begin{pmatrix} C \\ Q^{\frac{1}{2}} \end{pmatrix},$$

let

$$\widetilde{G}(z) := \widetilde{C}(zI - F)^{-1}B = \begin{pmatrix} G(z) \\ G_Q(z) \end{pmatrix}.$$

Since $\|\widetilde{G}\|_\infty \leq \|G\|_\infty + \|G_Q\|_\infty = \gamma_s + \|G_Q\|_\infty$, then $\|\widetilde{G}\|_\infty < \gamma$. By Proposition 13, as F is Schur and $(F, \widetilde{C})$ is observable, $\exists\, P > 0$ such that system (5.13) is satisfied. In particular, such a P satisfies the discrete–time algebraic Riccati equation

$$P = F'PF + F'PB(\gamma^2 I - B'PB)^{-1}B'PF + C'C + Q\,. \tag{5.17}$$

Let $V(x) := x'Px$ and $\Delta V(x) := V(x^+) - V(x)$ (the dependence on t is omitted), then

$$\begin{aligned}\Delta V(x) = \;& 2x'F'PBe + e'B'PBe - x'F'PB(\gamma^2 I - B'PB)^{-1}B'PFx - y'y - x'Qx = \\ = \;& 2x'F'PBe - e'(\gamma^2 I - B'PB)e + \gamma^2 e'e - y'y - x'Qx - \\ & -x'F'PB(\gamma^2 I - B'PB)^{-1}B'PFx = \\ = \;& \gamma^2 e'e - y'y - x'Qx - (e - e^*)'(\gamma^2 I - B'PB)(e - e^*)\,,\end{aligned}$$

where

$$e^* = (\gamma^2 I - B'PB)^{-1}B'PFx\,.$$

Hence,

$$\Delta V(x) \le \gamma^2 e'e - y'y = \gamma^2\|e\|_2^2 - \|y\|_2^2\,.$$

Let x be such that $y = Cx$ satisfies $\|y\|_2 > \varrho_0$: by the hypothesis a, $\|e\|_2^2 \le \gamma_e^2\|y\|_2^2$, therefore

$$\Delta V(x) \le \|y\|_2^2\big((\gamma\cdot\gamma_e)^2 - 1\big) < 0\,. \tag{5.18}$$

Assume instead that $\|y\|_2 \le \varrho_0$. With[3]

$$S := F'PB(\gamma^2 I - B'PB)^{-1}B'PF + C'C + Q > 0\,, \tag{5.19}$$

P satisfies the Lyapunov equation

$$F'PF - P = -S \tag{5.20}$$

(see equation (5.17)). By the hypothesis b, $\|e\|_2 \le E_0$, we can hence follow arguments that are similar to those we used to prove Proposition 9. Indeed, $\Delta V(x)$ can be written as

$$\Delta V(x) = -x'Sx + 2x'F'PBe + e'B'PBe$$

so that

$$\Delta V(x) \le -\lambda_{\min}(S)\|x\|_2^2 + 2E_0\|F'PB\|_2\cdot\|x\|_2 + E_0^2\,\|B'PB\|_2 := f(\|x\|_2)\,. \tag{5.21}$$

Thus, if $\|x\|_2 > R$, where R is defined as in equation (3.51) by

$$\begin{cases} R = \frac{E_0}{\lambda_{\min}(S)}\,\alpha(P)\,, \quad \text{where} \\ \alpha(P) = \|F'PB\|_2 + \sqrt{\|F'PB\|_2^2 + \lambda_{\min}(S)\|B'PB\|_2}\,, \end{cases}$$

then $\Delta V(x) < 0$. Therefore, the region where $\Delta V(x)$ may be positive is $\mathcal{D}_0 := \mathcal{B}_R \cap \{x \,|\, \|Cx\|_2 \le \varrho_0\}$. In order to analyze the behavior of the trajectories starting from $\mathcal{D}_0$, $\forall\,\epsilon \ge 0$ let $\mathcal{D}_\epsilon := \mathcal{B}_{R+\epsilon} \cap \{x \,|\, \|Cx\|_2 \le \varrho_0\}$ and

$$M_\epsilon^2 := \max_{x\in\mathcal{D}_\epsilon} V(x^+)\,.$$

[3] At this stage, the matrix $Q > 0$ is needed to guarantee that $S > 0$. We will be back on the importance of calling Q into the question later on.

It holds that $\forall\, r^2 \geq M_0^2$, the ellipsoid $\mathcal{E}_{P,r^2}$ is invariant. In fact,

$$\mathcal{E}_{P,r^2} = (\mathcal{E}_{P,r^2} \cap \mathcal{D}_0) \cup (\mathcal{E}_{P,r^2} \setminus \mathcal{D}_0)$$

and $x^+ \in \mathcal{E}_{P,r^2}$ if and only if $V(x^+) \leq r^2$: if $x \in \mathcal{E}_{P,r^2} \cap \mathcal{D}_0$, then $V(x^+) \leq M_0^2 \leq r^2$; if instead $x \in \mathcal{E}_{P,r^2} \setminus \mathcal{D}_0$, then $V(x^+) < V(x) \leq r^2$.
Let us provide an upper bound for M_ϵ^2. As in the proof of Proposition 9, for $x \in \mathcal{D}_\epsilon$,

$$\begin{aligned} V(x^+) = &\ \Delta V(x) + V(x) = \\ = &\ x'(P-S)x + 2x'F'PBe + e'B'PBe \leq \\ \leq &\ \lambda_{\max}(P-S)\|x\|_2^2 + 2E_0\|F'PB\|_2 \cdot \|x\|_2 + E_0^2\|B'PB\|_2 := g(\|x\|_2)\,. \end{aligned}$$

By equation (5.20), $\lambda_{\max}(P-S) = \lambda_{\max}(F'PF) \geq 0$, then $\max\limits_{x\in\mathcal{B}_{R+\epsilon}} g(\|x\|_2) = g(R+\epsilon)$.
Therefore[4],

$$\max_{x\in\mathcal{D}_\epsilon} g(\|x\|_2) \leq g(R+\epsilon)$$

and

$$\forall\, \epsilon \geq 0\,,\ \ M_\epsilon^2 \leq g(R+\epsilon)\,. \tag{5.22}$$

In particular $\big($following the same arguments that led to equation (3.52)$\big)$,

$$M_0^2 \leq g(R) = R^2\big(\lambda_{\max}(P-S) + \lambda_{\min}(S)\big)\,.$$

Hence, the proof of part $\imath$ is achieved with

$$\begin{aligned} &r_{\mathrm{i}}^2 := R^2\big(\lambda_{\max}(P-S) + \lambda_{\min}(S)\big)\,, \quad \text{where} \\ &\begin{cases} R = \frac{E_0}{\lambda_{\min}(S)}\,\alpha(P)\,, \\ \alpha(P) = \|F'PB\|_2 + \sqrt{\|F'PB\|_2^2 + \lambda_{\min}(S)\|B'PB\|_2}\,, \\ P \text{ is the solution of equation (5.17) and} \\ S = F'PB(\gamma^2 I - B'PB)^{-1}B'PF + C'C + Q\,. \end{cases} \end{aligned} \tag{5.23}$$

To prove part $\imath\imath$ let us first show that the following claim holds true: $\forall\, \epsilon > 0$ and $\forall\, x(0) \in \mathbb{R}^n \setminus \mathcal{D}_\epsilon$, $\exists\, t > 0$ such that $x(t) \in \mathcal{D}_\epsilon$. In fact, with $\mathcal{C}_1 := \{x \,|\, \|Cx\|_2 > \varrho_0\}$ and $\mathcal{C}_2 := \{x \,|\, \|Cx\|_2 \leq \varrho_0 \text{ and } \|x\|_2 > R+\epsilon\}$, it holds that $\mathbb{R}^n \setminus \mathcal{D}_\epsilon = \mathcal{C}_1 \cup \mathcal{C}_2$. By equation (5.18), if $x \in \mathcal{C}_1$ then $\Delta V(x) < \varrho_0^2\big((\gamma\cdot\gamma_e)^2 - 1\big) < 0$; if instead $x \in \mathcal{C}_2$, then, by equation (5.21), $\Delta V(x) \leq f(R+\epsilon) < 0$. Hence, with $\delta := \max\big\{f(R+\epsilon)\,;\, \varrho_0^2\big((\gamma\cdot\gamma_e)^2 - 1\big)\big\} < 0$, it holds that

$$\forall\, x \in \mathbb{R}^n \setminus \mathcal{D}_\epsilon\,,\ \Delta V(x) \leq \delta < 0\ :$$

the convergence in finite time to $\mathcal{D}_\epsilon$ easily follows because V is a positive definite quadratic form.

[4]Actually, if $C \in \mathbb{R}^{p\times n}$ with $p < n$, then $\max\limits_{x\in\mathcal{D}_\epsilon} g(\|x\|_2) = g(R+\epsilon)$. In fact, in this case $\{x \,|\, \|Cx\|_2 \leq \varrho_0\}$ is unbounded and hence $\exists\, x \in \mathcal{D}_\epsilon$ such that $\|x\|_2 = R+\epsilon$.

Given $r_1^2 \geq r_2^2 > r_{\mathrm{i}}^2$, as the invariance of both $\mathcal{E}_{P,r_1^2}$ and $\mathcal{E}_{P,r_2^2}$ has already been verified, to prove $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\right)$–stability we have only to show that $\forall\, x(0) \in \mathcal{E}_{P,r_1^2}$, $\exists\, t > 0$ such that $x(t) \in \mathcal{E}_{P,r_2^2}$. Since g is a continuous function and $g(R) = r_{\mathrm{i}}^2 < r_2^2$, $\exists\, \epsilon > 0$ such that $g(R+\epsilon) \leq r_2^2$. By the claim, $\forall\, x(0) \in \mathcal{E}_{P,r_1^2}$, $\exists\, t \geq 0$ such that $x(t) \in \mathcal{D}_\epsilon$, therefore, by definition of M_ϵ^2, it holds that $x(t+1) \in \mathcal{E}_{P,M_\epsilon^2}$. The thesis follows because $\mathcal{E}_{P,M_\epsilon^2} \subseteq \mathcal{E}_{P,r_2^2}$: in fact, by equation (5.22), $M_\epsilon^2 \leq g(R+\epsilon) \leq r_2^2$. ■

A classical version of the small–gain theorem is a particular case of Theorem 9, indeed:

Corollary 5 (Classical small–gain in H_∞ for asymptotic stability) *If $\forall\, y \in \mathbb{R}^p$, $\|q_e(y)\|_2 \leq \gamma_e \|y\|_2$ and $\gamma_s \cdot \gamma_e < 1$, then $x = 0$ is a globally asymptotically stable equilibrium for system (5.15).*

Proof. By the assumption on q_e, $\forall\, \varrho_0 > 0$, q_e has ϱ_0–external gain γ_e and $\|y\|_2 \leq \varrho_0 \Rightarrow \|q_e(y)\|_2 \leq \gamma_e \cdot \varrho_0 := E_0(\varrho_0)$. In particular, $q_e(0) = 0$ so that 0 is an equilibrium. In the proof of Theorem 9 we have shown that

$$\forall\, x \in \mathbb{R}^n \setminus \mathcal{B}_R\,, \quad \Delta V(x) < 0\,,$$

where

$$R = E_0(\varrho_0) \frac{\alpha(P)}{\lambda_{\min}(S)}\,.$$

Since $\lim_{\varrho_0 \to 0} E_0(\varrho_0) = 0$ and ϱ_0 can be chosen arbitrarily small, then $\forall\, x \in \mathbb{R}^n \setminus \{0\}$, $\Delta V(x) < 0$. ■

If q_e is bounded in norm, the practical stability analysis becomes simpler as there is no need to resort to small–gain arguments.

Proposition 14 (Uniformly bounded q_e) *Assume that $\forall\, y \in \mathbb{R}^p$, $\|q_e(y)\|_2 \leq E_0$. For any $\mathbb{R}^{n\times n} \ni S > 0$, let P be the solution of the Lyapunov equation*

$$F'PF - P = -S$$

and

$$\begin{cases} r_{\mathrm{i}}^2 := R^2\big(\lambda_{\max}(P-S) + \lambda_{\min}(S)\big)\,, \ \textit{where} \\ R = \frac{E_0}{\lambda_{\min}(S)}\,\alpha(P)\,, \ \textit{and} \\ \alpha(P) = \|F'PB\|_2 + \sqrt{\|F'PB\|_2^2 + \lambda_{\min}(S)\|B'PB\|_2}\,. \end{cases}$$

Then the following properties hold for system (5.15):

ι) $\forall\, r^2 \geq r_{\mathrm{i}}^2$, the ellipsoid $\mathcal{E}_{P,r^2}$ is positively invariant;
ιι) $\forall\, r_1^2 \geq r_2^2 > r_{\mathrm{i}}^2$ the system is $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\right)$–stable.

Proof. Part *ι* is a direct consequence of Proposition 9.
The proof of part *ιι* can be obtained with slight modifications of the proof of Theorem 9.*ιι*. Specifically, it is sufficient to replace $\mathcal{D}_\epsilon$ with $\mathcal{B}_{R+\epsilon}$ (and, accordingly, to modify the definition of M_ϵ). Correspondingly, one has to change the statement of the claim needed to prove convergence as follows: $\forall\, \epsilon > 0$ and $\forall\, x(0) \in \mathbb{R}^n \setminus \mathcal{B}_{R+\epsilon}$, $\exists\, t > 0$ such that $x(t) \in \mathcal{B}_{R+\epsilon}$. The proof of this fact is trivial because $\forall\, x(0) \in \mathbb{R}^n \setminus \mathcal{B}_{R+\epsilon}$, $\Delta V(x) \leq f(R+\epsilon) < 0$. ■

5.3.2 Practical stabilization of quantized input systems via $\boldsymbol{H_\infty}$–control

Let us show how Theorem 9 and Proposition 14 can be used to synthesize practically stabilizing controllers and to analyze the resulting closed loop dynamics for quantized input linear systems. Consider a system

$$\begin{cases} x^+ = Ax + Bu \\ x \in \mathbb{R}^n,\ u \in \mathcal{U} \subset \mathbb{R}^m, \end{cases} \tag{5.24}$$

where the pair (A, B) is supposed to be stabilizable. For the moment, the quantized control set $\mathcal{U}$ is assumed to be given, nevertheless, most of the theory we are going to present can be easily adjusted on the case where also $\mathcal{U}$ can be chosen (see also Remark 24 at the end of this section and Example 24 in Section 5.3.3).
The goal is to design a constant feedback matrix $K \in \mathbb{R}^{m\times n}$ and an input quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ so that the control law

$$u(x) = q_{\mathcal{U}}(Kx)$$

practically stabilizes system (5.24).
With the quantization error $q_e : \mathbb{R}^m \to \mathbb{R}^m$ defined by $q_e(y) = q_{\mathcal{U}}(y) - y$ (see Definition 4 in Section 2.1), the closed–loop dynamics induced by $u(x)$ is

$$x^+ = (A + BK)x + Bq_e(Kx). \tag{5.25}$$

First notice that, if K is such that $A + BK$ is Schur, then we are in the right framework to apply Theorem 9 or Proposition 14: it is sufficient to let $F := A + BK$ and $C := K$.
Let us begin by considering the simple case in which the quantization error is uniformly bounded. The typical example where this property occurs is provided by systems under uniform input quantization (see Definition 7 in Section 2.1). In this case, the suitable tool to face the problem is Proposition 14. All the details can be found in the discussion and in the examples presented in Section 3.1.4. The result presented in that section must be simply combined with the convergence property stated in Proposition 14.*ii*.
Let us consider now the general case. Suppose that q_e is such that $\exists\, \varrho_0 > 0$, $\gamma_e \geq 0$ and $E_0 \geq 0$ as in the hypotheses a–b of Theorem 9 (we will be back on this later on). If $K \in \mathbb{R}^{m\times n}$ is such that also hypothesis c is satisfied, then Theorem 9 guarantees the practical stability of the closed loop system (5.25) and provides the final invariant ellipsoid $\mathcal{E}_{P,r_i^2}$ to which convergence can be ensured[5]. Therefore, the problem is now reduced to find such a K. In other words, with $G_K(z) := K(zI - A - BK)^{-1}B$, we have to solve the following problem: given $\gamma_\infty \leq \frac{1}{\gamma_e}$, find $K \in \mathbb{R}^{m\times n}$ such that

$$\begin{cases} A + BK \quad \text{is Schur} & \text{(5.26a)} \\ \|G_K\|_\infty < \gamma_\infty\,. & \text{(5.26b)} \end{cases}$$

The solution of the practical stabilization problem and the analysis of the resulting closed loop dynamics can be summarized in the following theoretical procedure:

[5]To be really precise, we have proved the convergence to any ellipsoid of the type $\mathcal{E}_{P,r_i^2+\epsilon}$, $\forall\, \epsilon > 0$.

Procedure 1 (Practical stabilization and closed loop analysis) *Given system* (5.24) *, do:*

1. Input quantization: *fix an input quantizer* $q_{\mathcal{U}}$ *and analyze the corresponding quantization error* q_e *by providing* γ_e *and* E_0 *;*
2. Control synthesis: *let* $\gamma_\infty \leq 1/\gamma_e$ *and find* $K \in \mathbb{R}^{m\times n}$ *that solves problem* (5.26) *;*
3. Closed loop analysis: *apply Theorem 9 with* $F := A + BK$ *and* $C := K$ *.*
   (a) *Consider the Riccati equation* (5.17) *. Choice of the parameters* γ *and* Q *: with* $\gamma_s = \|G_K\|_\infty$ *, let* γ *be such that* $\gamma_s < \gamma < \frac{1}{\gamma_e}$ *and fix* $\mathbb{R}^{n\times n} \ni Q > 0$ *such that condition* (5.16) *is satisfied. Find* P *that solves equation* (5.17) *;*
   (b) *Compute* r_{i}^2 $\big($*which is depending on* $E_0$$\big)$ *according to equation* (5.23) *;*
4. Final result: $\forall\, r_1^2 \geq r_2^2 > r_{\mathrm{i}}^2$ *, system* (5.24) *controlled with* $u(x) = q_{\mathcal{U}}\big(Kx\big)$ *is* $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\big)$*–stable.*

In order to implement such a procedure, a deeper analysis of steps 1 , 2 and 3a is needed. Step 1 is essentially a geometry issue whose study is postponed to the end of the discussion of the other two steps.

• ***Control synthesis: implementation of step 2***

Problem (5.26) is a standard control problem in H_∞ and a wide literature is available for that (see, inter alia, [59, 116, 33]) . Actually, the formulation in (5.26) is a particular case of the so called *state feedback* H_∞ control problem known as the "actuator disturbance" case (see, e.g., [13]) . Namely, the noise term directly affects the input of the system or, with an equivalent terminology, there is an *input–matched* disturbance term. We recall here a classical result which is the particularization to our case of one of the several solutions for the general state feedback problem.

Definition 29 *A matrix* A *is said to be* unmixed *iff all its eigenvalues* $\lambda(A)$ *are such that* $|\lambda(A)| \neq 1$ *.*

Lemma 15 (The actuator disturbance case) *Consider a discrete–time system*

$$\begin{cases} x^+ = Ax + B(u + e) \\ y = u \\ x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \end{cases}$$

and assume that the matrix A *is unmixed. With* $u = Kx$ *the dynamics gets*

$$\begin{cases} x^+ = (A + BK)x + Be \\ y = Kx\,. \end{cases}$$

Let $G_K(z) = K(zI - A - BK)^{-1}B$ *be the corresponding transfer matrix. There exists* $K \in \mathbb{R}^{m\times n}$ *such that* $A + BK$ *is Schur and* $\|G_K\|_\infty < \gamma_\infty$ *if and only if there exists* $\mathbb{R}^{n\times n} \ni P^* \geq 0$ *such that the following conditions hold:*

$$\begin{cases} P^* = A'\Big(P^* - \frac{\gamma_\infty^2-1}{\gamma_\infty^2}P^*B\big(I + \frac{\gamma_\infty^2-1}{\gamma_\infty^2}B'P^*B\big)^{-1}B'P^*\Big)A & \text{(5.27a)} \\ \Big(I - BB'\big(I + \frac{\gamma_\infty^2-1}{\gamma_\infty^2}P^*BB'\big)^{-1}P^*\Big)A \quad \textit{is Schur} & \text{(5.27b)} \\ \gamma_\infty^2 I - B'P^*B > 0\,. & \text{(5.27c)} \end{cases}$$

In this case, a feasible choice for K *is the central* H_∞ *controller, namely*

$$K_{\rm c}(\gamma_\infty) := -B'\big(I + \tfrac{\gamma_\infty^2-1}{\gamma_\infty^2}P^*BB'\big)^{-1}P^*A\,. \tag{5.28}$$

Proof. See e.g., [33, 117, 24]. Here we limit ourselves to notice that condition (5.27b) is tantamount to requiring that $A + BK_{\rm c}(\gamma_\infty)$ is Schur. ■

If A is Schur the problem is trivial because $K = 0$ ensures $\|G_K\|_\infty = 0$. Hence, unless otherwise stated, from now on we always assume that the matrix A is not Schur.
Let us suppose that A is unmixed: in order that conditions (5.27) be satisfied, it is necessary that $\gamma_\infty > 1$: in fact, for $\gamma_\infty = 1$, equation (5.27a) becomes $P^* = A'P^*A$ and, because A is unmixed, the only positive semi–definite solution is $P^* = 0$. For $\gamma_\infty = 1$ and $P^* = 0$, condition (5.27b) becomes A is Schur[6]. For $\gamma_\infty > 1$, $\exists\, P^* \geq 0$ such that conditions (5.27a) and (5.27b) hold if and only if the pair (A, B) is stabilizable and A is unmixed (see [117]). However, condition (5.27c) might not be satisfied for some $\gamma_\infty > 1$. Hence, the problem raises of the computation of

$$\gamma_{\rm inf} := \inf\{\gamma_\infty > 1 \,|\, \exists\, P^* \geq 0 \text{ such that conditions (5.27) hold}\}\,.$$

For single–input systems, it has been proved in [51] that

$$\gamma_{\rm inf} = \Pi_{\lambda_{\rm u}(A)\in\mathcal{S}_{\rm u}(A)}|\lambda_{\rm u}(A)|\,, \tag{5.29}$$

where $\mathcal{S}_{\rm u}(A) := \big\{\lambda(A)\,|\,|\lambda(A)| > 1\big\}$. In other words, the supremum of the feasible gains for the quantization error, that is $\frac{1}{\gamma_{\rm inf}}$, only depends on the unstable poles of the open loop system. As expected, the more unstable the system is, namely the larger $\Pi_{\lambda_{\rm u}(A)\in\mathcal{S}_{\rm u}(A)}|\lambda_{\rm u}(A)|$ is, the smaller the allowed values for γ_e are. Indeed, smaller values of γ_e correspond to input sets more densely quantized.
For multi–input systems it has been proved in [13] that

$$1 \leq \gamma_{\rm inf} \leq 1 + \sqrt{1 + \lambda_\infty} := \gamma_{A,B}\,, \tag{5.30}$$

[6] Actually, the property that if A is not Schur, then $\|G_K\|_\infty \geq 1$ holds true also if A is not unmixed (see Lemma 16).

where $\lambda_\infty = \|B'P_\infty B\|_2$ and P_∞ is the stabilizing solution of the standard Riccati equation

$$P = A'\big(P - PB\big(I + B'PB\big)^{-1}B'P\big)A\,.$$

It will be shown in Example 23 in Section 5.3.3 that, for multi–input systems, $\gamma_{\inf}$ can be strictly smaller than $\Pi_{\lambda_{\rm u}(A)\in\mathcal{S}_{\rm u}(A)}|\lambda_{\rm u}(A)|$. That is, to obtain practical stability properties for a multi–input system, a lower input quantization density is tolerated (i.e., a larger external gain γ_e is allowed) than for a single–input system with the same amount of open loop instability[7].

To sum up, for systems $\Sigma(A,B,\mathcal{U})$ with A unmixed, Lemma 15 provides the necessary and sufficient condition for the existence of a solution K to the H_∞ control problem (5.26). An explicit solution is the central H_∞ controller $K = K_{\rm c}(\gamma_\infty)$. For single–input systems, problem (5.26) is feasible if and only if

$$\gamma_\infty \in \Big]\, \Pi_{\lambda_{\rm u}(A)\in\mathcal{S}_{\rm u}(A)}|\lambda_{\rm u}(A)|\,,\, \frac{1}{\gamma_e}\Big]\,. \tag{5.31}$$

For multi–input systems, by inequalities (5.30), a sufficient condition for the feasibility of problem (5.26) is

$$\gamma_\infty \in \Big]\, \gamma_{A,B}\,,\, \frac{1}{\gamma_e}\Big]\,,$$

a necessary condition is $\gamma_\infty \in \big]1\,,\,\frac{1}{\gamma_e}\big]$ or, equivalently, $\gamma_e < 1$.

Remark 18 *Expression* (5.29) *had already been found in [39] with reference to the search of the coarsest quantizer guaranteeing asymptotic stability. Indeed, in that paper a method equivalent to the approach presented in [51], but formally different, was used. The same expression also appears in the stabilizability condition under bit–rate constraint presented, for instance, in [89, 126, 80]* (*see equation* (1.2) *in Section 1.2; the continuous time version of this condition is the one in equation* (3.40) *of Section 3.1.3, see also [5]*).

Remark 19 (Recalibration of the Riccati equation) *When problem* (5.26) *is solved by means of system* (5.27), *not only one obtains* $\|G_{K(\gamma_\infty)}\|_\infty < \gamma_\infty$, *but in some cases such a norm can be significantly smaller than* γ_∞. *This phenomenon may be undesirable because the reduction of the* H_∞*–norm of the system can result in the increase of the* H_2*–norm. Therefore, one may be interested in using Lemma 15 with the purpose of obtaining a closed loop system with a* H_∞*–norm approximately equal to a specified value* $\bar\gamma$. *In this case, one has to solve system* (5.27) *with some proper* $\gamma_\infty > \bar\gamma$. *The determination of such a* γ_∞ *is referred to in the literature as the recalibration of the Riccati equation. A thorough reference for this problem is [13].*

Remark 20 (Algorithmic computation of $\gamma_{\inf}$) *If* A *is unmixed and not Schur, thanks to Lemma 15, an easy algorithm for the computation of* $\gamma_{\inf}$ *can be carried out:*

[7] Notice that we are only talking about the dispersion of the control values, we are not comparing the number of control values needed to obtain a desired practical stability objective.

- *Let* $\gamma_0^{(0)} := 1$ *and* $\gamma_1^{(0)} := \gamma_{A,B} + \epsilon$ *(for any* $\epsilon > 0$*);*
- *Assume that* $\gamma_0^{(h)}$ *and* $\gamma_1^{(h)}$ *are given, let us construct* $\gamma_0^{(h+1)}$ *and* $\gamma_1^{(h+1)}$*.*
  Let $\gamma^{(h+1)} := \frac{\gamma_0^{(h)} + \gamma_1^{(h)}}{2}$ *and try to solve system* (5.27) *with* $\gamma_\infty = \gamma^{(h+1)}$ *:*
  - *If system* (5.27) *is not feasible, let* $\gamma_0^{(h+1)} := \gamma^{(h+1)}$ *and* $\gamma_1^{(h+1)} := \gamma_1^{(h)}$*;*
  - *If system* (5.27) *is feasible, let* $\gamma_0^{(h+1)} := \gamma_0^{(h)}$ *and* $\gamma_1^{(h+1)} := \gamma^{(h+1)}$*.*

It holds that, $\forall\, h \in \mathbb{N}$*,* $\gamma_0^{(h)} \leq \gamma_{\inf} < \gamma_1^{(h)}$*: this follows by Lemma 15 because,* $\forall\, h \in \mathbb{N}$*, system* (5.27) *is not feasible for* $\gamma_\infty = \gamma_0^{(h)}$ *and it is feasible for* $\gamma_\infty = \gamma_1^{(h)}$*. This property holds for* $h = 0$ *thanks to inequality* (5.30)*, and for* $h \geq 1$ *by construction. Thus,*

$$\lim_{h \to +\infty} \gamma_1^{(h)} = \gamma_{\inf}\,.$$

Using equation (5.28)*, the algorithm can also incorporate a sequence* $K^{(h)} := K_{\mathrm{c}}\big(\gamma_1^{(h)}\big)$ *of stabilizing control gains such that*

$$\forall\, h \in \mathbb{N}\,, \quad \gamma_{\inf} \leq \|G_{K^{(h)}}\|_\infty < \gamma_1^{(h)}\,.$$

Moreover, we can take advantage of the fact that $\|G_{K^{(h)}}\|_\infty < \gamma_1^{(h)}$ *to reduce the number of iterations needed to reach a suitable approximation of* $\gamma_{\inf}$*. To this end, in case that system* (5.27) *is feasible, the updating rule of* $\gamma_1^{(h)}$ *can be modified as follows:*

- *if system* (5.27) *is feasible, let* $\gamma_0^{(h+1)} := \gamma_0^{(h)}$ *and, with* $\tilde{\gamma}^{(h+1)} := \|G_{K_{\mathrm{c}}(\gamma^{(h+1)})}\|_\infty$*, let* $\gamma_1^{(h+1)}$ *be any real number such that* $\tilde{\gamma}^{(h+1)} < \gamma_1^{(h+1)} < \gamma^{(h+1)}$*.*

If A is not unmixed, then Lemma 15 cannot be applied for the implementation of step 2 of the practical stabilization procedure. In this case, one has to resort to other techniques allowing one to solve problem (5.26): a reference including a detailed treatment for this case is [59]. Here we limit ourselves to notice that also in this case a necessary condition in order that problem (5.26) is feasible is that $\gamma_e < 1$, in fact:

Lemma 16 *If* A *is not Schur and* $K \in \mathbb{R}^{m \times n}$ *is such that* $A + BK$ *is Schur, then*

$$\|G_K\|_\infty \geq 1\,.$$

Proof. Assume that there exists K such that $A + BK$ is Schur and $\gamma_s := \|G_K\|_\infty < 1$. Let $q_{\mathcal{U}}(y) \equiv 0$, then $q_e(y) = q_{\mathcal{U}}(y) - y = -y$ and $\forall\, y \in \mathbb{R}^m$, $\|q_e(y)\|_2 = \gamma_e \|y\|_2$ with $\gamma_e = 1$. Therefore, by Corollary 5, system $x^+ = (A + BK)x + Bq_e(Kx)$ is asymptotically stable. But $x^+ = (A + BK)x + Bq_e(Kx) = (A + BK)x - BKx = Ax$ that contradicts the fact that A is not Schur. ■

Remark 21 *Given a quantized set* $\mathcal{U} \subset \mathbb{R}^m$ *and any input quantizer* $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$*, the classical* ℓ_2*–gain of the corresponding quantization error* q_e *is greater than or equal to 1.*

In fact, since $0 \in \mathcal{U}$ *is an isolated point (in fact,* $\mathcal{U}$ *is discrete),* $\exists\, z \in \mathbb{R}^m \setminus \{0\}$ *such that* $\operatorname{argmin}_{u\in\mathcal{U}} \|u - z\|_2 = 0$, *therefore*

$$\frac{\|q_e(z)\|_2}{\|z\|_2} = \frac{\|q_{\mathcal{U}}(z) - z\|_2}{\|z\|_2} \geq \frac{\|0 - z\|_2}{\|z\|_2} = 1\,.$$

This fact, together with Lemma 16, implies that if A *is not Schur, it is not possible to design a controller* $q_{\mathcal{U}}(Kx)$ *so that the classical small–gain condition given in Corollary 5 is satisfied. This is not surprising as we know that asymptotic stabilization of an open loop unstable linear system by means of a quantized control law is not possible (see Example 2 in Section 2.3).*

• ***Closed loop analysis: implementation of step 3a***

The implementation of step 3a requires some indications on the way to choose the parameters γ and Q to be considered in equation (5.17). Different choices of these parameters give rise to different results of the practical stability analysis. Some heuristic rules are provided in order to reduce conservativeness in the practical stability analysis.
Remind that $\gamma_s = \|G_K\|_\infty$. We have observed that, as γ varies in $\left]\|G_K\|_\infty, \frac{1}{\gamma_e}\right[$, the solution P of equation (5.17) changes slowly. Then, according to equation (5.23), the size of the final invariant ellipsoid $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ mainly depends on the choice of $Q > 0$. In particular, the value of $\lambda_{\min}(S)$ appearing in the denominator of the expression for r_{i}^2 is determined by Q: the more $\lambda_{\min}(S)$ is large, the more the final ellipsoid is small. Since S is a matrix of the form $S = Q + W(P,\gamma)$ $\big($where $W(P,\gamma)$ is a positive (semi) definite matrix$\big)$, then, with the purpose of increasing $\lambda_{\min}(S)$, γ and $Q > 0$ can be chosen so as to maximize $\lambda_{\min}(Q)$. To this end, a suboptimal choice consists of taking $\gamma \in \left]\gamma_s, \frac{1}{\gamma_e}\right[$ and letting $Q = \lambda \cdot I_n$, with $\lambda > 0$ such that $\|G_{\lambda\cdot I_n}\|_\infty < \gamma - \gamma_s$. Clearly, the larger is γ, the larger is $\lambda_{\min}(Q) = \lambda$. We can hence obtain a better result by choosing γ and $Q > 0$ as follows: fix $\epsilon > 0$ such that $\epsilon \ll \frac{1}{\gamma_e} - \gamma_s$ and let

$$\gamma := \frac{1}{\gamma_e} - \epsilon\,.$$

Then, according to condition (5.16), let

$$Q := \operatorname*{argmax}_{\left\{\begin{array}{l}\mathbb{R}^{n\times n} \ni X > 0\\ \|G_X\|_\infty \leq \gamma - \gamma_s - \epsilon\end{array}\right.} \lambda_{\min}(X)\,. \tag{5.32}$$

We have also observed that, when K is such that $\|G_K\|_\infty$ is close to $\gamma_{\inf}$ and $Q \simeq 0$, then $\lambda_{\min}(S) \simeq \lambda_{\min}(Q) \simeq 0$ so that the final invariant ellipsoid provided by Theorem 9 can be really large. This situation can not be avoided if $\gamma_{\inf} < 1/\gamma_e$ but $\gamma_{\inf} \cdot \gamma_e \simeq 1$: in this case, in fact, $\|G_K\|_\infty \simeq \gamma_{\inf}$ and $Q \simeq 0$ (see Fig. 5.2). This is consistent with the fact that, in this case, $\gamma_s \cdot \gamma_e \simeq 1$ (see case 2 of Example 22 in Section 5.3.3). As it will be discussed in Remark 24, it is important to take this fact into account when Theorem 9 is employed to deal with a stabilization problem in which the choice of the quantized set $\mathcal{U}$ is part of the design.

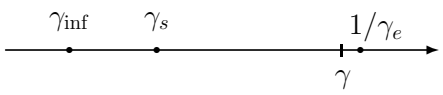

Figure 5.2: Representation of the mutual relations between $\gamma_{\inf}$, $\gamma_s = \|G_K\|_\infty$, γ $\big($appearing in equation (5.17)$\big)$ and $1/\gamma_e$ in the implementation of step 3a.

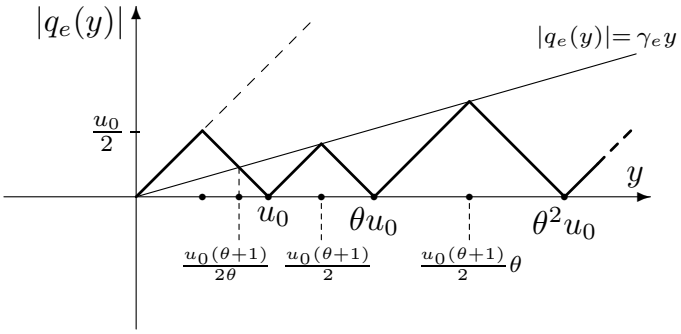

Figure 5.3: Logarithmic quantization of $\mathbb{R}$, the behavior of $|q_e|$ for $\theta = 1.8$.

• ***Input quantization: implementation of step 1***

Let us go back to the analysis of the external gain of the quantization error q_e and to the possibility of satisfying the hypotheses a–b of Theorem 9. The analysis of q_e consists of two steps: in order to determine an external gain, one has to study the function

$$\mathbb{R}^m \setminus \{0\} \ni y \mapsto \gamma(y) := \frac{\|q_e(y)\|_2}{\|y\|_2}$$

and, for fixed positive values of ϱ_0, to find an upper bound for $\sup_{\|y\|_2>\varrho_0} \gamma(y)$; in order to determine the absolute quantization error E_0 near 0, one has to find an upper bound for $\sup_{\|y\|_2\le\varrho_0} \|q_e(y)\|_2$. This study, at least theoretically, can be done for any arbitrarily assigned input set $\mathcal{U}$ and input quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$. Some typical examples of input quantizers are considered below. For the sake of clarity of presentation, only the results are presented while all the technical details can be found in Appendix A.4.1.

Example 19 (The logarithmic quantization of $\mathbb{R}$) *Let $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ be a logarithmic quantization of $\mathbb{R}$ with parameters (u_0, θ), where $u_0 > 0$ and $\theta > 1$ (see Definition 8 in Section 2.1). The corresponding quantization error $q_e(y) = q_{\mathcal{U}}(y) - y$ is standard with natural*

external gain $\gamma_e = \frac{\theta-1}{\theta+1}$ *. More precisely,* q_e *satisfies the hypotheses* a–b *of Theorem 9 with*

$$\begin{cases} \varrho_0 = \frac{u_0(\theta+1)}{2\theta} \\ \gamma_e = \frac{\theta-1}{\theta+1} \\ E_0 = \frac{u_0}{2} \end{cases} \tag{5.33}$$

and $\frac{u_0(\theta+1)}{2\theta}$ *is the smallest value of* ϱ_0 *so that the corresponding external gain is the natural one.*

As it is required for the feasibility of problem (5.26)*, it holds that* $\gamma_e < 1$*.*

Actually, the analysis of q_e *can be done according to the variation of* ϱ_0*. In general, changes of* ϱ_0 *induce changes of the parameters* E_0 *and* γ_e*. Consequently, Procedure 1 is modified beginning from step 1 and different closed loop systems having different practical stability properties may be obtained. In particular, the size of the final invariant ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *depends on these parameters. Let us analyze the way the final invariant set* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *changes with respect to the case where* ϱ_0 *takes the value in equation* (5.33)*. As* q_e *is standard, the* ϱ_0*–external gain does not change if* $\varrho_0 > \frac{u_0(\theta+1)}{2\theta}$*. On the contrary, the absolute quantization error near the origin*

$$E_0(\varrho_0) := \sup_{\|y\|_2 \le \varrho_0} \|q_e(y)\|_2$$

is non decreasing with ϱ_0 *and* $\lim_{\varrho_0 \to +\infty} E_0(\varrho_0) = +\infty$*. This behavior reflects on the size of the final invariant ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$*, in fact: since the external gain is constant, the matrix* P *can be held constant (i.e., steps 2 and 3a of Procedure 1 do not vary by changing* ϱ_0*); on the other hand* r_{i}^2 *increases quadratically with* E_0 *(see equation* (5.23)*). If instead* $\varrho_0 < \frac{u_0(\theta+1)}{2\theta}$*, then, as* ϱ_0 *decreases, the* ϱ_0*–external gain increases until it reaches the maximal value equal to* 1 *for* $\varrho_0 = \frac{u_0}{2}$ *(see Fig. 5.3 as well as Fig. A.2 in Appendix A.4.1 where all the details are given). Whilst, correspondingly, the absolute quantization error remains constant. Hence, the only effect of decreasing* ϱ_0 *is that of restricting the range of feasible choices for* γ *in step 2 of Procedure 1. Thus, there is no improvement in reducing* ϱ_0*. The bottom line is that the least conservative practical stability result is obtained by choosing* γ_e *and* E_0 *for step 1 of Procedure 1 according to equation* (5.33)*.*

The problem consisting of tuning the parameters defining the control set $\mathcal{U}$ *so that desired values of* E_0 *and* ϱ_0*–external gain* γ_e *are obtained will be referred to as the* inverse problem*. This question is important when the control design includes also the possibility of choosing the quantized set* $\mathcal{U}$ *(see Remark 24 for a discussion on these problems and Example 24 in Section 5.3.3). As far as the "inverse problem" for the logarithmic quantization of* $\mathbb{R}$ *is concerned, both the external gain* γ_e *and the absolute quantization error near the origin* E_0 *can be made arbitrarily small by choosing* θ *sufficiently close to* 1 *and* u_0 *sufficiently close to* 0*, respectively.* ♣

There are two main types of quantized multi–input sets depending on whether $\mathcal{U} \subset \mathbb{R}^m$ is in the form of a cartesian product, that is $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ (with $\mathcal{U}_i \subset \mathbb{R}$), or not. An

in–depth discussion on the main differences between the two structures is given in Remark 22 at the end of Example 21. Here we only recall that if $\mathcal{U}$ is in the form of a cartesian product, then it is possible to deal with input quantizers $q_\mathcal{U}$ acting separately on each component of the input vector (such a $q_\mathcal{U}$ is called a componentwise quantizer). In the other case instead, the input components are not decoupled and the multi–input space must be jointly quantized. The following two examples illustrate these two cases for a planar control space and allow us to point out some features of the two schemes.

Example 20 (The componentwise logarithmic quantization of $\mathbb{R}^2$) *Let $q_\mathcal{U} : \mathbb{R}^2 \to \mathcal{U}$ be a logarithmic quantization of $\mathbb{R}^2$ with parameters $\big((u_{01},\theta_1),(u_{02},\theta_2)\big)$, where for $i = 1,2$, $u_{0i} > 0$ and $\theta_i > 1$ (see Definition 8 in Section 2.1). Assume, for simplicity, that*

$$\begin{array}{rccl} q_\mathcal{U} : & \mathbb{R}^2 & \to & \mathcal{U} \\ & y & \mapsto & \big(q_{\mathcal{U}_1}(y_1)\,, q_{\mathcal{U}_2}(y_2)\big), \end{array}$$

with $q_{\mathcal{U}_i} : \mathbb{R} \to \mathcal{U}_i$ being a logarithmic quantization of $\mathbb{R}$ with parameters (u_{0i},θ_i). In this case

$$q_e(y) = q_\mathcal{U}(y) - y = \big(q_{e1}(y_1)\,, q_{e2}(y_2)\big),$$

where[8] *$q_{ei}(y_i) = q_{\mathcal{U}_i}(y_i) - y_i$.*
For $i = 1,2$, let $\varrho_{0i} := \frac{u_{0i}(\theta_i+1)}{2\theta_i}$, then $\forall\, \varrho_0 \geq \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$, q_e satisfies the hypotheses a–b *of Theorem 9 with*

$$\begin{cases} \gamma_e(\varrho_0) = \max\left\{ \sqrt{\gamma_{e1}^2 + \left(\frac{\varrho_{02}}{\varrho_0}\right)^2 (1-\gamma_{e1}^2)}\,,\ \sqrt{\gamma_{e2}^2 + \left(\frac{\varrho_{01}}{\varrho_0}\right)^2 (1-\gamma_{e2}^2)} \right\} \\ E_0(\varrho_0) = \sqrt{\mathtt{E}_{01}(\varrho_0)^2 + \mathtt{E}_{02}(\varrho_0)^2}\,, \end{cases} \tag{5.34}$$

where $\gamma_{ei} := \frac{\theta_i - 1}{\theta_i + 1}$ and $\mathtt{E}_{0i}(\varrho_0) := \max\limits_{|y_i| \leq \varrho_0} |q_{ei}(y_i)|$.

An explicit formula for $\mathtt{E}_{0i}(\varrho_0)$ is given as follows: with $n_i(\varrho_0) := \left\lfloor \log_{\theta_i} \frac{2\varrho_0}{u_{0i}(\theta_i+1)} \right\rfloor$, it holds that

$$\mathtt{E}_{0i}(\varrho_0) = \max\left\{ \frac{u_{0i}}{2}\,,\ \gamma_{ei} \frac{u_{0i}(\theta_i+1)}{2} \theta_i^{n_i(\varrho_0)}\,,\ |u_{0i}\theta_i^{n_i(\varrho_0)+1} - \varrho_0| \right\}. \tag{5.35}$$

Let us analyze the main properties of these quantities. The external gain γ_e given in equation (5.34) is a decreasing function and it is such that $\forall\, \varrho_0 \geq \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$,

$$\max\{\gamma_{e1}\,,\, \gamma_{e2}\} < \gamma_e(\varrho_0) < 1 \tag{5.36}$$

with $\lim_{\varrho_0 \to +\infty} \gamma_e(\varrho_0) = \max\{\gamma_{e1}\,,\, \gamma_{e2}\}$. In particular, the componentwise logarithmic quantization is not standard[9]*. More in detail: $\gamma_e(\varrho_0) > \max\{\gamma_{e1}\,,\, \gamma_{e2}\}$ because of what we call the*

[8] For a general nearest neighbor quantizer, the functions q_{ei} are not well–defined but equations (5.34) and (5.35) still hold true.

[9] Actually, the expression for $\gamma_e(\varrho_0)$ given in equation (5.34) is just an upper bound for the smallest feasible value of γ_e to be a ϱ_0–external gain (i.e., for *the* ϱ_0–external gain). Nevertheless, it is not difficult to show that the decreasing property holds true also for such a smallest feasible value.

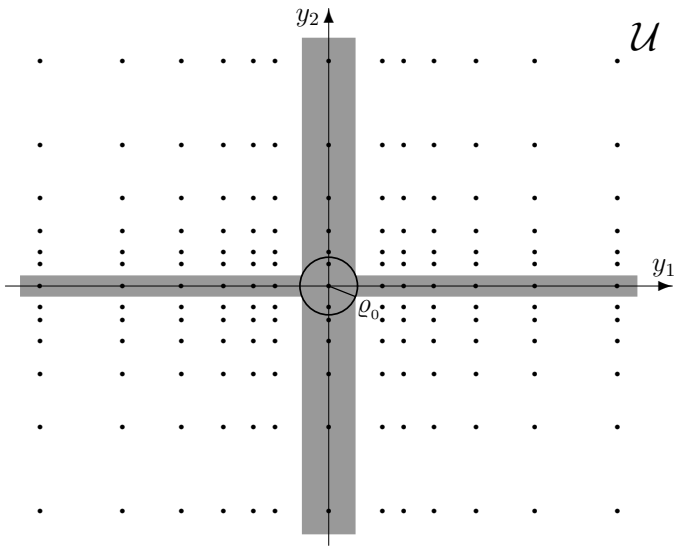

Figure 5.4: Example of a set $\mathcal{U} \subset \mathbb{R}^2$ logarithmically quantized with parameters $\big((u_{01},\theta_1),(u_{02},\theta_2)\big)$: $u_{02} = \frac{2}{5}\cdot u_{01}$, $\theta_1 = 1.4$ and $\theta_2 = 1.6$. The shaded region is the portion of the input space affected by the truncation of a nearest neighbor quantizer.

anisotropy *of the componentwise quantization. Namely, while in the logarithmic quantization of* $\mathbb{R}$ *the truncation of the quantizer does not affect* $q_{\mathcal{U}}$ *out of a bounded neighborhood of* 0*, in the componentwise case instead there are special unbounded regions showing the traces of truncation (see Fig. 5.4). This happens when* $\|y\|_2 \geq \varrho_0$ *but* $\exists\, i \in \{1,2\}$ *such that* $|y_i| < \varrho_{0i}$ *(i.e., when only one of the components of* y *is truncated by the corresponding scalar quantizer). On the other hand this truncation effect fades away as the distance from the origin increases and this accounts for the decrease of* γ_e *at the increase of* ϱ_0*.*
Thus, the external gain can be reduced by increasing ϱ_0 *but, similarly to Example 19,* E_0 *is a non decreasing function of* ϱ_0 *such that*[10] $\lim_{\varrho_0\to+\infty} E_0(\varrho_0) = +\infty$*. Also in this case it is natural to expect that large values of the absolute quantization error* E_0 *lead to a large size of the final invariant ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *and hence to weak practical stability properties of the closed loop system. Nevertheless, in this case the analysis is much more involved than the one we did in Example 19 because* γ_e *is not constant. There is indeed a trade off between* γ_e *and* E_0 *which are, respectively, decreasing and non decreasing with* ϱ_0*. We do not analyze this trade off because the complexity of the relations between the parameters contributing to the definition of the final invariant set* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *makes really hard the study of the way such a final set varies with* ϱ_0*. Such a complexity is apparent by looking at Procedure 1 where, differently from Example 19, the results of all the steps are affected by the variation of* ϱ_0*. Here we limit ourselves to claim that if* $\varrho_0 = \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$ *(i.e., if* ϱ_0 *takes the minimal allowed value*

[10] Also the expression for $E_0(\varrho_0)$ given in equation (5.34) is just an upper bound for $\sup_{\|y\|_2\leq\varrho_0}\|q_e(y)\|_2$, but the qualitative behavior of these two quantities is the same.

in equation (5.34)) *, then*

$$\gamma_e(\varrho_0) = \max\left\{ \frac{\sqrt{\gamma_{e1}^2\varrho_{01}^2 + \varrho_{02}^2}}{\varrho_0}\,,\, \frac{\sqrt{\varrho_{01}^2 + \gamma_{e2}^2\varrho_{02}^2}}{\varrho_0} \right\} > \frac{\sqrt{2}}{2}\,,$$

more precisely,

$$\begin{cases} \inf & \gamma_e\left(\sqrt{\varrho_{01}^2 + \varrho_{02}^2}\right) = \dfrac{\sqrt{2}}{2} \\ u_{01} > 0 \\ u_{02} > 0 \\ \theta_1 > 1 \\ \theta_2 > 1 \end{cases} \tag{5.37}$$

(as usual, the proof can be found in Appendix A.4.1) .

This property provides a spin–off for the "inverse problem" as it is clarified by inequality (5.38) *below. Also in this case both the external gain* $\gamma_e(\varrho_0)$ *and the absolute quantization error near the origin* $E_0(\varrho_0)$ *can be made arbitrarily small. This is achieved by fixing* $\varrho_0 > 0$ *, then by choosing* θ_i *sufficiently close to* 1 *and* u_{0i} *sufficiently close to* 0 *. In fact, according to equation* (5.34) *, it is sufficient to have both* $\frac{\varrho_{0i}}{\varrho_0}$ *and* γ_{0i} *close to* 0 *. More details can be found in Appendix A.4.1 . Notice that to obtain a small* ϱ_0*–external gain,* $\frac{\varrho_{0i}}{\varrho_0}$ *must be small and, consistently with equation* (5.37) *, if* $\frac{\varrho_{0i}}{\varrho_0} < \epsilon \ll 1$ *(for* $i = 1, 2$ *) , then*

$$\varrho_0 > \frac{\sqrt{\varrho_{01}^2 + \varrho_{02}^2}}{\epsilon\sqrt{2}} \gg \sqrt{\varrho_{01}^2 + \varrho_{02}^2}\,. \tag{5.38}$$

♣

Finally, let us present an example of a quantized control set $\mathcal{U} \subset \mathbb{R}^2$ where the different input components are not independently quantized.

Example 21 (The joint radial logarithmic quantization of $\mathbb{R}^2$) *Let* $q_{\mathcal{U}} : \mathbb{R}^2 \to \mathcal{U}$ *be a radial logarithmic quantization of* $\mathbb{R}^2$ *with parameters* (N, u_0, θ) *, where* $\mathbb{N} \ni N \geq 3$ *,* $u_0 > 0$ *and* $\theta > 1$ *(see Definition 9 in Section 2.1) . The corresponding quantization error* $q_e(y) = q_{\mathcal{U}}(y) - y$ *is standard with natural external gain* $\gamma_e = \sqrt{1 - \frac{4\theta\cos^2(\pi/N)}{(\theta+1)^2}}$ *. More precisely,* q_e *satisfies the hypotheses* a–b *of Theorem 9 with*

$$\begin{cases} \varrho_0 = \frac{u_0(\theta+1)}{2\theta\cos(\pi/N)} \\ \gamma_e = \sqrt{1 - \frac{4\theta\cos^2(\pi/N)}{(\theta+1)^2}} \\ E_0 = \max\left\{ \frac{u_0}{2\cos(\pi/N)}\,;\, \frac{u_0}{2\theta}\sqrt{(\theta-1)^2 + (1+\theta)^2\tan^2(\pi/N)} \right\} \end{cases} \tag{5.39}$$

and $\frac{u_0(\theta+1)}{2\theta\cos(\pi/N)}$ *is the smallest value of* ϱ_0 *so that the corresponding external gain is the natural one.*

Also in this case, $\gamma_e < 1$ *as desired. Because* q_e *is standard, there is no benefit in choosing* $\varrho_0 > \frac{u_0(\theta+1)}{2\theta\cos(\pi/N)}$ *. Nevertheless, differently from the logarithmic quantization of* $\mathbb{R}$ *, it*

is possible that a suitable choice of $\varrho_0 < \frac{u_0(\theta+1)}{2\theta\cos(\pi/N)}$ *results in a less conservative practical stability result. In fact, for such a* ϱ_0 *the external gain increases, but it can be shown that, if* $\frac{u_0}{2\theta}\sqrt{(\theta-1)^2+(1+\theta)^2\tan^2(\pi/N)} > \frac{u_0}{2\cos(\pi/N)}$ *, then* E_0 *decreases (whereas in the scalar case* E_0 *was constant). Again, we do not carry out the analysis of this trade off because the interdependence between the parameters implicated in the definition of the final invariant ellipsoid is really involved. However, considering* γ_e *and* E_0 *for step 1 of Procedure 1 according to equation* (5.39) *is an effective choice.*

Let us consider the "inverse problem": both the external gain γ_e *and the absolute quantization error near the origin* E_0 *can be made arbitrarily small by choosing* N *sufficiently large,* θ *sufficiently close to* 1 *and* u_0 *sufficiently close to* 0*.* ♣

Remark 22 (Componentwise quantization vs joint quantization) *Let us analyze the main differences between the componentwise quantization (i.e.,* $\mathcal{U}$ *is in the form of a cartesian product) and the joint quantization of a multi–input set. While doing that, we point out how some peculiarities of Examples 20 and 21 are related to the different structure of these two types of quantization.*

The first point we want to emphasize is concerned with the structure assumed by the controller. Since in the componentwise quantization the input quantizer $q_{\mathcal{U}}$ *acts separately on each component of the control vector, then the* m *input channels can be independently run by the controller without information exchange among them. In the joint quantization case instead, the independence between the input components is lost and a central intelligence is needed in the controller to run the different input channels.*

As it has been clarified in [64], the same stability performance can be obtained with a coarser quantization if a joint quantizer is used rather than a componentwise one. Nevertheless, the advantage of considering a joint quantization may vanish by slightly changing the assumptions. In this respect, an example is provided by the design of the so called "minimum distortion" quantizer, a critical issue in the framework of control under communication constraints [39, 18]. The problem consists of designing the quantizer that minimize the average norm of the quantization error while guaranteeing stability (under some constraints on the number of control values or on the density of quantization). Well, it has been shown in [52] that, if on the one hand it is true that the optimal result is obtained with a joint quantization, on the other hand the improvements with respect to the componentwise case are modest whilst complexity grows up exponentially with the dimension of the input space.

Indeed, there is in general a computational complexity problem with joint quantization that often renders more attractive the componentwise structure. As a matter of fact, many works recently appeared in the literature and dealing with joint quantization are limited to planar input sets (see e.g., [40]). Also Example 21 provided here is for $\mathcal{U} \subset \mathbb{R}^2$ *and it is not easily generalizable to an input space of higher dimension. On the contrary, the definition of the componentwise quantization can be directly extended to any dimension* $m \in \mathbb{N}$*.*

Finally, there is a structural problem with componentwise quantizations, which is what in Example 20 was referred to as anisotropy*. Since each input component is independently quan-*

tized, then each component is also separately affected by the truncation near 0 *of the corresponding scalar quantizer. On the overall input space, this causes the propagation of the truncation effects in special directions and on unbounded regions (see Fig. 5.4). In fact, as we have already noticed in Example 20, if* $y \in \mathbb{R}^m$ *is such that* $\exists\, i \in \{1, 2, \ldots, m\}$ *so that* y_i *is in a neighborhood of* 0 *where the* i*–th scalar quantizer is truncated (and this happens for* $y \in \mathbb{R}^m$ *of arbitrarily large norm), then such an* y *is affected by the truncation of the combined quantizer. This causes deterioration of performance with respect to a joint quantizer where instead it is possible to limit the drawbacks of truncation to a bounded neighborhood of* 0*. Moreover, the analysis itself of a componentwise quantizer is complicated by anisotropy. For instance, the estimate of* γ_e *and* E_0*, becomes more and more involved as the dimension of the input space increases. This happens because the truncation can affect any combination of the* m *components of* y*.*

Anisotropy has an influence also for the "inverse problem" (i.e., the design of a quantized set $\mathcal{U}$ *so that desired values of* E_0 *and of* ϱ_0*–external gain can be obtained). We have seen in Example 20 that, in order to obtain a small value of the* ϱ_0*–external gain, it is necessary that* $\varrho_0 \gg \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$*. That is, the external gain must be measured in a region sufficiently far from the origin where the truncation effect is attenuated. For joint logarithmic quantizers, the matter is completely different: because of the standard property, the design of quantization ensuring a small external gain can be done irrespective of* ϱ_0 *(see Examples 19 and 21).*

Remark 23 (Generalized logarithmic quantization) *As we have shown in Example 3 of Section 2.3, closed loop asymptotic stability, rather than mere practical stability, can be achieved if generalized quantized control sets are considered. An easy method allowing for the asymptotic stabilization of a linear system under generalized input quantization consists in satisfying a small–gain condition.*

For instance, let $q_{\mathcal{U}}$ *be a generalized (radial) logarithmic quantization of* $\mathbb{R}$ *(*$\mathbb{R}^2$*): in this case, we have shown in Section 2.1 that the quantization error* q_e *has classical* ℓ_2*–gain* $\gamma_* < 1$ *(the value of* γ_* *is provided in Lemma 3 and Lemma 4, respectively). Therefore, Corollary 5 can be applied and, provided that the* H_∞ *control problem of finding* $K \in \mathbb{R}^{m \times n}$ *such that* $\|G_K\|_\infty < 1/\gamma_*$ *is feasible, asymptotic stability in the Lyapunov sense can be ensured.*

A separate discussion is needed for the componentwise logarithmic quantization of $\mathbb{R}^m$*. Although we noticed in Section 2.1 that the cartesian product of generalized logarithmically quantized sets is not a generalized quantized set, it is interesting to consider also this case. For such a generalized componentwise logarithmic "quantization" of* $\mathbb{R}^m$*, it is straightforward to see that the "quantization" error* q_e *has classical* ℓ_2*–gain* $\gamma_* = \max\{\gamma_{e1}, \gamma_{e2}, \ldots, \gamma_{em}\}$*, where* γ_{ei} *is the classical* ℓ_2*–gain of the quantization error along the* i*–th component.*

Notice that, because of the absence of truncation, the problems deriving from anisotropy are removed and also the generalized componentwise logarithmic "quantization" gives rise to a "quantization" error which is standard with a natural gain. This is the case considered in [39, 51].

Let us conclude this paragraph with a couple of remarks concerned with related issues: first, the problem where the quantized set $\mathcal{U}$ is not assigned and its choice is part of the design is discussed; afterwards, a backward step is done to the analysis of controlled invariance.

Remark 24 (On the design of the quantized set $\mathcal{U}$) *In the examples above, we have mentioned the importance of the "inverse problem" with reference to the case where the design of the quantized set $\mathcal{U}$ is part of the control synthesis. Let us discuss a couple of interesting cases where results based on the small–gain theorem can be usefully applied. Both of them rise in the context of the control under communication constraints.*

Given system (5.24), suppose that a stabilizing matrix K is given while both the quantized input set $\mathcal{U} \subset \mathbb{R}^m$ and the quantizer $q_{\mathcal{U}}$ must be designed so that the feedback law $u(x) = q_{\mathcal{U}}(Kx)$ guarantee desired practical stability properties. Theorem 9 allows one to solve this problem. A rough application of it consists of making the quantization error q_e satisfy the following properties: $\gamma_e < \frac{1}{\|G_K\|_\infty}$ (in order to ensure convergence properties) and E_0 is sufficiently small (so that a desired size of the final invariant ellipsoid $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ be achieved). Nonetheless, as it has been clarified in the discussion of the implementation of step 3a of Procedure 1, the convergence issue and the assignment of the size of the final set as resulting from Theorem 9 are not decoupled, therefore a more aware application of this result is recommended. There is indeed a trade off between γ_e and E_0, let us give some intuition of this: while the choice of a maximal allowed value for γ_e permits to minimize the density of quantization, on the other hand it also makes the gap $\frac{1}{\gamma_e} - \|G_K\|_\infty$ small. As such a gap becomes more and more narrow, in order to obtain a small invariant ellipsoid $\mathcal{E}_{P,r_{\mathrm{i}}^2}$, E_0 must be smaller and smaller[11]: to this end, the quantizer must be truncated nearer and nearer 0. In other words, by increasing γ_e, a quantized set that tends to accumulate towards 0 is needed and, despite the decrease of quantization density, the number of control values within fixed neighborhoods of the origin may increase. This fact has an obvious drawback to the problem of control under communication constraints where the goal is to minimize the number of control values that allow one to achieve a desired stability property.

Once the proper values of γ_e and E_0 to be associated to some quantization error q_e have been identified, the construction of $\mathcal{U}$ and of $q_{\mathcal{U}}$ that accomplish these parameters is mainly a geometrical issue. The choice of a standard logarithmic quantizer is the natural one for two interrelated reasons: first, because a standard logarithmic quantizer allows one to obtain convergence by minimizing the density of quantization [39]; secondly, because a standard logarithmic quantizer is characterized by its natural external gain and this makes the "inverse problem" easy to solve. Of course, also a componentwise logarithmic quantization is a feasible option as well as other types of quantizers, including a uniform one. In Examples 19, 20 and 21 we have provided all the necessary tools to solve the problem when $m = 1$ or $m = 2$.

[11] The discussion on the implementation of step 3a of Procedure 1 is useful to gain insight on this fact: if $\frac{1}{\gamma_e} - \|G_K\|_\infty$ is small, then also the matrix Q resulting from the solution of problem (5.32) is small. Namely, there isn't enough room to add a non negligible term $Q > 0$ in equation (5.17) with the purpose of increasing $\lambda_{\min}(S)$ (see case 2 of Example 22 in Section 5.3.3). Since, in general, $\lambda_{\min}(S)$ can be quite small (especially if $\|G_K\|_\infty \simeq \gamma_{\inf}$), then E_0 must be small accordingly (see equation (5.23)).

A numerical example is reported in Section 5.3.3 (see Example 24).
Other methods for the synthesis of $\mathcal{U}$ *and* $q_{\mathcal{U}}$ *have been also considered in the literature, as for instance those related with the theory of "minimum distortion quantization" and "locational optimization" problems [36]. A paper where the link between stabilization under quantized control and the locational optimization problem is pointed out is [18]. That paper also provides comprehensive references to the wide literature on the subject.*

A variation of this problem consists in designing both K *and* $\mathcal{U}$ *so as to guarantee practical stability properties and to minimize the number of control values or the density of quantization. This problem has been stated in [39]: the solution presented in that paper is restricted to SISO systems and has not a direct interpretation in terms of a small–gain condition. Successively, the same problem has been considered again in [51] where a solution has been given for MIMO systems. In this case the approach is based on a "sector bound" method [67], a technique directly related to the small–gain approach, but only "componentwise" input quantizations have been taken into account. Moreover, only asymptotic stabilization is considered so that the issues inherent with practical stability are bypassed.*

Remark 25 (Back to controlled invariance analysis) *The results of this section, in particular Theorem 9.ı, expand to a broader class of quantizers the kind of controlled invariance analysis presented in Section 3.1.4. This wider ensemble includes logarithmically quantized sets and, more in general, quantized sets so that it is possible to construct an input quantizer* $q_{\mathcal{U}}$ *such that the* relative *quantization error* $\frac{\|q_e(y)\|_2}{\|y\|_2}$*, rather than the* absolute *one, is bounded.*

5.3.3 Numerical examples

The practical stabilization technique and some aspects discussed in the previous section are illustrated in the numerical examples below.

Example 22 *Let us consider the following quantized input system:*

$$\begin{cases} x^+ = Ax + Bu = \begin{pmatrix} 0 & 1 \\ -1 & 5/2 \end{pmatrix} x + \begin{pmatrix} 1 \\ 2 \end{pmatrix} u \\ \\ u \in \mathcal{U} \subset \mathbb{R}\,, \end{cases} \tag{5.40}$$

where $\mathcal{U}$ *is a logarithmically quantized set with parameters* $(u_0, \theta) = (1\,,\,2)$*.*
Two practically stabilizing control laws are synthesized and the corresponding closed loop dynamics are analyzed through the implementation of Procedure 1.
First, notice that the pair (A, B) *is not reachable but it is stabilizable. The eigenvalues of* A *are* $\lambda_1(A) = 1/2$ *and* $\lambda_2(A) = 2$*, hence* A *is unmixed. According to equation* (5.29)*,*

$$\gamma_{\inf} = \inf\big\{\, \|G_K\|_\infty \,|\, K \in \mathbb{R}^{1\times 2} \text{ is such that } A + BK \text{ is Schur} \big\} = 2\,.$$

Case 1:

1. Input quantization: *we consider a nearest neighbor input quantizer* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$. *According to equation* (5.33), *the corresponding quantization error* q_e *satisfies the hypotheses* a–b *of Theorem 9 with*
$$\begin{cases} \varrho_0 = \frac{3}{4} \\ \gamma_e = \frac{1}{3} \\ E_0 = \frac{1}{2}\,. \end{cases}$$

2. Control synthesis: *problem* (5.26) *is feasible if and only if* $\gamma_\infty \in \left]\, \gamma_{\inf}\,,\, \frac{1}{\gamma_e} \,\right] = \left]\, 2\,,\, 3 \,\right]$ (*see equation* (5.31)). *Let us apply Lemma 15: we choose* γ_∞ *ensuring that the closed loop system has a* H_∞*–norm close to* $\gamma_{\inf}$. *With* $\gamma_\infty = 2.01$, *the Riccati equation* (5.27a) *is solved by*
$$P^* = \begin{pmatrix} 0.4430 & -0.8860 \\ -0.8860 & 1.7719 \end{pmatrix}$$
and, according to equation (5.28),
$$K := K_{\mathrm{c}}(2.01) = \begin{pmatrix} 0.6645 & -1.3289 \end{pmatrix}.$$

3. Closed loop analysis: *let us apply Theorem 9 for the analysis of the resulting closed loop dynamics. It holds that*
$$F := A + BK = \begin{pmatrix} 0.6645 & -0.3289 \\ 0.3289 & -0.1579 \end{pmatrix}$$
whose eigenvalues are $\lambda_1(F) = 1/2$ *and* $\lambda_2(F) = 0.0066$. *It holds that* $\|G_K\|_\infty = 2.0066 < 2.01$.

   (a) *Following the discussion on the implementation of step 3a, we consider equation* (5.17) *with* $\gamma = 2.999$ *and* $Q = 0.194 \cdot I_2$ *(we made the suboptimal choice of considering matrices* Q *of the type* $Q = \lambda \cdot I_2$*). With these choices, equation* (5.17) *is solved by*
$$P = \begin{pmatrix} 0.7990 & -0.9630 \\ -0.9630 & 1.9992 \end{pmatrix}$$
   whose eigenvalues are $\lambda_1(P) = 0.2645$ *and* $\lambda_2(P) = 2.5338$.

   (b) *According to equation* (5.23), $r_{\mathrm{i}}^2 = 4.0579$.

4. Final result: $\forall\, r_1^2 \geq r_2^2 > 4.0579$, *system* (5.40) *controlled with* $u(x) = q_{\mathcal{U}}(Kx)$ *is* $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\right)$*–stable.*

This result becomes more expressive by providing the lengths of the semi–axes of the final invariant ellipsoid $\mathcal{E}_{P,r_{\rm i}^2}$ *. The* j*–th semi–axis is* $\mathrm{s}_j = \frac{r_{\rm i}}{\sqrt{\lambda_j(P)}}$ *, in this case:*

$$\begin{cases} \mathrm{s}_1 = 3.9170 \\ \mathrm{s}_2 = 1.2655\,. \end{cases}$$

Case 2: *let us modify the control synthesis step and let us show how performance change when the gap* $\frac{1}{\gamma_e} - \|G_K\|_\infty$ *is reduced. The controller is synthesized so that* $\|G_K\|_\infty \cdot \gamma_e = 0.9$ *(in case 1 we have* $\|G_K\|_\infty \cdot \gamma_e = 0.6687$*). We hence look for* K *such that* $\|G_K\|_\infty = 2.7$*. To this end, we use Lemma 15 together with a recalibration of the Riccati equation (see Remark 19), so we let* $\gamma_\infty = 4.243$ *in system* (5.27)*. With this choice of* γ_∞ *we obtain*

$$K := K_{\rm c}(4.243) = \begin{pmatrix} 0.5294 & -1.0588 \end{pmatrix}$$

and $\|G_K\|_\infty = 2.7$ *as desired. By repeating the steps done in case 1, we find*

$$P = \begin{pmatrix} 0.3957 & -0.7764 \\ -0.7764 & 1.5638 \end{pmatrix}$$

$\big($*with* $\lambda_1(P) = 0.0082$ *and* $\lambda_2(P) = 1.9513\big)$ *and* $r_{\rm i}^2 = 1.0420 \cdot 10^4$*. Thus, the semi–axes of the final ellipsoid* $\mathcal{E}_{P,r_{\rm i}^2}$ *are*

$$\begin{cases} \mathrm{s}_1 = 1126.8 \\ \mathrm{s}_2 = 73.1\,. \end{cases}$$

The deterioration of the practical stability result is evident. It must be stressed that, as a consequence of the reduced gap $\frac{1}{\gamma_e} - \|G_K\|_\infty$*, in step 3a we find a matrix* Q *which is much smaller than the one we found in case 1: in this case, in fact,* $Q = 0.0061 \cdot I_2$*. Moreover,* $\lambda_{\min}(S) = 0.0062 \simeq \lambda_{\min}(Q)$*: this means that the size of* $\lambda_{\min}(S)$ *is dictated by the size of* $\lambda_{\min}(Q)$*. Hence, the main responsibility for the large size of the final invariant ellipsoid fall in the lack of room to find a* $Q > 0$ *causing the increase of* $\lambda_{\min}(S)$*. Other numerical simulations provide a clear evidence of this trend to the worsening of the result as* $\|G_K\|_\infty$ *approaches* $1/\gamma_e$*.*

We will be back on this example at the end of Section 6.2.2 (see Example 29). In both cases 1 and 2, we will show that the analysis of the closed loop dynamics can be significantly improved by supplementing it with arguments based on a small–gain theorem in the ℓ_1 *functional space. For case 1, the simulation of a trajectory will be also reported in Fig. 6.3.* ♣

Example 23 *Let us consider again the system defined by the pair* (A, B) *considered in Example 11 of Section 3.1.4, that is*

$$\begin{cases} x^+ = Ax + Bu = \begin{pmatrix} 2 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & -4 & 4 \end{pmatrix} x + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} u \\ u \in \mathcal{U} \subset \mathbb{R}^2, \end{cases} \tag{5.41}$$

where $\mathcal{U} \subset \mathbb{R}^2$ *is a radially logarithmically quantized set with parameters* $(N, u_0, \theta) = (32\,,\,1\,,\,5/4)$ *.*
In this example, a stabilizing control gain $K^* \in \mathbb{R}^{m\times n}$ *is determined which realizes the minimum value* $\gamma_{\inf}$ *that can be attained by* $\|G_K\|_\infty$ *. Then, the practical stability properties of the closed loop system* $x^+ = Ax + Bq_{\mathcal{U}}(K^*x)$ *are analyzed, where* $q_{\mathcal{U}} : \mathbb{R}^2 \to \mathcal{U}$ *is a nearest neighbor quantizer.*
Recall that the pair (A, B) *is reachable, thus stabilizable, and neither the pair* $(A, B^{[1]})$ *nor* $(A, B^{[2]})$ *are stabilizable (where* $B^{[i]}$ *denotes the* i*–th column of* B*). The eigenvalues of* A *are all equal to* 2*, hence* A *is unmixed. By taking advantage of the algorithm for the computation of* $\gamma_{\inf}$ *provided in Remark 20, we find that* $\gamma_{\inf} = 4$ *and that such a value is attained for*

$$K^* = \begin{pmatrix} -8/5 & -8/5 & 0 \\ 0 & 4 & -16/5 \end{pmatrix}.$$

The closed loop dynamics is $x^+ = Fx + Bq_e(K^*x)$*, where*

$$F := A + BK^* = \begin{pmatrix} 2/5 & 2/5 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 4/5 \end{pmatrix}$$

and its eigenvalues are $\lambda_1(F) = 2/5$*,* $\lambda_2(F) = 4/5$ *and* $\lambda_3(F) = 0$*.*
We can apply Theorem 9 for the practical stability analysis of the closed loop system, in fact: according to equation (5.39)*, the quantization error* q_e *associated to the quantizer* $q_{\mathcal{U}}$ *satisfies the hypotheses* a–b *of Theorem 9 with*

$$\begin{cases} \varrho_0 = 0.9044 \\ \gamma_e = 0.1478 \\ E_0 = \max\Big\{\, 0.5024\,;\, 0.1336 \,\Big\} = 0.5024\,. \end{cases}$$

Since $\|G_{K^*}\|_\infty \cdot \gamma_e = 0.5911 < 1$*, then also hypothesis* c *of Theorem 9 is satisfied. The quantitative analysis of practical stability can hence be obtained by implementing step 3 of Procedure 1. The method is the same we used in Example 22: with* $\gamma = 1/\gamma_e - \epsilon = 6.7665$ *and* $Q = 0.1007 \cdot I_3$*, equation* (5.17) *is solved by*

$$P = \begin{pmatrix} 3.2137 & 3.1130 & -0.0671 \\ 3.1130 & 19.2137 & -12.8671 \\ -0.0671 & -12.8671 & 16.8894 \end{pmatrix}$$

whose eigenvalues are $\lambda_1(P) = 1.7738$*,* $\lambda_2(P) = 6.3744$ *and* $\lambda_3(P) = 31.1687$*. Finally, according to equation* (5.23)*,* $r_{\mathrm{i}}^2 = 2390.7$ *and the semi–axes of the final ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *are*

$$\begin{cases} \mathrm{s}_1 = 36.7118 \\ \mathrm{s}_2 = 19.3661 \\ \mathrm{s}_3 = 8.7580\,. \end{cases}$$

♣

Finally, let us give an example where the quantized set $\mathcal{U}$ is part of the design.

Example 24 *Let us consider a scalar system*

$$x^+ = \frac{5}{2}x + u\,,$$

assume that $K = -2$ *and that the problem consists in designing a quantized set* $\mathcal{U} \subset \mathbb{R}$ *and an input quantizer* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *such that the following properties hold for the closed loop system*

$$x^+ = \frac{5}{2}x + q_{\mathcal{U}}(Kx) \ :$$

ı) $\forall\, r_1^2 \geq r_2^2 > 1$ *, the system is* $\big([-r_1, r_1], [-r_2, r_2]\big)$*–stable;*
ıı) $\forall\, x$ *such that* $|x| > 1$ *, it holds that* $|x^+| \leq \frac{3}{4}|x|$ *.*

With the terminology of Theorem 9, the former property corresponds to impose that the final invariant ellipsoid $\mathcal{E}_{P,r_{\rm i}^2}$ *is the interval* $[-1, 1]$ *. The latter condition is a performance requirement corresponding to enforce a lower bound on the speed of convergence irrespective to the quantization error.*
As usual, the closed loop dynamics can be rewritten in the form

$$x^+ = \frac{1}{2}x + q_e(-2x)\,.$$

It holds that $\gamma_s = \big\| -2\big(z - \frac{1}{2}\big)^{-1}\big\|_\infty = 4$ *.*
To solve the problem, we apply Theorem 9: a value of γ_e *is determined such that* $\gamma_s \cdot \gamma_e < 1$ *and the performance requirement in ıı is fulfilled; then we look for* E_0 *so as to obtain the desired size of the final invariant set; finally, we choose* $\mathcal{U}$ *and* $q_{\mathcal{U}}$ *so that the corresponding quantization error attains the determined parameters* γ_e *and* E_0 *(i.e., we solve an "inverse problem") .*
Since

$$\frac{|x^+|}{|x|} = \frac{\big|\frac{1}{2}x + q_e(-2x)\big|}{|x|} \leq \frac{1}{2} + 2\frac{\big|q_e(-2x)\big|}{|-2x|}\,,$$

then, in order that for $|x| > 1$ *, be* $|x^+| \leq \frac{3}{4}|x|$ *, it is sufficient to impose that* q_e *has a* ϱ_0*–external gain* γ_e *with* $\varrho_0 = 2$ *and* $\gamma_e \leq 1/8$ *(as far as the condition on* γ_e *is concerned, see also inequality* (2.5) *with* γ_e *in place of* $\frac{\theta-1}{\theta+1}$ *and* $3/4$ *in place of* 1 *in the right–hand side). Any choice of* γ_e *in this range ensures that* $\gamma_s \cdot \gamma_e < 1$*: we pick* $\gamma_e = 1/8$ *, which is the value maximizing the dispersion of the quantized set. The implementation of step 3a of Procedure 1 yields* $Q = 3.996$ *,* $\gamma = 7.999$ *and* $P = 11.5015$ *. In order that* $\mathcal{E}_{P,r_{\rm i}^2} = [-1, 1]$ *, it must be* $r_{\rm i}^2 = P$ *. Therefore, the inverse implementation of step 3b (i.e., where* $r_{\rm i}^2$ *is given and* E_0 *is unknown) allows one to determine the needed value of* E_0*: by equation* (5.23) *, an easy computation provides* $E_0 = 1/2$ *(take advantage of equation* (5.20) *and of the fact that all the matrices indeed are scalars) .*
Let us construct $\mathcal{U}$ *and* $q_{\mathcal{U}}$*: we look for a logarithmic quantization of* $\mathbb{R}$ *with parameters* (u_0, θ) *such that the corresponding quantization error* q_e *has* 2*–external gain equal to* $1/8$

and absolute quantization error within $\mathcal{B}_2$ *bounded by* $1/2$. *To this end, we solve the "inverse problem" for the logarithmic quantization of* $\mathbb{R}$ *(see Example 19). A solution is given by* $\theta = 9/7$ *and* $u_0 = 1$, *in fact: with this choice of the parameters, according to equation* (5.33), $\varrho_0 = 8/9 < 2$ *and* $\gamma_e = 1/8$, *then the* 2*–external gain is equal to* $1/8$ *as desired. Let us check that* $\max_{|y|\leq 2} |q_e(y)| \leq \frac{1}{2}$*: by equation* (5.33), *for* $|y| \leq 8/9$, $|q_e(y)| \leq u_0/2 = 1/2$, *whereas, for* y *such that* $8/9 < |y| \leq 2$ *it holds that* $|q_e(y)| \leq \frac{1}{8}|y| \leq \frac{1}{4}$. ♣

Chapter 6

The small–gain approach in ℓ_1: quantized multi–input

In this chapter, we lay down the theory for a different solution to the practical stabilization problem for quantized input systems. The main tool is still a generalized small–gain theorem, but here the signals are viewed as elements of the functional space ℓ_∞. As explained in Section 5.1, this case gives rise to the so called control problem in ℓ_1 and it is the natural approach to the problem when quantized control sets are considered. Nevertheless, the main limitations of the theory on ℓ_1 control lie in the fact that, although it has been the subject of a certain amount of literature [28, 34, 113, 65, 38, 7], this theory is not as elegant and exhaustive as the one for the H_∞ control.
The chapter is organized as follows: in Section 6.1, we focus on the study of the ℓ_∞–gain of linear systems. A novel approach to the computation of an upper bound for the ℓ_∞–gain and to the control synthesis in ℓ_1 under static output feedback is presented. The subsequent Section 6.2 is the counterpart of Section 5.3: in Section 6.2.1 we prove results on the practical stability analysis of feedback systems which rely on a small–gain theorem in ℓ_1 (i.e., involving the ℓ_∞–gains of the operators forming the system). These results can be merged together with those from Section 5.3.1, thus providing mixed H_∞/ℓ_1 analysis tools. In Section 6.2.2, we illustrate how the developed theory can be applied for the practical stabilization of quantized input systems. We also show that the practical stabilization technique based on invariant hypercubes, presented in Chapter 4, can be interpreted as the outcome of the generalized small–gain theorem in ℓ_1: this observation allows us to extend that technique to a larger class of controllers than the quantized deadbeat only. Examples are reported showing the weight of the contribution brought by the ℓ_1 theory to the stabilization problem, in particular, as far as the analysis of the steady–state performance are concerned. Although some results are not as general as those presented in the context of the H_∞ control, the following sections provide significant improvements to the theory developed so far and disclose interesting issues for further investigations.

6.1 A factorization approach to the analysis and control synthesis in ℓ_1

In this section, we study the ℓ_∞–gain of a linear system and the problem of the synthesis of a static output feedback ensuring that the ℓ_∞–gain of the closed loop dynamics is below a desired threshold. As remarked in the above introduction, the theory of the ℓ_1 control still presents some gaps to be filled. For instance, closed formulae for the ℓ_∞–gain of a system are missing. In [15], an upper bound for the ℓ_∞–gain of a linear system is given in terms of the singular values of the Hankel operator [53]. This result has been the basis to carry out efficient numerical algorithms for the computation of $\|\mathcal{G}\|_\infty$ (see [6, 58]). Although these methods allow one to find a good estimate of the ℓ_∞–gain of a system, they do not appear to be practical to deal with control synthesis problems. As far as control synthesis is concerned, the approaches proposed in [28, 34, 113, 65, 38] take advantage of the convex structure of the set of all stabilizing controllers and, either the problem is transformed into an infinite dimensional linear optimization, or a linear (or quadratic) programming formulation is presented. Also in this case, algorithmic procedures are carried out for the numerical approximation of the solution. When the synthesis is restricted to static output feedback controllers, the problem becomes more complex because the set of stabilizing control gains is not convex. Actually, the minimization of the closed loop ℓ_∞–gain by means of static output feedback has been less investigated.
The main contribution of this section consists in providing an easy method for the computation of an upper bound for the ℓ_∞–norm of the input/output operator $\mathcal{G}$ associated to a BIBO–stable system (5.7). Although the proposed bound is not always feasible (i.e., it can be computed only for some particular systems) and often quite conservative, yet it turns out to be useful in some interesting cases. In particular, the bound is proved to be tight for single–input *positive* systems [49]. Furthermore, the proposed method is suitable to deal with control synthesis: a sufficient criterion is provided that allows one to find a linear static output feedback so that the ℓ_∞–gain of the closed loop dynamics is below a desired threshold. This can be done by solving a system of linear inequalities.
The upper bound is obtained by factorizing the overall dynamics in terms of subsystems whose computation of the ℓ_∞–gain is simple. To obtain the desired factorization, it is convenient to resort to the representation of system (5.7) in the frequency domain, namely to the transfer matrix $G(z)$ of the system. Let us stress once more that the norm $\|G\|_\infty$ considered in the previous section should not be confused with the norm $\|\mathcal{G}\|_\infty$ whose study is the main subject of this section (see Remark 17 in Section 5.2).

Let us recall the definitions of the norms of signals in ℓ_∞ and in ℓ_1. For $\vec{v} : \mathbb{N} \to \mathbb{R}^k$,

$$\|\vec{v}\|_\infty := \sup_{t\in\mathbb{N}} \|v(t)\|_\infty ;$$

while for $\vec{g} : \mathbb{N} \to \mathbb{R}^{q\times m}$,

$$\|\vec{g}\|_1 := \sum_{t=1}^{+\infty} \|g(t)\|_\infty .$$

6.1.1 Analysis in ℓ_1

Under the assumption of BIBO–stability, let us consider the problem of computing $\|\mathcal{G}\|_\infty$ (see equation (5.11)). The following Proposition provides an expression of $\|\mathcal{G}\|_\infty$ in terms of the impulse response.

Proposition 15 *If system* (5.7) *is BIBO–stable, then*

$$\|\mathcal{G}\|_\infty = \max_{i=1,\dots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| . \tag{6.1}$$

Proof. Although this is a well known fact in systems theory, its proof is explicitly reported in Appendix A.5.1 for completeness. ■

Corollary 6 *The following relation holds between* $\|\mathcal{G}\|_\infty$ *and the* ℓ_1*–norm of the impulse response:*

$$\frac{1}{q}\|\vec{g}\,\|_1 \le \|\mathcal{G}\|_\infty \le \|\vec{g}\,\|_1 .$$

In particular, for single–output systems, $\|\mathcal{G}\|_\infty = \|\vec{g}\,\|_1$.

Proof. In fact:

$$\|\mathcal{G}\|_\infty = \max_{i=1,\dots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| \le \sum_{\tau=0}^{+\infty} \max_{i=1,\dots,q} \sum_{j=1}^{m} |g_{i,j}(\tau)| = \|\vec{g}\,\|_1 .$$

On the other hand,

$$\begin{aligned} q \cdot \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| \ge & \quad \sum_{i=1}^{q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| = \\ = & \quad \sum_{\tau=0}^{+\infty} \sum_{i=1}^{q} \sum_{j=1}^{m} |g_{i,j}(\tau)| \ge \\ \ge & \quad \sum_{\tau=0}^{+\infty} \max_{i=1,\dots,q} \sum_{j=1}^{m} |g_{i,j}(\tau)| = \|\vec{g}\,\|_1 . \end{aligned}$$

■

By reversing the order of the summations, equation (6.1) can be rewritten as

$$\|\mathcal{G}\|_\infty = \left\| \begin{pmatrix} \|\vec{g}_{1,1}\|_1 & \cdots & \|\vec{g}_{1,m}\|_1 \\ \vdots & \ddots & \vdots \\ \|\vec{g}_{q,1}\|_1 & \cdots & \|\vec{g}_{q,m}\|_1 \end{pmatrix} \right\|_\infty . \tag{6.2}$$

Thus, the analysis of the ℓ_∞–gain of a MIMO system can be reduced to the study of the ℓ_∞–gain for SISO systems.

The expression for $\|\mathcal{G}\|_\infty$ given in Proposition 15 is not practical because, in general, it requires the computation of an infinite series. There are two classes of systems where $\|\mathcal{G}\|_\infty$ can be exactly computed: externally positive systems and FIR systems.

Definition 30 *(See [49]) System* (5.7) *is said to be* externally positive *iff its impulse response* $\vec{g}$ *is positive.*

Clearly, $\vec{g}$ is positive if and only if $\forall\, i = 1, \ldots, q$ and $\forall\, j = 1, \ldots, m$, $\vec{g}_{i,j}$ is positive. Hence, also the analysis of external positivity can be reduced to SISO systems.

Lemma 17 *Consider system* (5.7), *assume that* $e \in \mathbb{R}$ *and* $y \in \mathbb{R}$. *Let*

$$G(z) = \frac{c_n z^{n-1} + c_{n-1} z^{n-2} + \cdots + c_1}{z^n - f_n z^{n-1} - \cdots - f_1}$$

be the transfer function of the system. A sufficient condition for the external positivity of system (5.7) *is that* $\forall\, k = 1, \ldots, n$, $c_k \geq 0$ *and* $f_k \geq 0$.

Proof. See also [49]. The transfer matrix of the linear system $\Sigma(F, B, C)$, with

$$F = \begin{pmatrix} 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ f_1 & f_2 & \ldots & f_n \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} c_1 & c_2 & \ldots & c_n \end{pmatrix},$$

is $G(z)$. As the entries of the matrices F, B and C are non–negative, then the impulse response is positive. ■

Proposition 16 ($\|\mathcal{G}\|_\infty$ of externally positive systems) *If system* (5.7) *is BIBO–stable and externally positive, then*

$$\|\mathcal{G}\|_\infty = \|G(1)\|_\infty .$$

Proof. Because $\vec{g}$ is positive, then $\forall\, i = 1, \ldots, q$ and $\forall\, j = 1, \ldots, m$,

$$G_{i,j}(1) = \sum_{t=0}^{+\infty} g_{i,j}(t) = \|\vec{g}_{i,j}\|_1 .$$

The thesis then immediately follows by equation (6.2). ■

In next Lemma 18, a well known relation between the H_∞ norm and the ℓ_∞–gain of a BIBO–stable SISO system is recalled. A consequence of this relation is that, for externally positive SISO systems, the two norms coincide.

Lemma 18 (Equivalence of H_∞ and ℓ_1 norms for externally positive SISO systems) *Consider a BIBO–stable system* (5.7), *assume that* $e \in \mathbb{R}$ *and* $y \in \mathbb{R}$. *Then*

$$\|G\|_\infty \leq \|\mathcal{G}\|_\infty .$$

Moreover, if the system is externally positive, then

$$\|G\|_\infty = \|\mathcal{G}\|_\infty = |G(1)| .$$

Proof. For $\theta \in [0, 2\pi[$, it holds that

$$|G(e^{i\theta})| = \left| \sum_{t=0}^{+\infty} g(t) \cdot \frac{1}{e^{i\theta t}} \right| \leq \sum_{t=0}^{+\infty} |g(t)| \cdot \left| \frac{1}{e^{i\theta t}} \right| = \|\vec{g}\|_1 = \|\mathcal{G}\|_\infty .$$

Therefore,

$$\|G\|_\infty = \max_{\theta \in [0,2\pi[} |G(e^{i\theta})| \leq \|\mathcal{G}\|_\infty .$$

For a positive system, $\|\mathcal{G}\|_\infty = |G(1)| \leq \|G\|_\infty$. The thesis follows. ■

Definition 31 *Consider system* (5.7) *and let* $\vec{g}$ *be its impulse response. The system is said to be* finite impulse response *(FIR) iff* $\exists\, r \in \mathbb{N}$, *such that* $\forall\, t > r$, $g(t) = 0$.

By definition, system (5.7) is FIR if and only if

$$G(z) = \frac{1}{z^r} \sum_{t=1}^{r} g(t) z^{r-t}$$

for some $r \in \mathbb{N}$. This is equivalent to say that all the poles of the system are in 0. Hence, a sufficient condition for system (5.7) to be FIR is that F is nilpotent, the condition is also necessary if the system is reachable and observable.
It is useful to introduce the following notation: if $G(z)$ is the transfer matrix of a FIR system, we let $\mathtt{G} \in \mathbb{R}^{q \times mr}$ be defined by

$$\mathtt{G} := \left[g(1) \,|\, \cdots \,|\, g(r) \right]. \tag{6.3}$$

For a FIR system, the computation of $\|\mathcal{G}\|_\infty$ is trivial as the series in equation (6.1) is a finite sum. According to equation (5.3),

$$\|\mathcal{G}\|_\infty = \max_{i=1,\ldots,q} \sum_{\tau=1}^{r} \sum_{j=1}^{m} |g_{i,j}(\tau)| = \|\mathtt{G}\|_\infty . \tag{6.4}$$

We are ready to introduce the main result of this section. Consider system (5.7), assume without loss of generality that $q \geq m$ and let $G(z)$ be the transfer matrix of the system: $G(z)$ is a strictly proper rational matrix. It is always possible to factorize $G(z)$ in the form

$$G(z) = N(z)\big(I_m + D(z)\big)^{-1}, \tag{6.5}$$

where $N(z)$ and $D(z)$ are the transfer matrices of FIR systems[1] (see Fig. 6.1). Three methods to obtain this factorization are described in next Remark 27.

Theorem 10 (Bound for $\|\mathcal{G}\|_\infty$) *Consider system* (5.7), *assume without loss of generality that* $q \geq m$ *and let the transfer matrix of the system be factorized as in equation* (6.5). *If* $\|\mathcal{D}\|_\infty < 1$, *then the system is BIBO–stable and*

$$\|\mathcal{G}\|_\infty \leq \frac{\|\mathcal{N}\|_\infty}{1 - \|\mathcal{D}\|_\infty} .$$

[1] If $q \leq m$, just consider a factorization of the type $G(z) = \big(I_q + D(z)\big)^{-1} N(z)$.

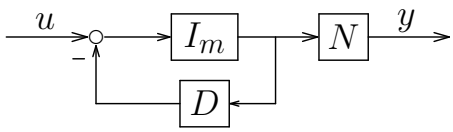

Figure 6.1: Block diagram representation of the factorization of $G(z)$ considered in Theorem 10.

Proof. Denote by $\mathcal{I}_m : \ell_\infty(\mathbb{R}^m) \to \ell_\infty(\mathbb{R}^m)$ the identity operator. Since $\|\mathcal{D}\|_\infty < 1$, then the operator $(\mathcal{I}_m+\mathcal{D})^{-1} : \ell_\infty(\mathbb{R}^m) \to \ell_\infty(\mathbb{R}^m)$ is well–defined and $\|(\mathcal{I}_m+\mathcal{D})^{-1}\|_\infty \leq \frac{1}{1-\|\mathcal{D}\|_\infty}$ (see Lemma 27 in Appendix A.5.1). Also, $\mathcal{N} : \ell_\infty(\mathbb{R}^m) \to \ell_\infty(\mathbb{R}^q)$ because $N(z)$ is a FIR system. From the factorization (6.5) of $G(z)$, it follows that $\mathcal{G} = \mathcal{N} \circ (\mathcal{I}_m+\mathcal{D})^{-1} : \ell_\infty(\mathbb{R}^m) \to \ell_\infty(\mathbb{R}^q)$ and hence,

$$\|\mathcal{G}\|_\infty \leq \|\mathcal{N}\|_\infty \|(\mathcal{I}_m + \mathcal{D})^{-1}\|_\infty \leq \frac{\|\mathcal{N}\|_\infty}{1 - \|\mathcal{D}\|_\infty} .$$

■

Remark 26 *For a given transfer matrix $G(z)$, a factorization of $G(z)$ of the type in equation (6.5) is not unique. Different factorizations of the same $G(z)$ give rise, in general, to different operators $\mathcal{N}$ and $\mathcal{D}$, and hence to different upper bounds for $\|\mathcal{G}\|_\infty$. Also, it may happen that condition $\|\mathcal{D}\|_\infty < 1$, which allows one to apply Theorem 10, is satisfied for some specific factorizations whilst it is not for some others. These phenomena, which are illustrated in next Examples 25 and 26, raise the issue of the search for conditions on $G(z)$ that guarantee the existence of a factorization (6.5) such that $\|\mathcal{D}\|_\infty < 1$ and the search for the factorization that minimizes the corresponding upper bound for $\|\mathcal{G}\|_\infty$. In this thesis, however, we do not face these points and we let them as open issues for future investigations.*

Remark 27 (Methods for the computation of the factorization (6.5)) *Let us provide some methods that, for any strictly proper transfer matrix $G(z)$ ($q \geq m$), allow one to obtain a factorization as in equation (6.5).*

• **Method 1:** *factorization (6.5) can be obtained in the form of a right coprime rational matrix factorization of $G(z)$ (see [68]). The standard state space approach to obtain such a factorization is the following: let $\Sigma(F, B, C)$ be a reachable and observable linear system whose transfer matrix is $G(z)$ and K be a matrix such that all the eigenvalues of $F + BK$ are in 0. Then*

$$\begin{cases} N(z) = C\big(zI - (F + BK)\big)^{-1}B \\ D(z) = K\big(zI - (F + BK)\big)^{-1}B \end{cases} \tag{6.6}$$

are such that equation (6.5) holds[2]. The details to determine such a $\Sigma(F, B, C)$ can be found, for instance, in [114].

[2]If $q \leq m$, just consider a left coprime rational matrix factorization: let L be such that all the eigenvalues of $F + LC$ are in 0 and let $N(z) = C\big(zI - (F + LC)\big)^{-1}B$ and $D(z) = C\big(zI - (F + LC)\big)^{-1}L$.

Next methods 2 and 3 are purely algebraic approaches that do not involve state space realizations of $G(z)$.

• **Method 2:** *let* $d(z)$ *be the monic least common multiple of the denominators of* $G(z)$ *and* $r := \deg(d)$. *Let*

$$\bar{D}(z) := \frac{d(z)}{z^r} I_m \,,$$

then equation (6.5) *holds with*

$$\begin{cases} N(z) = G(z)\bar{D}(z) \\ D(z) = \bar{D}(z) - I_m \,. \end{cases} \tag{6.7}$$

• **Method 3:** *for* $j = 1, \ldots, m$, *let* $d_j(z)$ *be the monic least common multiple of the denominators appearing in the* j*–th column of* $G(z)$ *and* $r_j := \deg(d_j)$. *Let*[3]

$$\bar{D}(z) := \operatorname{diag}\left\{ \frac{d_1(z)}{z^{r_1}}, \ldots, \frac{d_m(z)}{z^{r_m}} \right\} ,$$

then equation (6.5) *holds with*

$$\begin{cases} N(z) = G(z)\bar{D}(z) \\ D(z) = \bar{D}(z) - I_m \,. \end{cases} \tag{6.8}$$

In general, $N(z)$ *and* $\bar{D}(z)$ *resulting from method 3 are not right coprime rational matrices. In case they are, the factorizations resulting from methods 1 and 3 coincide.*

Next Proposition 17 is just a particularization of Theorem 10 to single–input systems. This particular case allows us to point out that, for a special class of externally positive single–input systems, the bound resulting from Theorem 10 is indeed an equality (see Corollary 7 below) .

Proposition 17 (ℓ_∞–gain of single–input systems) *Consider a strictly proper rational transfer matrix* $G(z)$ *of a linear system with* $u \in \mathbb{R}$ *and* $y \in \mathbb{R}^q$. *Let* $d(z) = z^n - \sum_{k=1}^{n} f_k z^{k-1}$ *be the polynomial of the poles of the system*[4] *and* $G^{(I)}(z)$ *be defined by* $G_k^{(I)}(z) := z^{k-1}/d(z)$, $k = 1, \ldots, n$. *Let* $C \in \mathbb{R}^{q \times n}$ *be such that* $G(z) = CG^{(I)}(z)$. *If* $f := \sum_{k=1}^{n} |f_k| < 1$, *then the system is BIBO–stable and*

$$\|\mathcal{G}\|_\infty \leq \frac{\|C\|_\infty}{1 - f} \,.$$

[3]Where $\operatorname{diag}\left\{ \frac{d_1(z)}{z^{r_1}}, \ldots, \frac{d_m(z)}{z^{r_m}} \right\} := \sum_{j=1}^{m} \frac{d_j(z)}{z^{r_j}} e_j e_j'$, $e_j \in \mathbb{R}^m$.

[4]It is easy to see that $d(z)$ is the monic least common multiple of the denominators of $G(z)$; see also [25] .

Proof. Since $\sum_{k=1}^{n} |f_k| < 1$, then the poles of $G(z)$ lie into the interior of the unit ball of $\mathbb{C}$ (see Lemma 28 in Appendix A.5.1) and the system is BIBO–stable. As $G(z) = CG^{(I)}(z)$, then $G(z)$ can be factorized in the form

$$\begin{cases} G(z) = N(z)\big(1 + D(z)\big)^{-1}, & \text{with} \\ N(z) = \frac{1}{z^n} C \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{n-1} \end{pmatrix} = \frac{1}{z^n} \sum_{t=1}^{n} C e_t z^{t-1} \\ D(z) = \frac{-\sum_{t=1}^{n} f_t z^{t-1}}{z^n} \end{cases} \tag{6.9}$$

(where e_t is the t–th vector of the canonical basis). By equation (6.4), $\|\mathcal{N}\|_\infty = \|C\|_\infty$ and $\|\mathcal{D}\|_\infty = \sum_{t=1}^{n} |f_t| < 1$. We can hence apply Theorem 10 which yields

$$\|\mathcal{G}\|_\infty \leq \frac{\|\mathcal{N}\|_\infty}{1 - \|\mathcal{D}\|_\infty} = \frac{\|C\|_\infty}{1 - f}\,.$$

■

Corollary 7 *Under the assumptions of Proposition 17, if* $\forall\, i = 1, \ldots, q$ *and* $\forall\, j = 1, \ldots, n$, $C_{i,j} \geq 0$ *and* $\forall\, k = 1, \ldots, n$, $f_k \geq 0$, *then*

$$\|\mathcal{G}\|_\infty = \frac{\|C\|_\infty}{1 - f}\,.$$

Proof. The system is externally positive because, $\forall\, i = 1, \ldots, q$, $G_i(z) = e_i' C G^{(I)}(z)$ satisfies the hypotheses of Lemma 17. Thus, by Proposition 16, $\|\mathcal{G}\|_\infty = \|G(1)\|_\infty$ and, by equation (6.9),

$$\|G(1)\|_\infty = \left\| \frac{1}{1 - f}\, C \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_\infty = \frac{\|C\|_\infty}{1 - f}\,.$$

■

Let us illustrate, through some numerical examples, how to take advantage of Theorem 10 to compute an upper bound for $\|\mathcal{G}\|_\infty$. The transfer matrix $G(z)$ is factorized according to the different methods presented in Remark 27 and the resulting bounds for $\|\mathcal{G}\|_\infty$ are compared.

Example 25 *Consider a MIMO system whose transfer matrix is*

$$G(z) = \begin{pmatrix} \frac{1}{z+1/3} & \frac{z+1}{(z-1/2)(z+1/4)} \\ \frac{-2}{z-1/2} & \frac{z}{(z-1/4)(z+1/4)} \end{pmatrix}.$$

Case 1: *Let us factorize* $G(z)$ *according to the method 1 in Remark 27. A reachable and*

observable linear system $\Sigma(F,B,C)$ *whose transfer matrix is* $G(z)$ *can be easily found to be*

$$\begin{cases} x(t+1) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/6 & 1/6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1/32 & 1/16 & 1/2 \end{pmatrix} x(t) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} u(t) \\ y(t) = \begin{pmatrix} -1/2 & 1 & -1/4 & 3/4 & 1 \\ -2/3 & -2 & 0 & -1/2 & 1 \end{pmatrix} x(t)\,. \end{cases}$$

With

$$K = \begin{pmatrix} -1/6 & -1/6 & 0 & 0 & 0 \\ 0 & 0 & 1/32 & -1/16 & -1/2 \end{pmatrix},$$

by equation (6.6) *one gets* $G(z) = N(z)\big(I_m + D(z)\big)^{-1}$, *where*

$$\begin{cases} N(z) = \begin{pmatrix} \frac{z-\frac{1}{2}}{z^2} & \frac{z^2+\frac{3}{4}z-\frac{1}{4}}{z^3} \\ \frac{-2z-\frac{2}{3}}{z^2} & \frac{z-\frac{1}{2}}{z^2} \end{pmatrix} = \\ \quad = \frac{1}{z^3}\left[\begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix} z^2 + \begin{pmatrix} -1/2 & 3/4 \\ -2/3 & -1/2 \end{pmatrix} z + \begin{pmatrix} 0 & -1/4 \\ 0 & 0 \end{pmatrix}\right] \\ D(z) = \begin{pmatrix} \frac{-\frac{1}{6}z-\frac{1}{6}}{z^2} & 0 \\ 0 & \frac{-\frac{1}{2}z^2-\frac{1}{16}z+\frac{1}{32}}{z^3} \end{pmatrix} = \\ \quad = \frac{1}{z^3}\left[\begin{pmatrix} -1/6 & 0 \\ 0 & -1/2 \end{pmatrix} z^2 + \begin{pmatrix} -1/6 & 0 \\ 0 & -1/16 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ 0 & 1/32 \end{pmatrix}\right]. \end{cases}$$

Thus,

$$\mathtt{N} = \big[\, n(1) \,|\, n(2) \,|\, n(3) \,\big] = \begin{pmatrix} 1 & 1 & -1/2 & 3/4 & 0 & -1/4 \\ -2 & 1 & -2/3 & -1/2 & 0 & 0 \end{pmatrix},$$

and, according to equation (6.4)*,* $\|\mathcal{N}\|_\infty = \|\mathtt{N}\|_\infty = \frac{25}{6}$. *Similarly,*

$$\mathtt{D} = \big[\, d(1) \,|\, d(2) \,|\, d(3) \,\big] = \begin{pmatrix} -1/6 & 0 & -1/6 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & -1/16 & 0 & 1/32 \end{pmatrix},$$

and $\|\mathcal{D}\|_\infty = \|\mathtt{D}\|_\infty = \frac{19}{32} < 1$. *Hence, by Theorem 10,*

$$\|\mathcal{G}\|_\infty \leq \frac{25/6}{1 - 19/32} = \frac{400}{39} \simeq 10.26\,.$$

Case 2: *Let us factorize* $G(z)$ *according to the method 2 in Remark 27. It holds that*

$$d(z) = (z + 1/3)(z - 1/2)(z + 1/4)(z - 1/4)\,,$$

so that, with

$$\bar{D}(z) = \frac{z^4 - \frac{1}{6}z^3 - \frac{11}{48}z^2 + \frac{1}{96}z + \frac{1}{96}}{z^4} I_2 \,,$$

by equation (6.7) *one gets* $G(z) = N(z)\big(I_2 + D(z)\big)^{-1}$ *, where*

$$\begin{cases} N(z) = \frac{1}{z^4} \begin{pmatrix} z^3 - \frac{1}{2}z^2 - \frac{1}{16}z + \frac{1}{32} & z^3 + \frac{13}{12}z^2 - \frac{1}{12} \\ -2z^3 - \frac{2}{3}z^2 + \frac{1}{8}z + \frac{1}{24} & z^3 - \frac{1}{6}z^2 - \frac{1}{6}z \end{pmatrix} \\ D(z) = \frac{1}{z^4}\left(-\frac{1}{6}I_2\, z^3 - \frac{11}{48}I_2\, z^2 + \frac{1}{96}I_2\, z + \frac{1}{96}I_2\right) . \end{cases}$$

Thus, $\|\mathcal{N}\|_\infty = \frac{25}{6}$ *and* $\|\mathcal{D}\|_\infty = \frac{5}{12} < 1$ *. Hence, by Theorem 10 ,*

$$\|\mathcal{G}\|_\infty \leq \frac{25/6}{1 - 5/12} = \frac{50}{7} \simeq 7.14 \,.$$

Case 3: *In this example, the factorization of* $G(z)$ *according to the method 3 in Remark 27 leads to the same factorization we found in Case 1 .*

Case 4: *Taking advantage of equation* (6.2) *and of external positivity properties of some of the components of* $\vec{g}$ *, the exact computation of* $\|\mathcal{G}\|_\infty$ *can be carried out. Since*

$$g_{1,1}(t) = \begin{cases} 0 & for \quad t = 0 \\ (-\frac{1}{3})^{t-1} & for \quad t \geq 1 \,, \end{cases}$$

then $\|\vec{g}_{1,1}\|_1 = \frac{1}{1-1/3} = \frac{3}{2}$ *. By Lemma 17, both* $\vec{g}_{1,2}$ *and* $\vec{g}_{2,2}$ *are positive; hence, by Proposition 16,* $\|\vec{g}_{1,2}\|_1 = G_{1,2}(1) = \frac{16}{5}$ *and* $\|\vec{g}_{2,2}\|_1 = G_{2,2}(1) = \frac{16}{15}$ *. As for* $\vec{g}_{2,1}$ *, it holds that* $\|\vec{g}_{2,1}\|_1 = 2\|\vec{g}^{\,+}_{2,1}\|_1$ *, where* $\vec{g}^{\,+}_{2,1}$ *is the impulse response of the externally positive system whose transfer function is* $G^+_{2,1}(z) := \frac{1}{z-1/2}$ *. Thus,* $\|\vec{g}_{2,1}\|_1 = 2G^+_{2,1}(1) = 4$ *. Therefore equation* (6.2) *yields*

$$\|\mathcal{G}\|_\infty = \left\| \begin{pmatrix} 3/2 & 16/5 \\ 4 & 16/15 \end{pmatrix} \right\|_\infty = \frac{76}{15} \simeq 5.067 \,.$$

♣

In Example 25 , the second method of factorization leads to a smaller upper bound for $\|\mathcal{G}\|_\infty$ than the one we found when we used the right coprime factorization. Nevertheless, as it is explicitly illustrated in next Example 26 , this is not true in general.
For a MIMO system, instead of applying Theorem 10 to the overall $G(z)$, one may use it to bound the ℓ_1–norm of each SISO impulse response $\vec{g}_{i,j}$ that composes the system, then to resort to equation (6.2) . While this approach may provide a less conservative upper bound for $\|\mathcal{G}\|_\infty$, on the other hand it does not appear to be profitable for control synthesis purposes.

Example 26 *Consider a MIMO system whose transfer matrix is*

$$G(z) = \begin{pmatrix} \frac{1}{z} & \frac{1}{z-1/2} \\ \frac{1}{z-1/4} & \frac{1}{z-1/5} \end{pmatrix} .$$

Also in this example, the factorization methods 1 and 3, presented in Remark 27, lead to the same factorization of $G(z)$. We hence consider only the algebraic methods.

Case 1: *Let us factorize $G(z)$ according to the method 2 in Remark 27. It holds that*

$$\begin{aligned} d(z) = &\quad z(z-1/2)(z-1/4)(z-1/5) = \\ = &\quad z^4 - \tfrac{19}{20}z^3 + \tfrac{11}{40}z^2 - \tfrac{1}{40}z\,. \end{aligned}$$

It immediately follows that,

$$\|\mathcal{D}\|_\infty = \frac{19}{20} + \frac{11}{40} + \frac{1}{40} = \frac{5}{4} > 1$$

and hence Theorem 10 cannot be applied.

Case 2: *If instead we factorize $G(z)$ according to method 3, then*

$$\begin{cases} d_1(z) = z(z-1/4) \\ d_2(z) = (z-1/2)(z-1/5)\,. \end{cases}$$

Hence, with

$$\bar{D}(z) = \operatorname{diag}\left\{\frac{d_1(z)}{z^2}\,,\,\frac{d_2(z)}{z^2}\right\} = \frac{1}{z^2}\begin{pmatrix} z^2 - \frac{1}{4}z & 0 \\ 0 & z^2 - \frac{7}{10}z + \frac{1}{10} \end{pmatrix},$$

by equation (6.8) one gets $G(z) = N(z)\big(I_2 + D(z)\big)^{-1}$, where

$$\begin{cases} N(z) = \frac{1}{z^2}\begin{pmatrix} z - \frac{1}{4} & z - \frac{1}{5} \\ z & z - \frac{1}{2} \end{pmatrix} \\ \\ D(z) = \frac{1}{z^2}\begin{pmatrix} -\frac{1}{4}z & 0 \\ 0 & -\frac{7}{10}z + \frac{1}{10} \end{pmatrix}. \end{cases}$$

Thus, $\|\mathcal{N}\|_\infty = \frac{5}{2}$ and $\|\mathcal{D}\|_\infty = \frac{4}{5} < 1$. Hence, by Theorem 10,

$$\|\mathcal{G}\|_\infty \leq \frac{5/2}{1 - 4/5} = \frac{25}{2}\,.$$

Case 3: *By Lemma 17, $G(z)$ is the transfer matrix of an externally positive system. Therefore, by Proposition 16,*

$$\|\mathcal{G}\|_\infty = \|G(1)\|_\infty = 3\,.$$

This gives evidence of the fact that the proposed bound may be quite conservative. Moreover, for multi–input systems also the tightness of the bound for positive systems is lost. ♣

6.1.2 Synthesis in ℓ_1

Let us consider now the following control synthesis problem in ℓ_1:

Problem 1 (Control synthesis: static output feedback) *For a given discrete–time system*

$$\begin{cases} x^+ = Ax + B(u+e) \\ y = Cx \\ x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m, \quad y \in \mathbb{R}^q \end{cases} \tag{6.10}$$

and $\gamma > 0$, find $K \in \mathbb{R}^{m\times q}$ such that, under the static output feedback $u = Ky$, the closed loop system

$$\begin{cases} x^+ = (A + BKC)x + Be \\ y = Cx\,. \end{cases} \tag{6.11}$$

is BIBO–stable and, denoted by $\mathcal{G}_K$ its input/output operator, it holds that $\|\mathcal{G}_K\|_\infty \le \gamma$.

As in Section 5.3.2, we deal with the "actuator disturbance" case. This is indeed the case that turns out to be useful for the practical stabilization of quantized input systems. One of the main features of Theorem 10 is that of being suitable to deal with this control synthesis problem.

Theorem 11 (Static output feedback in ℓ_1) *Consider system (6.10), assume without loss of generality that $q \ge m$ and let $G(z) = C(zI-A)^{-1}B$ be factorized in the form $G(z) = N(z)\big(I_m + D(z)\big)^{-1}$ as in equation (6.5). Consider the closed loop dynamics (6.11) under the static output feedback $u = Ky$, let $G_K(z) := C(zI-A-BKC)^{-1}B$ be the corresponding transfer matrix and, if system (6.11) is BIBO–stable, denote by $\mathcal{G}_K$ its input/output operator. If $K \in \mathbb{R}^{m\times q}$ is such that $\|\mathcal{D}_K\|_\infty < 1$, where $D_K(z) := D(z) - KN(z)$, then system (6.11) is BIBO–stable and*

$$\|\mathcal{G}_K\|_\infty \le \frac{\|\mathcal{N}\|_\infty}{1 - \|\mathcal{D}_K\|_\infty}\,. \tag{6.12}$$

Before proving the theorem, let us derive the solution to Problem 1 in terms of linear inequalities.

Corollary 8 (Linear inequalities formulation) *With the same notation of Theorem 11, let $\gamma \ge \|\mathcal{N}\|_\infty$. As in equation (6.3), let $\mathtt{N} = \big[\, n(1) \,|\, \cdots \,|\, n(r) \,\big] \in \mathbb{R}^{q\times mr}$ and $\mathtt{D} = \big[\, d(1) \,|\, \cdots \,|\, d(r) \,\big] \in \mathbb{R}^{m\times mr}$ (for suitable $r \in \mathbb{N}$) be the matrices associated to the FIR systems $N(z)$ and $D(z)$ appearing in the factorization (6.5). If $\exists\, K \in \mathbb{R}^{m\times q}$ such that*

$$\forall\, i = 1, \dots, m\,, \quad \sum_{j=1}^{mr} \left| \mathtt{D}_{i,j} - \sum_{l=1}^{q} K_{i,l}\mathtt{N}_{l,j} \right| \le 1 - \frac{\|\mathcal{N}\|_\infty}{\gamma}\,, \tag{6.13}$$

then

$$\|\mathcal{G}_K\|_\infty \le \gamma\,.$$

Proof. By Theorem 11 , a sufficient condition in order that $K \in \mathbb{R}^{m\times q}$ is such that $\|\mathcal{G}_K\|_\infty \leq \gamma$ is that $\|\mathcal{D}_K\|_\infty < 1$ and $\frac{\|\mathcal{N}\|_\infty}{1-\|\mathcal{D}_K\|_\infty} \leq \gamma$. This is equivalent to find $K \in \mathbb{R}^{m\times q}$ such that

$$\|\mathcal{D}_K\|_\infty \leq 1 - \frac{\|\mathcal{N}\|_\infty}{\gamma} .$$

Because $D_K(z) = D(z) - KN(z) = \frac{1}{z^r}\sum_{t=1}^{r}\big(d(t) - Kn(t)\big)z^{r-t}$, by equation (6.4) it holds that $\|\mathcal{D}_K\|_\infty = \|\mathtt{D} - K\mathtt{N}\|_\infty$. Condition (6.13) is tantamount to requiring that $\|\mathtt{D} - K\mathtt{N}\|_\infty \leq 1 - \frac{\|\mathcal{N}\|_\infty}{\gamma}$ $\big($see equation (5.3)$\big)$. ■

Proof of Theorem 11. It is sufficient to show that $G_K(z) = N(z)\big(I_m + D_K(z)\big)^{-1}$, the thesis then follows by Theorem 10 . The transfer matrix $G_K(z)$ can be rewritten as $G_K(z) = \big(I_q - G(z)K\big)^{-1}G(z)$. Thus,

$$\begin{aligned}
G_K(z) = &\ \big(I_q - G(z)K\big)^{-1}G(z) = \\
\overset{(a)}{=} &\ G(z)\big(I_m - KG(z)\big)^{-1} = \\
= &\ N(z)\big(I_m + D(z)\big)^{-1}\Big(I_m - KN(z)\big(I_m + D(z)\big)^{-1}\Big)^{-1} = \\
= &\ N(z)\big(I_m + D(z) - KN(z)\big)^{-1} = \\
= &\ N(z)\big(I_m + D_K(z)\big)^{-1},
\end{aligned}$$

where equality (a) follows by Lemma 29 in Appendix A.5.2 . ■

Theorem 11 and Corollary 8 provide a sufficient criterion for the solution of Problem 1 . However, because the upper bound (6.12) may be in general quite conservative, in some cases the linear inequalities (6.13) are not feasible even if a solution to the control synthesis problem exists. Moreover, the issues related with the non uniqueness of the factorization (6.5) (see Remark 26 and Examples 25 and 26) also affects the control synthesis problem.
Nonetheless, there is a special case of Problem 1 where the proposed control synthesis technique turns out to be particularly useful, namely, that of state feedback for single–input systems.

Proposition 18 (State feedback for single-input systems) *Consider a strictly proper rational transfer matrix $G(z)$ of a linear system with $u \in \mathbb{R}$ and $y \in \mathbb{R}^n$. Let $d(z) = z^n - \sum_{k=1}^{n} a_k z^{k-1}$ be the polynomial of the poles of the system and $G^{(I)}(z)$ be defined by $G_k^{(I)}(z) := z^{k-1}/d(z)$, $k = 1, \ldots, n$. Let $C \in \mathbb{R}^{n\times n}$ be such that $G(z) = CG^{(I)}(z)$. Then, $\forall\, \gamma \geq \|C\|_\infty$, a control gain $K \in \mathbb{R}^{1\times n}$ can be determined by solving a system* (6.13) *such that the closed loop dynamics with $u = Kx$ is BIBO–stable and $\|\mathcal{G}_K\|_\infty \leq \gamma$.*
Moreover, if $\forall\, i, j = 1, \ldots, n$, $C_{i,j} \geq 0$, then a solution exists to system (6.13) *so that the equality $\|\mathcal{G}_K\|_\infty = \gamma$ is satisfied.*

Proof. First notice that, because the system is of order n and $y \in \mathbb{R}^n$, then $C \in \mathbb{R}^{n\times n}$ is invertible. It holds that $G(z) = N(z)\big(1 + D(z)\big)^{-1}$, where $N(z)$ and $D(z)$ are defined

as in equation (6.9) (with a_t in place of f_t). Hence, $\mathtt{N} = C$, $\mathtt{D} = \begin{pmatrix} -a_1 & -a_2 & \cdots & -a_n \end{pmatrix}$ and $\|\mathcal{N}\|_\infty = \|C\|_\infty$. For $\gamma \geq \|C\|_\infty$, there exists a solution to system (6.13) if and only if $\exists\, K \in \mathbb{R}^{1\times n}$ such that $\|\mathtt{D} - K\mathtt{N}\|_\infty \leq 1 - \frac{\|C\|_\infty}{\gamma}$. Since $\mathtt{D} - K\mathtt{N} = (\mathtt{D}C^{-1} - K)C$, then $\|\mathtt{D} - K\mathtt{N}\|_\infty \leq \|\mathtt{D}C^{-1} - K\|_\infty \|C\|_\infty$. Therefore, K can be chosen so as to make $\|\mathtt{D}C^{-1} - K\|_\infty$, and hence $\|\mathtt{D} - K\mathtt{N}\|_\infty$, arbitrarily small.
Notice that the row vector $\mathtt{D} - K\mathtt{N}$ collects the coefficients of the polynomial of the poles of $G_K(z)$. The last statement is hence a direct consequence of Corollary 7: let $\gamma \geq \|C\|_\infty$, it is sufficient to pick $K \in \mathbb{R}^{1\times n}$ so that $\mathtt{D} - K\mathtt{N} = (\, -f_1 \; -f_2 \; \ldots \; -f_n \,)$, with $f_k \geq 0 \;\; \forall\, k = 1, \ldots, n$ and $\sum_{k=1}^n f_k = 1 - \frac{\|C\|_\infty}{\gamma}$. This is achieved with $K = \big(\mathtt{D} + (\, f_1 \; f_2 \; \ldots \; f_n \,)\big)C^{-1}$. ■

Let us provide an example where the control synthesis technique based on Corollary 8 allows one to solve Problem 1 for a MIMO system.

Example 27 *Let us consider the system*

$$\begin{cases} x(t+1) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 3/4 \end{pmatrix} x(t) + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \big(u(t) + e(t)\big) \\ \\ y(t) = \begin{pmatrix} 0 & -1 & 2 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix} x(t)\,. \end{cases}$$

The goal is to find $K \in \mathbb{R}^{2\times 2}$ such that, with $u = Ky$, the closed loop system is BIBO–stable with $\|\mathcal{G}_K\|_\infty \leq 10$.
It is a reachable and observable system whose poles are 1 and $-1/4$ (both with double multiplicity), therefore the system is not BIBO–stable (see property 2 of Lemma 14). The transfer matrix of the system is

$$G(z) = \begin{pmatrix} \frac{2z}{(z-1)(z+1/4)} & -\frac{1}{(z-1)(z+1/4)} \\ \frac{1/2}{(z-1)(z+1/4)} & \frac{z}{(z-1)(z+1/4)} \end{pmatrix} = \begin{pmatrix} 2z & -1 \\ 1/2 & z \end{pmatrix} \frac{1}{z^2 - \frac{3}{4}z - \frac{1}{4}}\,.$$

Hence, $G(z) = N(z)\big(I_2 + D(z)\big)^{-1}$ with

$$N(z) := \frac{1}{z^2}\left[\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} z + \begin{pmatrix} 0 & -1 \\ 1/2 & 0 \end{pmatrix}\right]$$

and

$$D(z) := \frac{1}{z^2}\left[\begin{pmatrix} -3/4 & 0 \\ 0 & -3/4 \end{pmatrix} z + \begin{pmatrix} -1/4 & 0 \\ 0 & -1/4 \end{pmatrix}\right].$$

Accordingly,

$$\mathtt{N} = \begin{pmatrix} 2 & 0 & 0 & -1 \\ 0 & 1 & 1/2 & 0 \end{pmatrix} \quad \textit{and} \quad \mathtt{D} = \begin{pmatrix} -3/4 & 0 & -1/4 & 0 \\ 0 & -3/4 & 0 & -1/4 \end{pmatrix}.$$

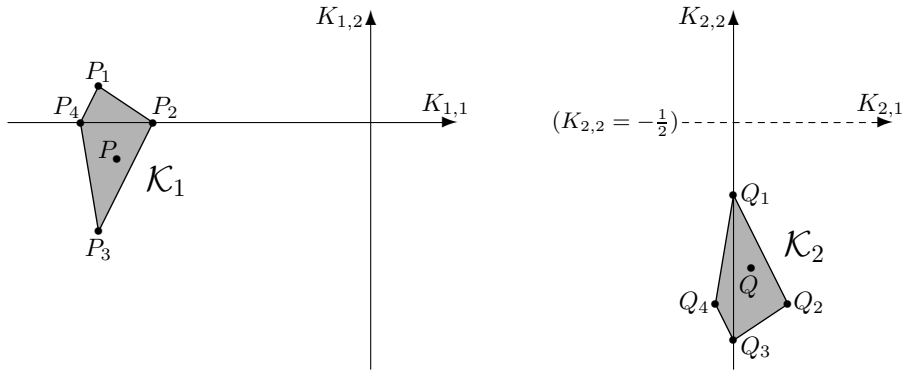

Figure 6.2: The feasibility regions of system (6.13) in the case considered in Example 27 .

As $\|\mathcal{N}\|_\infty = 3$ *, according to Corollary 8, we look for* $K \in \mathbb{R}^{2\times 2}$ *such that* $\|\mathtt{D} - K\mathtt{N}\|_\infty \leq 1 - \frac{3}{10} = \frac{7}{10}$ *. We have*

$$\mathtt{D} - K\mathtt{N} = \begin{pmatrix} -\frac{3}{4} - 2K_{1,1} & -K_{1,2} & -\frac{1}{4} - \frac{K_{1,2}}{2} & K_{1,1} \\ -2K_{2,1} & -\frac{3}{4} - K_{2,2} & -\frac{K_{2,2}}{2} & -\frac{1}{4} + K_{2,1} \end{pmatrix},$$

thus system (6.13) *takes the form of*

$$\begin{cases} \left|\frac{3}{4} + 2K_{1,1}\right| + |K_{1,2}| + \left|\frac{1}{4} + \frac{K_{1,2}}{2}\right| + |K_{1,1}| \leq \frac{7}{10} \\ |2K_{2,1}| + \left|\frac{3}{4} + K_{2,2}\right| + \left|\frac{K_{2,2}}{2}\right| + \left|K_{2,1} - \frac{1}{4}\right| \leq \frac{7}{10}\,. \end{cases}$$

The system is solved for $(K_{1,1}, K_{1,2}) \in \mathcal{K}_1 \subset \mathbb{R}^2$ *, where* $\mathcal{K}_1$ *is the quadrilateral whose vertices are*

$$P_1 = \left(-\tfrac{3}{8}\,,\,\tfrac{1}{20}\right), \quad P_2 = \left(-\tfrac{3}{10}\,,\,0\right), \quad P_3 = \left(-\tfrac{3}{8}\,,\,-\tfrac{3}{20}\right), \quad P_4 = \left(-\tfrac{2}{5}\,,\,0\right),$$

and for $(K_{2,1}, K_{2,2}) \in \mathcal{K}_2 \subset \mathbb{R}^2$ *, where* $\mathcal{K}_2$ *is the quadrilateral whose vertices are*

$$Q_1 = \left(0\,,\,-\tfrac{3}{5}\right), \quad Q_2 = \left(\tfrac{3}{40}\,,\,-\tfrac{3}{4}\right), \quad Q_3 = \left(0\,,\,-\tfrac{4}{5}\right), \quad Q_4 = \left(-\tfrac{1}{40}\,,\,-\tfrac{3}{4}\right),$$

see Fig. 6.2. One feasible choice for K *is*

$$K = \begin{pmatrix} -7/20 & -1/20 \\ 1/40 & -7/10 \end{pmatrix}$$

(which is identified by the points P *and* Q *represented in Fig. 6.2). For such a* K *we have* $\|\mathtt{D} - K\mathtt{N}\|_\infty = 27/40$ *and hence* $\|\mathcal{G}_K\|_\infty \leq \frac{3}{1-27/40} = \frac{120}{13} \simeq 9.2308$ *.*
Notice that, because the closed loop system is reachable and observable, then $A + BKC$ *is a Schur matrix and the resulting dynamics is* ℓ_∞*–stable.* ♣

6.2 ℓ_∞/ℓ_∞ small–gain for practical stability of multi–input systems

6.2.1 Practical stability analysis in ℓ_1 and mixed H_∞/ℓ_1 analysis

In this section, the practical stability analysis for feedback systems is studied in the ℓ_1 functional space. We first introduce a generalized notion of ℓ_∞–gain for a static nonlinearity, then we prove some practical stability results that are based on a small–gain condition in terms of the ℓ_∞–gain of the operators forming the feedback system. These results can be merged together with those from the analogous Section 5.3.1 , thus providing mixed H_∞/ℓ_1 analysis tools.

For a given function $\psi : \mathbb{R}^n \to \mathbb{R}^m$, let us introduce some quantities which will be useful for our analysis. For $\Omega \subseteq \mathbb{R}^n$, let

$$\mathscr{E}(\Omega) := \sup_{x\in\Omega} \|\psi(x)\|_\infty .$$

We always assume that the function ψ is *regular*, namely that if $\Omega \subset \mathbb{R}^n$ is bounded, then $\mathscr{E}(\Omega) < +\infty$.
In the particular case of $\Omega = Q_n(\Delta)$, $\Delta \geq 0$, we use the notation

$$\mathscr{E}(\Delta) := \sup_{x\in Q_n(\Delta)} \|\psi(x)\|_\infty .$$

The function $\mathscr{E}(\Delta)$ is non–decreasing with Δ , we can hence define the right continuous function

$$\mathscr{E}^+(\Delta) := \lim_{\epsilon\to 0^+} \mathscr{E}(\Delta+\epsilon) .$$

Definition 32 *For $\Delta > 0$, let the* generalized ℓ_∞–gain *of the function ψ be defined by*[5]

$$\gamma_e(\Delta) := \frac{\mathscr{E}^+(\Delta)}{\Delta/2} .$$

Theorem 12 (Small–gain in ℓ_1: (X_0, X_1, Ω)–stability analysis) *Let us consider a linear system*

$$\begin{cases} x(t+1) = Fx(t) + Be(t) \\ y(t) = Cx(t) \\ x \in \mathbb{R}^n,\ e \in \mathbb{R}^m,\ y \in \mathbb{R}^p, \end{cases}$$

where F is a Schur matrix. Let $\mathcal{G}^{(I)}$ be the input/state operator associated to the system. Assume that $q_e : \mathbb{R}^p \to \mathbb{R}^m$ is such that the closed–loop dynamics

$$x(t+1) = Fx(t) + Bq_e\big(Cx(t)\big) \tag{6.14}$$

is (X_0, X_1, Ω)–stable. Consider the function $\psi := q_e \circ C : \mathbb{R}^n \to \mathbb{R}^m$ and let $\Delta_1 := 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(\Omega)$. Then,

[5] In the definition of γ_e , $\mathscr{E}^+(\Delta)$ is divided by $\Delta/2$ because $x \in Q_n(\Delta) \Leftrightarrow \|x\|_\infty \leq \Delta/2$.

i) $\forall\,\Delta > \Delta_1$, *system* (6.14) *is* $\big(X_0, X_1, Q_n(\Delta)\big)$*–stable.*
Moreover, let $\gamma_e(\Delta)$ *be the generalized* ℓ_∞*–gain of the function* ψ *:*
ii) if

$$\|\mathcal{G}^{(I)}\|_\infty \cdot \gamma_e(\Delta_1) < 1\,, \tag{6.15}$$

then it is well–defined

$$\Delta_{\inf} := \begin{cases} \max\big\{\Delta < \Delta_1 \,|\, \|\mathcal{G}^{(I)}\|_\infty \cdot \gamma_e(\Delta) = 1\big\} & \text{if} \\ \quad \big\{\Delta < \Delta_1 \,|\, \|\mathcal{G}^{(I)}\|_\infty \cdot \gamma_e(\Delta) = 1\big\} \neq \emptyset & \\ 0 \quad \text{otherwise}\,, & \end{cases} \tag{6.16}$$

and $\forall\,\Delta_\star > \Delta_{\inf}$, *system* (6.14) *is* $\big(X_0, X_1, Q_n(\Delta_\star)\big)$*–stable.*

Proof. The proof is given below, after some remarks and some preliminary results. ■

In Theorem 12 , the closed loop dynamics (6.14) is assumed to be (X_0, X_1, Ω)–stable and, by taking advantage of a small–gain condition in the ℓ_1 space, a new stability property is deduced, that is $\big(X_0, X_1, Q_n(\Delta_\star)\big)$–stability. Compared with Theorem 9 in Section 5.3.1 , where practical stability properties can be asserted with no a priori assumptions on the practical stability of the closed loop dynamics, Theorem 12 is a weaker result. Nevertheless, the theorem can be used to supplement the practical stability analysis of a dynamics that has been proved to be practically stable through some other techniques. For instance, the combination of Theorem 12 with the analysis in H_∞ leads to a mixed H_∞/ℓ_1 analysis (see Corollary 9 below and Example 29 in next Section 6.2.2) . Moreover, this result, when used in the control synthesis perspective, gives rise to a special type of mixed H_∞/ℓ_1 control synthesis problem (whose exact formulation, and an example, is provided in Section 6.2.2) . It is worth noting that, if $Q_n(\Delta_\star) \subseteq \Omega$, then the application of the theorem allows one to improve the steady–state analysis (meaning that, according to the containment relation, the convergence of the trajectories to within a smaller neighborhood of the equilibrium is proved) . In general, this is not true. However, because system (6.14) is both (X_0, X_1, Ω)–stable and $\big(X_0, X_1, Q_n(\Delta_\star)\big)$–stable, then it is $\big(X_0, X_1, \Omega \cap Q_n(\Delta_\star)\big)$–stable. As $\Omega \cap Q_n(\Delta_\star) \subseteq \Omega$, then the application of Theorem 12 can only improve the practical stability analysis. The subsequent presentation and some examples further clarify the relevance of the contribution brought by this theorem.
The proof of Theorem 12 is based on the following

Lemma 19 (Main tool) *With the notation of Theorem 12, consider* $x(0) \in \mathbb{R}^n$ *and its evolution under the closed loop dynamics* (6.14) *. Let* $S \subseteq \mathbb{R}^n$ *be such that* $\mathscr{E}(S) < +\infty$ *and* $\exists\,\hat{t} \geq 0$ *such that* $\forall\, t \geq \hat{t}$, $x(t) \in S$ *. Then,* $\forall\,\Delta > 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(S)$, $\exists\, t_1 \geq 0$ *such that* $\forall\, t \geq t_1$, $x(t) \in Q_n(\Delta)$ *.*

Proof. To prove the result it is sufficient to show that $\limsup_{t\to+\infty} \|x(t)\|_\infty \leq \|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(S)$. First notice that, for $x(0) \in \mathbb{R}^n$ and $\vec{e} := \big\{\psi\big(x(t)\big)\big\}_{t\in\mathbb{N}}$, it holds that

$$\forall\, t \geq 0 \text{ and } \forall\, k \geq 0\,, \quad x(t+k) = F^k x(t) + (\vec{g}^{\,(I)} * \sigma^t \vec{e})(k)\,,$$

where $\vec{g}^{(I)}$ is the impulse response associated to system $\Sigma(F,B,I)$. Since $\forall\, t \geq \hat{t}$, $x(t) \in S$, then $\forall\, t \geq \hat{t}$, $\|e(t)\|_\infty \leq \mathscr{E}(S)$ or, equivalently, $\|\sigma^{\hat{t}}\vec{e}\|_\infty \leq \mathscr{E}(S)$. Thus,

$$\begin{aligned}
\limsup_{t\to+\infty} \|x(t)\|_\infty = &\quad \limsup_{k\to+\infty} \|x(\hat{t}+k)\|_\infty \leq \\
\leq &\quad \limsup_{k\to+\infty} \left(\|F^k x(\hat{t})\|_\infty + \|(\vec{g}^{(I)} * \sigma^{\hat{t}}\vec{e})(k)\|_\infty\right) \leq \\
\overset{(a)}{\leq} &\quad \|(\vec{g}^{(I)} * \sigma^{\hat{t}}\vec{e})\|_\infty \leq \\
\leq &\quad \|\mathcal{G}^{(I)}\|_\infty \|\sigma^{\hat{t}}\vec{e}\|_\infty \leq \\
\leq &\quad \|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(S)\,,
\end{aligned}$$

where in inequality (a) we used the fact that, as F is a Schur matrix, then

$$\lim_{k\to+\infty} \|F^k x(\hat{t})\|_\infty = 0\,.$$

■

To prove Theorem 12 we also need the following technical result:

Lemma 20 *With the notation of Theorem 12, for any fixed $\Delta_1 \geq 0$, consider the sequence $\{\Delta_k\}_{k\in\mathbb{N}\setminus\{0\}}$ defined by $\Delta_{k+1} := 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}^+(\Delta_k)$. Then, $\forall\, \hat{m} \in \mathbb{N}\setminus\{0\}$, the following property holds: $\forall\, \epsilon' > 0$, $\exists\, \{\epsilon_k\}_{k=1,\dots,\hat{m}}$, with $\epsilon_k > 0$ $\forall\, k = 1\dots,\hat{m}$, such that*

$$\Delta_{\hat{m}} < \Delta^+_{\hat{m}} < \Delta_{\hat{m}} + \epsilon'\,,$$

where

$$\begin{cases} \Delta^+_1 := \Delta_1 + \epsilon_1 \\ \Delta^+_{k+1} := 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(\Delta^+_k) + \epsilon_{k+1} \quad (\textit{for } 1 \leq k < \hat{m})\,. \end{cases} \tag{6.17}$$

Proof. See in Appendix A.5.3. ■

We are now ready for the proof of the theorem:

Proof of Theorem 12. To prove part *i*, since the closed loop dynamics (6.14) is already known to be (X_0, X_1, Ω)–stable, we have only to show that $\forall\, x(0) \in X_0$ and $\forall\, \Delta > \Delta_1$, $\exists\, t_1 \geq 0$ such that $\forall\, t \geq t_1$, $x(t) \in Q_n(\Delta)$. Such a property follows by the application of Lemma 19 with $S = \Omega$.

To prove part *ii*, let us first show that $\Delta_{\inf}$ is well–defined. To this end, for $\Delta \in [0, \Delta_1]$, consider

$$\phi(\Delta) := 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}^+(\Delta)\,. \tag{6.18}$$

The following properties hold for ϕ: $\phi(\Delta) \geq 0$ and it is a non–decreasing and right continuous function (because these properties hold for $\mathscr{E}^+$); moreover, by definition,

$$\phi(\Delta) < \Delta \iff \|\mathcal{G}^{(I)}\|_\infty \cdot \gamma_e(\Delta) < 1\,, \tag{6.19}$$

so that, by the small–gain assumption (6.15), $\phi(\Delta_1) < \Delta_1$. Hence, by Lemma 11 of Section 4.1, the sequence $\{\Delta_k\}_{k\in\mathbb{N}\setminus\{0\}}$ defined by $\Delta_{k+1} := \phi(\Delta_k)$ is non–increasing. By equation (4.4) of Lemma 11 and relation (6.19), it immediately follows that $\lim_{k\to+\infty} \Delta_k = \Delta_{\inf}$,

with $\Delta_{\inf}$ as defined in equation (6.16).
Now, let $\Delta_\star > \Delta_{\inf}$: to prove the $\big(X_0, X_1, Q_n(\Delta_\star)\big)$–stability, we construct a sequence $\{\Delta_k^+\}_{k=1,\dots,\hat m}$ such that $\Delta_{\hat m}^+ < \Delta_\star$ and, iterating Lemma 19, we show that, $\forall\, k = 1, \dots, \hat m$, any trajectory starting from X_0 eventually remains confined within $Q_n(\Delta_k^+)$. In detail, as $\lim_{k\to+\infty} \Delta_k = \Delta_{\inf}$, $\exists\, \hat m \in \mathbb{N}$ such that $\Delta_{\hat m} < \Delta_\star$. Let $\epsilon' := \Delta_\star - \Delta_{\hat m}$: by Lemma 20, $\exists\, \{\epsilon_k\}_{k=1,\dots,\hat m}$ (with $\epsilon_k > 0 \;\; \forall\, k = 1, \dots, \hat m$) such that

$$\Delta_{\hat m} < \Delta_{\hat m}^+ < \Delta_\star\,,$$

where the sequence $\{\Delta_k^+\}_{k=1,\dots,\hat m}$ is defined in equation (6.17). Applying Lemma 19 with $S = Q_n(\Delta_k^+)$, if $\exists\, t_k \geq 0$ such that $\forall\, t \geq t_k$, $x(t) \in Q_n(\Delta_k^+)$, then $\exists\, t_{k+1} \geq 0$ such that $\forall\, t \geq t_{k+1}$, $x(t) \in Q_n(\Delta_{k+1}^+)$. Since in part $\imath$ we have shown that this is true for $k = 1$, the thesis follows by iterating the application of Lemma 19. ■

Remark 28 *If system* (6.14) *is* (X_0, Ω)*–stable, then Theorem 12 states that the system is* $\big(X_0, X_0, Q_n(\Delta_\star)\big)$*–stable (for* $\Delta_\star > \Delta_{\inf}$*). The reason why the weakest notion of* $\big(X_0, X_0, Q_n(\Delta_\star)\big)$*–stability is provided by Theorem 12* $\big($*rather than* $\big(X_0, Q_n(\Delta_\star)\big)$*–stability, see Remark 1 in Section 2.3.1*$\big)$ *lies in the fact that this result is based only on the analysis of the forced component of the state, hence providing information only on the asymptotic properties of the trajectories. Since the transient behavior is not taken into account, the positive invariance of* $Q_n(\Delta_\star)$ *is not guaranteed.*

Theorem 12 can be combined with the practical stability results presented so far. For instance, Theorem 9 of Section 5.3.1 together with Theorem 12 yield a mixed H_∞/ℓ_1 stability analysis result:

Corollary 9 (Mixed H_∞/ℓ_1 analysis) *Under the assumptions of Theorem 9, let* $\psi := q_e \circ C$ *and* $\mathcal{G}^{(I)}$ *be the input/state operator associated to system* (5.14). *For* $r_2^2 \geq r_{\rm i}^2$, *let* $\mathscr{E}(r_2^2) := \mathscr{E}(\mathcal{E}_{P,r_2^2})$ *and*

$$\Delta_1 := \inf_{r_2^2 > r_{\rm i}^2} 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(r_2^2)\,. \tag{6.20}$$

Then,

ı) $\forall\, r_1^2 > r_{\rm i}^2$ *and* $\forall\, \Delta > \Delta_1$, *system* (5.15) *is* $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_n(\Delta)\big)$*–stable.*

Moreover, let $\gamma_e(\Delta)$ *be the generalized* ℓ_∞*–gain of the function* ψ*:*

ıı) if

$$\|\mathcal{G}^{(I)}\|_\infty \cdot \gamma_e(\Delta_1) < 1\,,$$

then $\forall\, \Delta_\star > \Delta_{\inf}$, *system* (5.15) *is* $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_n(\Delta_\star)\big)$*–stable, where* $\Delta_{\inf}$ *is defined in equation* (6.16).

Proof. Part $\imath$: by Theorem 9.$\imath\imath$, $\forall\, r_1^2 \geq r_2^2 > r_{\rm i}^2$, system (5.15) is $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\big)$–stable. Hence, by Theorem 12.$\imath$, $\forall\, \Delta > 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(r_2^2)$, system (5.15) is $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_n(\Delta)\big)$–stable. Since this property holds $\forall\, r_2^2 > r_{\rm i}^2$, then the $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_n(\Delta)\big)$–stability is guaranteed $\forall\, \Delta > \Delta_1$.

Part ii is simply a restatement of Theorem 12.ii suited to system (5.15). ■

The combination of Theorem 12 with the analysis result in the case of uniformly bounded q_e (see Proposition 14 in Section 5.3.1) provides the following

Corollary 10 (Uniformly bounded q_e) *Under the assumptions of Proposition 14, let $\psi := q_e \circ C$ and denote by $\mathcal{G}^{(I)}$ the input/state operator associated to system (5.14), then $\forall\, r^2 \geq r_{\mathrm{i}}^2$ and $\forall\, \Delta > 2\|\mathcal{G}^{(I)}\|_\infty E_0$, system (5.15) is $\big(\mathcal{E}_{P,r^2}, \mathcal{E}_{P,r^2}, Q_n(\Delta)\big)$–stable.*

Proof. It is a direct consequence of Proposition 14 and of Theorem 12.i: it is sufficient to notice that $\forall\, \Omega \subseteq \mathbb{R}^n$, it holds that $\mathscr{E}(\Omega) \leq E_0$. Indeed, $\forall\, x \in \mathbb{R}^n$, $\|q_e(Cx)\|_2 \leq E_0$ and $\forall\, z \in \mathbb{R}^m$, $\|z\|_\infty \leq \|z\|_2$. ■

As already noticed, Theorem 12 is a weaker result than Theorem 9, because practical stability properties are a priori assumed for the closed loop dynamics (6.14). A stronger result can be instead proved for single–input reachable systems. In this case, in fact, practical stability properties are derived by a small–gain condition in ℓ_1 without a priori stability assumptions on the closed loop dynamics. Moreover, the stronger notion of (X_0, Ω)–stability is ensured.

Proposition 19 (Small–gain in ℓ_1: (X_0, Ω)–stability analysis of single–input reachable systems) *Let us consider a single–input linear system*

$$\begin{cases} x(t+1) = Fx(t) + Be(t) \\ y(t) = Cx(t) \\ x \in \mathbb{R}^n, \quad e \in \mathbb{R}, \quad y \in \mathbb{R}^p, \end{cases} \tag{6.21}$$

where the pair (F, B) is assumed to be in the controller form (see equation (3.1) in Section 3.1.1). Let $z^n - f_n z^{n-1} - \cdots - f_2 z - f_1$ be the characteristic polynomial of F and suppose that $f := \sum_{i=1}^n |f_i| < 1$. For a given $q_e : \mathbb{R}^p \to \mathbb{R}$, consider the control law

$$e(t) = q_e\big(y(t)\big) \ :$$

the corresponding closed loop dynamics is

$$x(t+1) = Fx(t) + Bq_e\big(Cx(t)\big)\,. \tag{6.22}$$

Let $\psi := q_e \circ C : \mathbb{R}^n \to \mathbb{R}$ and denote by $\gamma_e(\Delta)$ the generalized ℓ_∞–gain of the function ψ. Then, the following properties hold for system (6.22):

i) $\forall\, \Delta > 0$ such that

$$\frac{\gamma_e(\Delta)}{1-f} \leq 1\,,$$

$Q_n(\Delta)$ is positively invariant;

ii) if $\Delta_0 > 0$ is such that

$$\frac{\gamma_e(\Delta_0)}{1-f} < 1\,, \tag{6.23}$$

then it is well–defined

$$\Delta_{\inf} := \begin{cases} \max\left\{\Delta < \Delta_0 \,|\, \frac{\gamma_e(\Delta)}{1-f} = 1\right\} & \text{if} \quad \left\{\Delta < \Delta_0 \,|\, \frac{\gamma_e(\Delta)}{1-f} = 1\right\} \neq \emptyset \\ 0 & \text{otherwise}\,, \end{cases} \tag{6.24}$$

and $\forall\, \Delta_\star \in]\Delta_{\inf}, \Delta_0]$*, the system is* $\big(Q_n(\Delta_0), Q_n(\Delta_\star)\big)$*–stable.*

Proof. For $\Delta \geq 0$, let $\phi(\Delta) := f\Delta + 2\mathscr{E}^+(\Delta)$. The following properties hold for ϕ: $\phi(\Delta) \geq 0$ and it is a non–decreasing and right continuous function (it is a consequence of $f \geq 0$ and of the analogous properties of $\mathscr{E}^+$); for $f \in [0, 1[$,

$$\phi(\Delta) < \Delta \iff \frac{\gamma_e(\Delta)}{1-f} < 1\,; \tag{6.25}$$

also, for $\Delta > 0$, $\phi(\Delta) \leq \Delta \Leftrightarrow \frac{\gamma_e(\Delta)}{1-f} \leq 1$ which is equivalent to

$$f\frac{\Delta}{2} + \mathscr{E}^+(\Delta) \leq \frac{\Delta}{2} \iff \frac{\gamma_e(\Delta)}{1-f} \leq 1\,. \tag{6.26}$$

To prove part $\imath$, let $\Delta > 0$ be such that $\frac{\gamma_e(\Delta)}{1-f} \leq 1$ and $x \in Q_n(\Delta)$: because the system is in controller form, the positive invariance of $Q_n(\Delta)$ is guaranteed by showing that $|x_n^+| \leq \frac{\Delta}{2}$ (see equation (3.4) in Section 3.1.1). As $x_n^+ = \sum_{i=1}^n f_i x_i + \psi(x)$, then

$$\begin{aligned} |x_n^+| \leq \;& f\|x\|_\infty + \mathscr{E}(\Delta) \leq \\ \leq \;& f\|x\|_\infty + \mathscr{E}^+(\Delta) \leq \\ \leq \;& f\tfrac{\Delta}{2} + \mathscr{E}^+(\Delta) \leq \\ \leq \;& \tfrac{\Delta}{2}\,, \end{aligned}$$

where the last inequality holds thanks to (6.26).

To prove part $\imath\imath$, let $\Delta_0 > 0$ be such that the small–gain condition (6.23) holds. In this case, we can apply Lemma 11 of Section 4.1: hence, the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ defined by $\Delta_{k+1} := \phi(\Delta_k)$ is non–increasing and, as in the proof of Theorem 12.$\imath\imath$, equation (4.4) and relation (6.25) imply that $\lim_{k\to+\infty} \Delta_k = \Delta_{\inf}$, with $\Delta_{\inf}$ as defined in equation (6.24).
Let $\Delta_\star \in]\Delta_{\inf}, \Delta_0]$: by Lemma 11, $\exists\, \hat{m} \in \mathbb{N}$ such that $\Delta_{\inf} \leq \Delta_{\hat{m}} < \Delta_\star$. Let us show that, $\forall\, \Delta \in [\Delta_{\inf}, \Delta_0]$, $Q_n(\Delta)$ is positively invariant. Indeed, $\phi(\Delta_{\inf}) = \Delta_{\inf}$ and, again by Lemma 11, $\forall\, \Delta \in]\Delta_{\inf}, \Delta_0]$, $\phi(\Delta) < \Delta$: the positive invariance then follows by relation (6.26) and part $\imath$. Now, to prove the $\big(Q_n(\Delta_0), Q_n(\Delta_\star)\big)$–stability, it is sufficient to show that, $\forall\, k = 0, \ldots, \hat{m} - 1$, system (6.22) is $\big(Q_n(\Delta_k), Q_n(\Delta_{k+1})\big)$–stable. This property holds true because $\forall\, k \in \mathbb{N}$, if $x(0) \in Q_n(\Delta_k)$, then $x(n) \in Q_n(\Delta_{k+1})$. Let us prove this fact: since the system is in controller form, it is sufficient to show that, $\forall\, x \in Q_n(\Delta_k)$, $|x_n^+| \leq \frac{\Delta_{k+1}}{2}$. Following the same arguments used to prove part $\imath$,

$$|x_n^+| \leq f\frac{\Delta_k}{2} + \mathscr{E}^+(\Delta_k) = \frac{\phi(\Delta_k)}{2} = \frac{\Delta_{k+1}}{2}\,.$$

■

Remark 29 *In Proposition 19, the ℓ_∞–norm of the input/state operator associated to system (6.21) does not explicitly appear in the small–gain condition (6.23). Nevertheless, $\frac{\gamma_e(\Delta_0)}{1-f} < 1$ is a small–gain condition in ℓ_1 because, by Proposition 17 in Section 6.1.1, it holds that $\|\mathcal{G}^{(I)}\|_\infty \leq \frac{1}{1-f}$.*

6.2.2 Practical stabilization of quantized input systems via ℓ_1–control: the hypercubes technique seen as the control synthesis in ℓ_1

Let us illustrate how the theory on the small–gain in the functional space ℓ_1 can be profitably used to address the practical stabilization problem for quantized input linear systems. Consider a system

$$\begin{cases} x^+ = Ax + Bu \\ x \in \mathbb{R}^n,\ u \in \mathcal{U} \subset \mathbb{R}^m, \end{cases} \tag{6.27}$$

where the pair (A, B) is supposed to be stabilizable and $\mathcal{U}$ is an assigned quantized set. As in Section 5.3.2, suitable adjustments of the proposed arguments allow one to deal with the case where also $\mathcal{U}$ can be chosen, nevertheless, the details of this problem are not discussed here.
We search for a constant feedback matrix $K \in \mathbb{R}^{m\times n}$ and an input quantizer $q_{\mathcal{U}} : \mathbb{R}^m \to \mathcal{U}$ so that the control law $u(x) = q_{\mathcal{U}}(Kx)$ practically stabilizes system (6.27). As usual, with the quantization error $q_e : \mathbb{R}^m \to \mathbb{R}^m$ defined by $q_e(y) = q_{\mathcal{U}}(y) - y$ (see Definition 4 in Section 2.1), the closed–loop dynamics induced by $u(x)$ is

$$x^+ = (A + BK)x + Bq_e(Kx)\,.$$

The practical stability analysis of this system can be done taking advantage of the results presented in the previous section, it is sufficient to let $F := A + BK$, $C := K$ and $\psi := q_e \circ K : \mathbb{R}^n \to \mathbb{R}^m$.
We propose two main solutions to the practical stabilization problem:

1. *Control synthesis in H_∞ and mixed H_∞/ℓ_1 closed loop analysis.*
   The synthesis of the controller is done according to the small–gain theorem in H_∞ (i.e., following Procedure 1). This yields a practically stable closed loop dynamics that can be analyzed with the mixed H_∞/ℓ_1 tools (namely, through Corollary 9 or Corollary 10, depending on the structure of $\mathcal{U}$).

2. *Mixed H_∞/ℓ_1 control synthesis and mixed H_∞/ℓ_1 closed loop analysis.*
   In view of Theorem 12, the control synthesis step of Procedure 1 (i.e., step 2) can be modified so as to guarantee that, not only the H_∞–norm of the closed loop system is below a desired threshold guaranteeing that the small–gain condition is met, but also the ℓ_∞–gain of the corresponding input/state operator is minimized (so as to reduce the size of the final hypercube $Q_n(\Delta_\star)$). Namely, step 2 is replaced with

2. Mixed H_∞/ℓ_1 control synthesis: *let* $\gamma_\infty \leq \frac{1}{\gamma_e}$, *find* $K \in \mathbb{R}^{m\times n}$ *such that*

$$K = \underset{\substack{X \in \mathbb{R}^{m\times n} \\ \text{such that} \\ \begin{cases} A+BX \text{ is Schur} \\ \|G_X\|_\infty < \gamma_\infty \end{cases}}}{\operatorname{argmin}} \|\mathcal{G}_X^{(I)}\|_\infty , \tag{6.28}$$

where $G_X(z) = X(zI - A - BX)^{-1}B$ *and* $\mathcal{G}_X^{(I)}$ *is the input/state operator of the system* $\Sigma(A+BX, B, I)$.

As in the former solution, the closed loop dynamics is then analyzed with the mixed H_∞/ℓ_1 tools.

In order to implement these solutions, one has to supplement the discussion on the implementation of the practical stabilization procedure presented in Section 5.3.2 with the aspects related to the ℓ_1 theory. Specifically, the first solution requires the analysis of the ℓ_∞–gain of the input/state operator associated to the closed loop dynamics and the analysis of the function ψ (in particular, of its generalized ℓ_∞–gain). For the second solution, it is also necessary to solve problem (6.28).
The analysis of the ℓ_∞–gain of the closed loop linear system has been extensively discussed in Section 6.1.1. Efficient numerical methods are also available (see [6, 58]). Similarly to the study of the external gain (see the implementation of step 1 of Procedure 1), the analysis of ψ is mainly a geometric issue that, in principle, can be done for any ψ (hence, for any input quantizer). However, for general input quantizers and large dimension of the input space, this analysis may be quite involved. At the end of this section, in Example 28, we explicitly analyze $\psi = q_e \circ K$ when q_e is the quantization error associated to a logarithmic quantization of $\mathbb{R}$. Finally, the one in equation (6.28) is a special kind of mixed H_∞/ℓ_1 control problem. In this thesis, we do not go into the details of this issue and we let it as an open point for further investigations. Here, we limit ourselves to mention that, in the framework of multi–objective control, there is a certain amount of literature dealing with mixed H_∞/ℓ_1 control problems (see [23, 37, 121]). Moreover, for the study of this problem, one can make the most of the equivalence between the H_∞ and the ℓ_1–norms of externally positive SISO systems (see Lemma 18). An associated open issue consists in the study of the the mixed H_∞/ℓ_1 control problem under the supplementary constraint that K is a feedback gain ensuring external positivity properties for the closed loop dynamics.
Details on the implementation of the two solutions can be found in Example 29 at the end of this section where a comparison between the two approaches is presented.

Both the stabilization methods that we proposed rely on the control synthesis in H_∞. Indeed, in order to apply Theorem 12, the closed loop dynamics must have a priori known practical stability properties and Theorem 9 does provide them. In accordance with the corresponding result on the practical stability analysis, the practical stabilization problem for single–input reachable systems, instead, can be addressed entirely relying on ℓ_1 theory. This result

is presented in Theorem 13 below and, not only it provides an interpretation in terms of a control problem in ℓ_1 of the stabilization technique based on the analysis of controlled invariant hypercubes presented in Theorem 6 of Section 4.1, but it also extends that approach to a wider class of controllers (rather than to the quantized deadbeat only).

In detail, let us consider the practical stabilization problem for system (6.27) where $m = 1$ and the pair (A, B) is reachable. Without loss of generality, we assume that the system is in controller form (see equation (3.1) of Section 3.1.1). If $K \in \mathbb{R}^{1\times n}$ is such that the matrix $F := A + BK$ satisfies $f := \sum_{i=1}^{n} |f_i| < 1$, where $z^n - f_n z^{n-1} - \cdots - f_2 z - f_1$ is the characteristic polynomial of F, then the practical stability properties of the closed loop dynamics arising from the control law $u(x) = q_{\mathcal{U}}(Kx)$ (for some input quantizer $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$) can be analyzed through Proposition 19. However, this approach has a drawback: in fact, in the small–gain condition $\frac{\gamma_e(\Delta)}{1-f} < 1$ (see equation (6.23)), the dependence on the control gain K is not limited to the term concerned with the ℓ_∞–gain of the ideal closed loop dynamics (i.e., $\frac{1}{1-f}$) but also involves the parameter that takes the quantization error into account (i.e., $\gamma_e(\Delta)$). When $q_{\mathcal{U}}$ is a nearest neighbor quantizer, it is possible to obtain a more general practical stabilization result where the small–gain condition $\frac{\gamma_e(\Delta)}{1-f} < 1$ is replaced by a similar condition but the dependence on K is restricted to the term taking the ideal closed loop dynamics into account. To state the result, we refer to the definition of $\rho(\Delta_0)$, $M(\Delta_0)$ and $m(\Delta_0)$ given in Section 3.1.1.

Theorem 13 (Small–gain in ℓ_1: (X_0, Ω)–stabilization of single–input reachable systems) *Consider system* (6.27), *assume* **A0** (*see equation* (3.1) *in Section 3.1.1*) *and that*[6] $\alpha := \sum_{i=1}^{n} |a_i| \geq 1$. *Let* $K \in \mathbb{R}^{1\times n}$ *be such that* $F := A + BK$ *satisfies* $f := \sum_{i=1}^{n} |f_i| < 1$. *Consider* $\Delta_0 > 0$ *such that*

$$\begin{cases} m(\Delta_0) < -\dfrac{\Delta_0}{2}(\alpha - 1) & \text{(6.29a)} \\ M(\Delta_0) > \dfrac{\Delta_0}{2}(\alpha - 1) & \text{(6.29b)} \\ \rho(\Delta_0) < (1 - f)\Delta_0\,, & \text{(6.29c)} \end{cases}$$

assume that[7] $\mathcal{U} = \mathcal{U}(\Delta_0)$ *and let* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}(\Delta_0)$ *be a nearest neighbor quantizer. Then it is well–defined*

$$\Delta_{\inf}(f) := \max\left\{\Delta < \Delta_0 \,|\, \rho(\Delta) = (1 - f)\Delta\right\} \tag{6.30}$$

and $\forall\, \Delta_\star \in]\Delta_{\inf}(f)\,,\, \Delta_0]$, *the control law* $u(x) = q_{\mathcal{U}}(Kx)$ *is* $\big(Q_n(\Delta_0), Q_n(\Delta_\star)\big)$*–stabilizing.*

Proof. The proof is reported in Appendix A.5.4. It is based on arguments similar to those we used to prove Theorem 6 in Section 4.1, as well as Theorem 12 and Proposition 19.*iii* in Section 6.2.1. ■

In conditions (6.29), f is the only term which is depending on the design parameter K. Conditions (6.29a-b) are concerned with the structure of the quantized control set $\mathcal{U}$. Thus,

[6] The case $\alpha < 1$ is trivial because, by Lemma 28 in Appendix A.5.1, the matrix A is Schur and $u(t) \equiv 0$ guarantees asymptotic stability.

[7] As far as the assumption $\mathcal{U} = \mathcal{U}(\Delta_0)$ is concerned, see the discussion in Remark 14 of Section 3.2.2.

similarly to the practical stabilization via the H_∞–control, the analysis of the nonlinearity due to the quantization error can be carried out apart from the problem of the design of the control gain K. Namely, if the quantized control set $\mathcal{U}$ is assigned, the search of K so that desired practical stability properties are ensured only consists of satisfying condition (6.29c) or, in other words, of finding K such that $\frac{1}{1-f}$, and hence the ℓ_∞–gain of the input/state operator associated to system (6.21), is sufficiently small.

Remark 30 (Optimality of the quantized deadbeat controller) *Let $\Delta_0 > 0$ be such that conditions (6.29) are satisfied for some $\hat{f} \in [0, 1[$. For such a Δ_0, by equation (6.30), $\Delta_{\inf}(f)$ is an increasing function of f. Hence, the size of the final invariant hypercube within which convergence is guaranteed by Theorem 13, starting from $Q_n(\Delta_0)$, can be reduced by choosing a control gain K making smaller the corresponding value of f. In this respect, the optimal choice of K, i.e., the one minimizing $\Delta_{\inf}(f)$, is the deadbeat controller $K = (\,-a_1 \;\; -a_2 \;\; \cdots \;\; -a_n\,)$: in this case, in fact, $f = 0$.*

Let us clarify the relation between the small–gain condition $\frac{\gamma_e(\Delta_0)}{1-f} < 1$ and conditions (6.29). Let $\Delta_0 > 0$ and assume that $\mathcal{U} = \mathcal{U}(\Delta_0)$: because $q_{\mathcal{U}}$ is a nearest neighbor quantizer and $\forall\, x \in Q_n(\Delta_0)$, $|Kx| \leq \|K\|_\infty \frac{\Delta_0}{2}$, it is straightforward to see that

$$\mathscr{E}^+(\Delta_0) = \max\left\{\frac{\rho(\Delta_0)}{2}\,,\, \|K\|_\infty\frac{\Delta_0}{2} - M(\Delta_0)\,,\, m(\Delta_0) + \|K\|_\infty\frac{\Delta_0}{2}\right\}.$$

This means that one condition in terms of $\mathscr{E}^+(\Delta_0)$ can be expressed as three conditions involving $\rho(\Delta_0)$, $M(\Delta_0)$ and $m(\Delta_0)$. Hence, taking advantage of relation (6.25),

$$\frac{\gamma_e(\Delta_0)}{1-f} < 1 \iff f\Delta_0 + 2\mathscr{E}^+(\Delta_0) < \Delta_0 \iff \left\{\begin{array}{l} m(\Delta_0) < -\frac{\Delta_0}{2}(f + \|K\|_\infty - 1) \\ M(\Delta_0) > \frac{\Delta_0}{2}(f + \|K\|_\infty - 1) \\ \rho(\Delta_0) < (1-f)\Delta_0\,. \end{array}\right.$$

Finally, as $\forall\, i = 1, \ldots, n$, $a_i = f_i - K_i$, then $\alpha \leq f + \|K\|_\infty$ so that the small–gain condition $\frac{\gamma_e(\Delta_0)}{1-f} < 1$ implies conditions (6.29). In particular, the latter conditions are less restrictive and the range of applicability of Theorem 13 is wider than the range of applicability of the practical stabilization technique which is based on a direct application of Proposition 19.

Example 28 (Analysis of ψ for logarithmic quantization of $\mathbb{R}$) *Let $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ be a logarithmic quantization of $\mathbb{R}$ with parameters (u_0, θ) (see Definition 8 in Section 2.1) and q_e be the corresponding quantization error. Let $K \in \mathbb{R}^{1\times n}$ and $\psi := q_e \circ K : \mathbb{R}^n \to \mathbb{R}$:*

ı) For a given $\mathbb{R}^{n\times n} \ni P > 0$, the function $\mathscr{E}(r_2^2) := \mathscr{E}(\mathcal{E}_{P,r_2^2})$ is continuous and, with

$$\mu_1 := \sqrt{r_2 K P^{-1} K'}\,,$$

it holds that

$$\mathscr{E}(r_2^2) = \left\{\begin{array}{ll} \mu_1 & \text{if} \quad \mu_1 < \frac{u_0}{2} \\ \max\left\{\frac{u_0}{2}\,,\, \gamma_e\frac{u_0(\theta+1)}{2}\theta^{n(\mu_1)}\,,\, |u_0\theta^{n(\mu_1)+1} - \mu_1|\right\} & \text{otherwise}\,, \end{array}\right. \tag{6.31}$$

where $\gamma_e := \frac{\theta-1}{\theta+1}$ *and* $n(\mu) := \left\lfloor \log_\theta \frac{2\mu}{u_0(\theta+1)} \right\rfloor$.

ıı) For $\Delta \geq 0$, *the function* $\mathscr{E}(\Delta) := \mathscr{E}\big(Q_n(\Delta)\big)$ *is continuous and, with*

$$\mu_2 := \|K\|_\infty \frac{\Delta}{2} ,$$

it holds that

$$\mathscr{E}(\Delta) = \begin{cases} \mu_2 & \text{if} \quad \mu_2 < \frac{u_0}{2} \\ \max\left\{ \frac{u_0}{2} \, , \, \gamma_e \frac{u_0(\theta+1)}{2} \theta^{n(\mu_2)} \, , \, |u_0 \theta^{n(\mu_2)+1} - \mu_2| \right\} & \text{otherwise} . \end{cases} \tag{6.32}$$

The proofs of these facts are reported in Appendix A.5.4 . ♣

Example 29 *Let us consider again the quantized input system studied in Example 22 of Section 5.3.3 :*

$$\begin{cases} x^+ = Ax + Bu = \begin{pmatrix} 0 & 1 \\ -1 & 5/2 \end{pmatrix} x + \begin{pmatrix} 1 \\ 2 \end{pmatrix} u \\ \\ u \in \mathcal{U} \subset \mathbb{R}, \end{cases}$$

where $\mathcal{U}$ *is a logarithmically quantized set with parameters* $(u_0, \theta) = (1 \, , \, 2)$.
For the solution of the practical stabilization problem, three cases are considered: in cases 1 and 2 , we perform the control synthesis in H_∞ *followed by a mixed* H_∞/ℓ_1 *analysis of the resulting closed loop dynamics. Specifically, we consider the controllers designed in cases 1 and 2 of Example 22 and we supplement the closed loop analysis with the results based on the small–gain in* ℓ_1 *(i.e., with Corollary 9) . In Case 3 instead, we perform a mixed* H_∞/ℓ_1 *control synthesis followed by a mixed* H_∞/ℓ_1 *analysis of the closed loop system.*

Let $K = (\, K_1 \;\; K_2 \,) \in \mathbb{R}^{1\times 2}$ *be such that* $A + BK$ *is Schur. As usual, we consider a control law* $u(x) = q_{\mathcal{U}}(Kx)$, *where* $q_{\mathcal{U}}$ *is a nearest neighbor quantizer. The transfer matrix* $G_K^{(I)}(z)$ *of the closed loop system* $\Sigma(A + BK, B, I)$ *can be easily computed to be*

$$G_K^{(I)}(z) = \begin{pmatrix} \frac{1}{z-(2+K_1+2K_2)} \\ \frac{2}{z-(2+K_1+2K_2)} \end{pmatrix} .$$

In particular, $2 + K_1 + 2K_2$ *is the only pole of the closed loop system. Therefore (see Section 6.1.1) ,*

$$\|\mathcal{G}_K^{(I)}\|_\infty = \frac{2}{1 - |2 + K_1 + 2K_2|} . \tag{6.33}$$

Case 1: *in case 1 of Example 22 , the control gain is* $K = (\, 0.6645 \;\; -1.3289 \,)$. *This* K *has been designed so that* $\|G_K\|_\infty$ *is close to* $\gamma_{\inf}$, *where*

$$\gamma_{\inf} = \inf\big\{ \, \|G_K\|_\infty \, | \, K \in \mathbb{R}^{1\times 2} \text{ is such that } A + BK \text{ is Schur} \big\} = 2 . \tag{6.34}$$

Namely, we have chosen an "authoritative" controller: for such a K, *we have* $\|G_K\|_\infty = 2.0066$ *and* $\|G_K\|_\infty \cdot \gamma_e = 0.6687$. *With*

$$P = \begin{pmatrix} 0.7990 & -0.9630 \\ -0.9630 & 1.9992 \end{pmatrix}$$

and $r_{\mathrm{i}}^2 = 4.0579$, *we have found that* $\forall\, r_1^2 \geq r_2^2 > r_{\mathrm{i}}^2$, *the resulting closed loop dynamics is* $\left(\mathcal{E}_{P,r_{\mathrm{i}}^2}, \mathcal{E}_{P,r_2^2}\right)$*–stable. The semi–axes of the final invariant ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *are*

$$\begin{cases} \mathrm{s}_1 = 3.9170 \\ \mathrm{s}_2 = 1.2655\,. \end{cases}$$

Let us complete the analysis through the application of Corollary 9 to the closed loop system (where, as usual, $F = A + BK$, $C = K$ *and* $\psi = q_e \circ K$*).*
Consider Corollary 9.ı: let us compute $\Delta_1 = \inf\limits_{r_2^2 > r_{\mathrm{i}}^2} 2\|\mathcal{G}_K^{(I)}\|_\infty \mathscr{E}(r_2^2)$ *as in equation* (6.20). *By equation* (6.33), *it holds that* $\|\mathcal{G}_K^{(I)}\|_\infty = 2.0133$. *As far as* $\mathscr{E}(r_2^2)$ *is concerned, we are in the right framework of Example 28: in particular, because* $\mathscr{E}(r_2^2)$ *is continuous and non–decreasing, then* $\inf\limits_{r_2^2 > r_{\mathrm{i}}^2} \mathscr{E}(r_2^2) = \mathscr{E}(r_{\mathrm{i}}^2)$. *Hence, by equation* (6.31),

$$\Delta_1 = 2\|\mathcal{G}_K^{(I)}\|_\infty \mathscr{E}(r_{\mathrm{i}}^2) = 2.0133\,.$$

In this case, it holds that $\Delta_1 = \|\mathcal{G}_K^{(I)}\|_\infty$ *just because* $\mathscr{E}(r_{\mathrm{i}}^2) = 1/2$. *We will see in next cases 2 and 3 that, in general, this is not true.*
Consider Corollary 9.ıı: let us check if the small–gain condition is satisfied. In Example 28, we have shown that $\mathscr{E}(\Delta)$ *is continuous, therefore* $\mathscr{E}^+(\Delta) = \mathscr{E}(\Delta)$ *and* $\gamma_e(\Delta) = \frac{\mathscr{E}(\Delta)}{\Delta/2}$. *By equation* (6.32), *one computes* $\mathscr{E}(\Delta_1) = 1/2$ *and* $\gamma_e(\Delta_1) = 1/\Delta_1$, *hence*

$$\|\mathcal{G}_K^{(I)}\|_\infty \cdot \gamma_e(\Delta_1) = 1\,.$$

Thus, Corollary 9.ıı cannot be applied.
Final result: with $u(x) = q_{\mathcal{U}}(Kx)$, $\forall\, r_1^2 > 4.0579$ *and* $\forall\, \Delta > 2.0133$, *the closed loop dynamics is* $\left(\mathcal{E}_{P,r_{\mathrm{i}}^2}, \mathcal{E}_{P,r_1^2}, Q_2(\Delta)\right)$*–stable.*
In order to appreciate the contribution to the closed loop analysis brought by the application of the ℓ_1 *theory, let us compare the diameters of the final hypercube* $Q_2(\Delta_1)$ *and of the final invariant ellipsoid* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$. *The half diagonal of* $Q_2(\Delta_1)$ *is* $\Delta_1/\sqrt{2} = 1.4236$ *whereas the largest semi–axis of* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *is* $\mathrm{s}_1 = 3.917$. *Fig. 6.3 provides a visual representation of such an improvement. It is also reported the simulation of a closed loop trajectory that bears evidence of the non–conservativeness of the obtained result. Notice that* $Q_2(\Delta_\star)$ *is not positively invariant but, eventually, the trajectories are guaranteed to remain confined therein (see Remark 1 in Section 2.3.1).*

Case 2: *in case 2 of Example 22, the control gain is* $K = (\,0.5294 \;\; -1.0588\,)$. *This* K *has been designed so that* $\|G_K\|_\infty \cdot \gamma_e = 0.9$ *(*$\|G_K\|_\infty = 2.7$*), namely we have chosen a less*

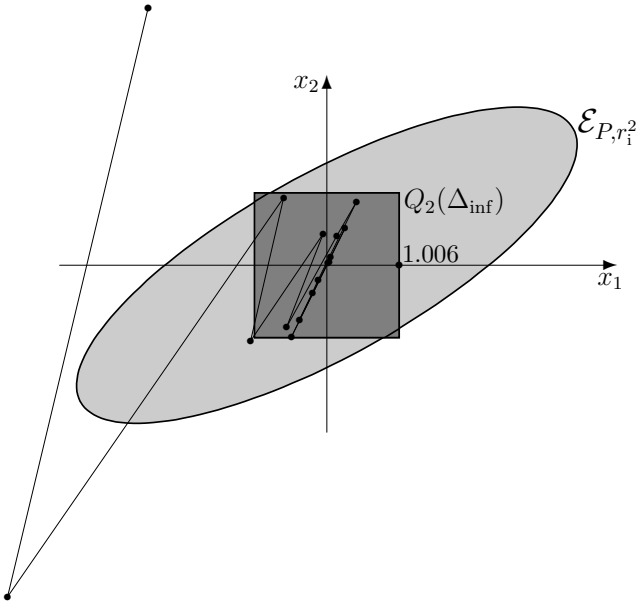

Figure 6.3: Comparison between the final invariant ellipsoid $\mathcal{E}_{P,r_i^2}$ and the final hypercube $Q_2(\Delta_1)$ in case 1 of Example 29 (control synthesis in H_∞ and mixed H_∞/ℓ_1 closed loop analysis). Representation of the trajectory starting from $x(0) = (-2.48 \;\; 3.57)$.

"authoritative" controller. With

$$P = \begin{pmatrix} 0.3957 & -0.7764 \\ -0.7764 & 1.5638 \end{pmatrix}$$

and $r_i^2 = 1.0420 \cdot 10^4$, *we have found that* $\forall\, r_1^2 \geq r_2^2 > r_i^2$, *the resulting closed loop dynamics is* $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\right)$*–stable. The semi–axes of the final invariant ellipsoid* $\mathcal{E}_{P,r_i^2}$ *are*

$$\begin{cases} \mathrm{s}_1 = 1126.8 \\ \mathrm{s}_2 = 73.1\,. \end{cases}$$

Let us complete the analysis by taking advantage of Corollary 9. By equation (6.33), *it holds that* $\|\mathcal{G}_K^{(I)}\|_\infty = 3.4001$. *With the same arguments of case 1, the application of Corollary 9.ıı provides*

$$\Delta_1 = 2\|\mathcal{G}_K^{(I)}\|_\infty \mathscr{E}(r_i^2) = 153.0816\,.$$

As for the application of Corollary 9.ıı, by equation (6.32), *one computes* $\mathscr{E}(\Delta_1) = 17.0424$ *and* $\gamma_e(\Delta_1) = 0.2227$. *Therefore,*

$$\|\mathcal{G}_K^{(I)}\|_\infty \cdot \gamma_e(\Delta_1) = 0.7571 < 1$$

and the small–gain condition in ℓ_1 *is met. In order to determine* $\Delta_{\inf}$ $\big($*see equation* (6.16)$\big)$, *according to Lemma 11, we can follow an iterative procedure: the sequence defined by* $\Delta_{k+1} =$

$2\|\mathcal{G}_K^{(I)}\|_\infty \mathscr{E}(\Delta_k)$ $\big($*see equation* (6.18)$\big)$ *converges to*

$$\Delta_{\inf} = 108.8029\,.$$

Final result: with $u(x) = q_{\mathcal{U}}(Kx)$, $\forall\, r_1^2 > 1.0420 \cdot 10^4$ *and* $\forall\, \Delta_\star > 108.8029$, *the closed loop dynamics is* $\big(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_2(\Delta_\star)\big)$*–stable.*
Not only the final hypercube $Q_2(\Delta_{\inf})$ *is such that*[8] $Q_2(\Delta_{\inf}) \subset \mathcal{E}_{P,r_{\mathrm{i}}^2}$, *but also the comparison between the largest semi–axis* $\mathrm{s}_1 = 1126.8$ *of* $\mathcal{E}_{P,r_{\mathrm{i}}^2}$ *and the half diagonal* $\Delta_{\inf}/\sqrt{2} = 76.9353$ *of* $Q_2(\Delta_{\inf})$ *gives evidence of the improvement to the stability analysis brought by the application of the* ℓ_1 *theory.*

Case 3: *let us solve the problem through mixed* H_∞/ℓ_1 *control synthesis.*
In this example, problem (6.28) *is equivalent to a simpler problem where* $G_X(z)$ *and* $\mathcal{G}_X^{(I)}(z)$ *can be replaced with a unique SISO operator. In fact: by equation* (6.33), $\|\mathcal{G}_K^{(I)}\|_\infty = \|\mathcal{G}_K^{(\mathtt{siso})}\|_\infty$, *where* $G_K^{(\mathtt{siso})}(z) := \frac{2}{z-(2+K_1+2K_2)}$, *and* $G_K(z) = \frac{-2}{z-(2+K_1+2K_2)} = -G_K^{(\mathtt{siso})}(z)$.
Therefore, the one in equation (6.28) *is equivalent to the following problem: for* $\gamma_\infty \leq \frac{1}{\gamma_e}$, *find* $K \in \mathbb{R}^{1\times 2}$ *such that*

$$K = \underset{\substack{X \in \mathbb{R}^{1\times 2} \quad \textit{such that} \\ \begin{cases} A+BX \textit{ is Schur} \\ \|G_X^{(\mathtt{siso})}\|_\infty < \gamma_\infty \end{cases}}}{\operatorname{argmin}} \|\mathcal{G}_X^{(\mathtt{siso})}\|_\infty\,. \tag{6.35}$$

Furthermore, because the system is SISO, by Lemma 18 it holds that $\|G_X^{(\mathtt{siso})}\|_\infty \leq \|\mathcal{G}_X^{(\mathtt{siso})}\|_\infty$.
Hence, $\forall\, \gamma_\infty > \gamma_{\inf} = 2$ $\big($*see equation* (6.34)$\big)$, *the solution to problem* (6.35) *is given by*

$$K = \underset{\substack{X \in \mathbb{R}^{1\times 2} \quad \textit{such that} \\ A+BX \textit{ is Schur}}}{\operatorname{argmin}} \|\mathcal{G}_X^{(\mathtt{siso})}\|_\infty\,.$$

According to equation (6.33), *a solution is* $K = (\,0 \;\; -1\,)$ *which yields*

$$\|\mathcal{G}_K^{(\mathtt{siso})}\|_\infty = \|G_K^{(\mathtt{siso})}\|_\infty = 2 = \gamma_{\inf}\,.$$

Actually, because it is a first order system, it holds that $\|\mathcal{G}_K^{(\mathtt{siso})}\|_\infty = \|G_K^{(\mathtt{siso})}\|_\infty$ *for any stabilizing* $K \in \mathbb{R}^{1\times 2}$.
For such a K, *let us analyze the closed loop dynamics induced by* $u(x) = q_{\mathcal{U}}(Kx)$. *First, according to the* H_∞ *analysis we find that, with*

$$P = \begin{pmatrix} 3.7165 & -1.7587 \\ -1.7587 & 2.0785 \end{pmatrix}$$

[8] This can be easily verified as $Q_2(\Delta) \subset \mathcal{E}_{P,r^2} \;\Leftrightarrow\; \frac{\Delta^2}{4}\big(P_{1,1} + P_{2,2} + 2|P_{2,1}|\big) \leq r^2$.

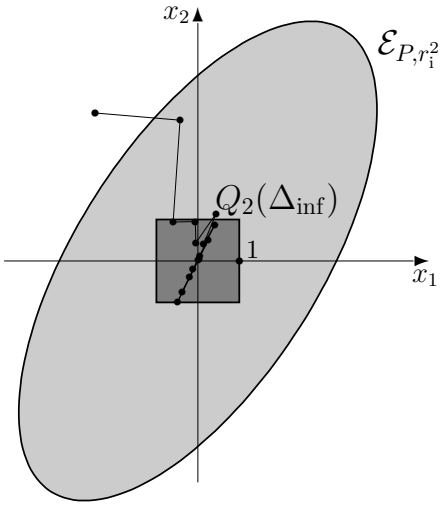

Figure 6.4: Comparison between the final invariant ellipsoid $\mathcal{E}_{P,r_i^2}$ and the final hypercube $Q_2(\Delta_{\inf})$ in case 3 of Example 29 (mixed H_∞/ℓ_1 control synthesis and mixed H_∞/ℓ_1 closed loop analysis). Representation of the trajectory starting from $x(0) = (-2.48 \;\; 3.57)$.

and $r_i^2 = 41.7306$, $\forall\, r_1^2 \geq r_2^2 > r_i^2$, *the closed loop system is* $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_2^2}\right)$*–stable. The semi–axes of the final invariant ellipsoid* $\mathcal{E}_{P,r_i^2}$ *are*

$$\begin{cases} s_1 = 6.6017 \\ s_2 = 2.9371\,. \end{cases}$$

The completion of the analysis with the ℓ_1 *theory yields the following results:*

$$\Delta_1 = 2\|\mathcal{G}_K^{(I)}\|_\infty \mathscr{E}(r_i^2) = 7.1457\,;$$

the small–gain condition in ℓ_1 *is satisfied because*

$$\|\mathcal{G}_K^{(I)}\|_\infty \cdot \gamma_e(\Delta_1) = 0.5598 < 1\,;$$

finally, with the iterative method described in case 2, we compute

$$\Delta_{\inf} = 2\,.$$

Final result: with $u(x) = q_{\mathcal{U}}(Kx)$, $\forall\, r_1^2 > 41.7306$ *and* $\forall\, \Delta_\star > 2$, *the closed loop dynamics is* $\left(\mathcal{E}_{P,r_1^2}, \mathcal{E}_{P,r_1^2}, Q_2(\Delta_\star)\right)$*–stable.*
Also in this case, $Q_2(\Delta_{\inf}) \subset \mathcal{E}_{P,r_i^2}$. *The comparison between the largest semi–axis* $s_1 = 6.6017$ *of* $\mathcal{E}_{P,r_i^2}$ *and the half diagonal* $\Delta_{\inf}/\sqrt{2} = \sqrt{2}$ *of* $Q_2(\Delta_{\inf})$ *gives further evidence of the improvement brought by the application of the* ℓ_1 *theory. This is shown also in Fig. 6.4. In the figure, the simulation is reported of the closed loop trajectory corresponding to the same*

initial condition considered in case 1 : the comments we made to Fig. 6.3 are valid also for this case.
Notice also that, compared with the previous cases, the mixed H_∞/ℓ_1 control synthesis yields the best stability result, that is the size of the final hypercube is minimized (even if there is only a slight improvement with respect to case 1) . Actually, taking advantage of the fact that the non–reachable dynamics of the system is vanishing and applying Theorem 3 of Section 3.1.2 to the first order system describing the reachable dynamics, it is not difficult to see that $Q_2(2)$ is indeed the smallest ball in the infinity norm within which can be ultimately bounded the trajectories of the system. ♣

Chapter 7

Performance vs complexity

In Section 2.4, an analysis was presented on the mutual dependence between *performance* of a closed loop system and *complexity* of the corresponding quantized controller. The notion of complexity of a quantizer has been introduced with reference to the problem of the control under communication constraints. If the plant can exchange information with the controller through a finite rate (noiseless) communication channel (e.g., because the plant and the controller are remotely located, see Fig. 7.1), then the need rises for the encoding of the variables into symbols suited for transmission over the channel. In order to reduce performance deterioration, we want such an encoding to be so that transmission delays are minimized. In this respect, as discussed at the end of Section 2.4, suitable analysis tools should take the statistics of the symbols to encode into account. The study of the relations between performance and complexity in this probabilistic framework is indeed the subject of this chapter.

We consider scalar linear systems and we analyze the dynamics of probability distributions when the system is controlled by a static quantizer. The probabilistic notion of practical stability considered in this chapter is the so called *mean–square practical convergence*. This property amounts to ensuring that, for any initial distribution belonging to a specified class $\mathcal{P}$, the energy of the evolved distribution definitively stays below a desired threshold. Many quantized control strategies achieving this target can be found in the literature [29, 131, 17, 39, 45, 108], as well as in the previous chapters of this thesis. In this chapter, instead, the focus is on providing the theoretical tools for the analysis of the achievable closed loop performance according to the complexity of the quantizer. In such a probabilistic framework, it is part of the problem to identify the suitable performance and complexity measures.
The study is developed so as to include the analysis of distributions with unbounded support. There are two main reasons for this choice: first, it allows us to treat standard cases such as the analysis of the evolution of Gaussian distributions; secondly, it makes the developed theory ready for extensions to the case where also noise terms affecting the system are considered. In fact, even under the assumption that the initial distribution has a bounded support, if the presence of unbounded noise terms is included in the model (e.g., a Gaussian noise), then

the evolving distributions have an unbounded support. Taking distributions with unbounded support into consideration poses interesting issues from both the theoretical and the practical point of view. In fact, whatever the control goal, an essential requirement is that the energy (i.e., the mean–square value) of the distributions remains bounded during the evolution. As it has been clarified in [89] , when dealing with open loop unstable plants and distributions with unbounded support, in order to keep bounded the energy, control laws taking infinite values are necessary. Because we consider static controllers (as opposed to time–varying techniques widely studied in the literature [123, 17, 125, 88, 89]) , infinite symbols are necessary to encode the control values. This poses technical problems concerned with transmission delays caused by the presence of arbitrarily long coding sequences, as well as theoretical questions on the definition of a proper complexity measure for a controller (indeed, a cardinality function as the one considered in Section 2.4 is no more meaningful) . The theoretical framework allowing us to address these issues in a formal mathematical way is provided by the *Information theory*. In particular, the measure of the coding complexity is a basic issue in Information theory and its characterization is strictly related to the notion of *entropy* of a probability distribution.

Let us summarize the main contributions of this chapter. First, the complexity of a quantizer is defined in terms of an entropy–like function $\mathcal{H}$, the so called "coding complexity function" . This function is defined through the solution of an infinite dimensional optimization problem to which an explicit solution is not available. However, a detailed analysis of this function is possible because a nice relation is proved between $\mathcal{H}$ and the Laplace transformation of an easily computable function providing a geometric description of the quantizer. The use of the Laplace transformation theory allows us to carry out the asymptotic analysis of the coding complexity function for a wide range of quantizers. Secondly, a lower bound on the achievable asymptotic value of the energy under an assigned quantization is found. This result, which establishes a relation between steady–state performance and complexity, is carried out by means of two Information theoretical inequalities: one that quantifies the counteraction to the increase of entropy, due to open loop instability, that can be obtained

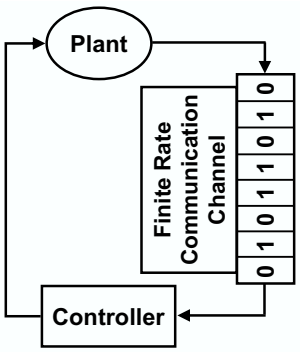

Figure 7.1: The control scheme considered in this chapter.

through a quantized controller; the other one that relates the energy to the entropy of a distribution. As for the analysis of performance in the transient behavior, we first show that if monotonic decrease of the energy is desired, then the complexity of the quantization must be at least that of a logarithmic quantizer. In the case of logarithmic quantizations, an asymptotic analysis of the relations between complexity and performance in the transient behavior is then provided. Finally, the applicability of the developed tools is borne out by their use to analyze performance and complexity for the class of so called nested quantizers, introduced in Section 7.5 .
We remark that the case of scalar linear systems is considered because it already contains the basic difficulties one encounters for more complex systems.

The chapter is organized as follows: in Section 7.1 , basic facts on the notion of entropy in Information theory are recalled. The considered models as well as the parameters to measure performance and complexity are defined in Section 7.2 . Subsequent Section 7.3 is devoted to the analysis of the complexity measure function and includes the study of the main quantizers considered in the previous chapters of the thesis. The relations between complexity and performance are studied in Section 7.4 . In Section 7.5 , the analysis for nested quantizers is presented.

Remark 31 (Notation) *For the notation and terminology we refer to the homonymous Section 1.5 in the introduction of the thesis, in particular to the paragraph "probability". Here, we stress that, differently from the usual convention in Information theory (where logarithms are in the base* 2 *), if not otherwise stated, the logarithms are in the base* e *.*

7.1 The entropy in Information theory

Let us briefly recall some basic facts on the notion of entropy in Information theory. The presentation is limited to the properties instrumental for the subsequent presentation, a more comprehensive treatment can be found in [26, 52] .

Definition 33 *Let* $X : \Omega \to \mathcal{X}$ *be a random variable taking values in a countable set* $\mathcal{X}$ *and let* $\mathcal{X} \ni x \mapsto p_x \in [0\,,\,1]$ *be its probability distribution. The* discrete entropy *of* X *is defined by*

$$H(X) := \sum_{x \in \mathcal{X}} -p_x \log p_x \,.$$

The entropy of a random variable only depends on its distribution, therefore, more in general, for a distribution defined on a countable set $\mathcal{X} \ni x \overset{\mathbf{p}_\mathcal{X}}{\longmapsto} p_x \in [0\,,\,1]$, we let $H(\mathbf{p}_\mathcal{X}) := -\sum_{x \in \mathcal{X}} p_x \log p_x$.
By a discrete random *source* we mean a sequence $\{X_i\}_{i \in \mathbb{N} \setminus \{0\}}$ of independent and identically distributed random variables defined on some probability space Ω and taking values in a finite set $\mathcal{X}$. Let $\mathcal{X} \ni x \mapsto p_x \in [0\,,\,1]$ be the probability distribution of any of the random variables defining the process and denote by $H(X)$ the discrete entropy of such a distribution:

$H(X)$ is referred to as the entropy of the source. We think of this process as the model for a device generating symbols according to some statistics. This process is the subject of the remaining part of this section.
We are interested in the following problem: suppose that another finite set of symbols $\mathcal{A}$ is given and that strings of symbols drawn by the source are to be transformed (say, *encoded*) into strings of symbols from the alphabet $\mathcal{A}$. We want this transformation to be invertible and so that the average length of the strings used to encode the symbols of the source is minimized. This problem is one of the basic issues raising in communication theory, where the symbols generated by some source (e.g., letters from the English alphabet or the elements $\{\mathcal{C}_u\}_{u\in\mathcal{U}}$ considered in Section 2.4) are to be transmitted through a communication bus capable of handling symbols from the alphabet $\mathcal{A}$ (e.g., a digital communication channel where $\mathcal{A}=\{0,1\}$). If only a finite number of symbols per unit of time can be transmitted over the channel (namely, communication happens at a finite rate), then the average length of the strings has to be minimized to reduce transmission delays.
Let us introduce the formal definitions and the main result concerned with the described problem.

Definition 34 *Let $\mathcal{A}$ be a finite set and consider $\mathcal{A}^* := \bigcup_{n\in\mathbb{N}\setminus\{0\}} \mathcal{A}^n$. The elements of $\mathcal{A}^*$ are called* strings *of symbols in the alphabet $\mathcal{A}$ and are denoted by $\bar{a}$. For $\bar{a}\in\mathcal{A}^*$, there exists a unique $n\in\mathbb{N}$ such that $\bar{a}\in\mathcal{A}^n$: such an n is called the* length *of the string $\bar{a}$ and it is denoted by $\mathrm{len}(\bar{a})$.*
Let $\mathcal{X}$ be a finite set. A map $c:\mathcal{X}\to\mathcal{A}^*$ is called a code *for the set of symbols $\mathcal{X}$. The elements of the set $c(\mathcal{X})$ are called* codewords. *For $\bar{x}=(x_1,\dots,x_n)\in\mathcal{X}^*$, let $\bar{c}(\bar{x}) := \big(c(x_1),\dots,c(x_n)\big)\in\mathcal{A}^*$. A code c is said to be* uniquely decodable *iff $\bar{c}:\mathcal{X}^*\to\mathcal{A}^*$ is an injective function.*
Let $X:\Omega\to\mathcal{X}$ be a random variable and let $\mathcal{X}\ni x\mapsto p_x\in[0\,,1]$ be its probability distribution. For a given code $c:\mathcal{X}\to\mathcal{A}^*$, let the average length *of the codewords be $\mathbb{E}\big[\mathrm{len}\big(c(X)\big)\big] = \sum_{x\in\mathcal{X}} p_x\cdot\mathrm{len}\big(c(x)\big)$.*

Proposition 20 (Source coding theorem) *Let $X:\Omega\to\mathcal{X}$ be a random variable taking values in a finite set and let $\mathcal{X}\ni x\mapsto p_x\in[0\,,1]$ be its probability distribution. If $c:\mathcal{X}\to\mathcal{A}^*$ is a decodable code, then*

$$\mathbb{E}\big[\mathrm{len}\big(c(X)\big)\big] \geq \frac{H(X)}{\log(\#\,\mathcal{A})}\,.$$

Moreover, if $\forall\, x\in\mathcal{X}$, $p_x>0$, then there exists a decodable code c such that

$$\mathbb{E}\big[\mathrm{len}\big(c(X)\big)\big] \leq \frac{H(X)}{\log(\#\,\mathcal{A})}+1\,.$$

Proof. See, e.g., [26] or [52]. ■

Thus, except for the constant factor $1/\log(\#\,\mathcal{A})$, which could be incorporated in the definition of the entropy by taking the logarithm to the base $\#\,\mathcal{A}$, the entropy of X fixes a lower

bound on the minimal achievable average length of the codewords. Such a lower bound can be approached if the code is defined on $\mathcal{X}^n$ (for sufficiently large n) namely, if a so called "block coding" of the source is done instead of considering the symbols one by one. In fact, for $n \neq 0$, consider

$$\begin{array}{rccl} X^{(n)}: & \Omega & \to & \mathcal{X}^n \\ & \omega & \mapsto & \big(X_1(\omega), \dots, X_n(\omega)\big)\,. \end{array}$$

Proposition 20, applied to the random variable $X^{(n)}$, ensures the existence of a coding $c : \mathcal{X}^n \to \mathcal{A}^*$ such that $\mathbb{E}\big[\mathrm{len}\big(c(X^{(n)})\big)\big] \leq \frac{H(X^{(n)})}{\log(\#\mathcal{A})} + 1$. By the independence of the $\{X_i\}_{i\in\mathbb{N}\setminus\{0\}}$ and standard properties of the entropy (see [26]), it holds that $\mathbb{E}\big[\mathrm{len}\big(c(X^{(n)})\big)\big] \leq n\frac{H(X)}{\log(\#\mathcal{A})} + 1$. Thus, dividing by n both sides of the inequality, we have

$$\bar{L}_n := \frac{\mathbb{E}\big[\mathrm{len}\big(c(X^{(n)})\big)\big]}{n} \leq \frac{H(X)}{\log(\#\mathcal{A})} + \frac{1}{n}\,.$$

$\bar{L}_n$ represents the minimal average length of the codewords per symbol of the source. Hence, by coding sufficiently long blocks (i.e., for sufficiently large n), such a length can be made as close as desired to the entropy of the source.
The code minimizing the average length of the codewords is the so called Huffman's code (see [52]). It can be algorithmically constructed by the knowledge of the distribution of the X_i's and it is such that long codewords are assigned to the least probable symbols and vice versa. Actually, for our purposes, it is enough to know the minimal average length of the strings rather than the code achieving such a minimum. Thus, entropy of the source contains all the relevant information.

We are now interested in counting the codewords needed to encode a discrete source. The number of codewords needed to encode the random variable $X^{(n)}$ is $(\#\mathcal{X})^n$. However, this number does not take the statistics of the symbols into account. A fair way to include statistics in the count of the codewords is provided by the so called *asymptotic equipartition property*. In details, let $\mathbf{p}_{X^{(n)}}$ be the distribution of $X^{(n)}$: by the independence assumption, it holds that

$$\mathcal{X}^n \ni (x_1, \dots, x_n) \mapsto \mathbf{p}_{X^{(n)}}\big((x_1, \dots, x_n)\big) = \Pi_{i=1}^n p_{x_i}\,.$$

In correspondence to the given process, consider the function

$$\begin{array}{rccl} W: & \mathcal{X} & \to & \mathbb{R}^+ \\ & x & \mapsto & -\log p_x \end{array}$$

and the sequence of random variables $\big\{W(X_i)\big\}_{i\in\mathbb{N}\setminus\{0\}}$. It holds that

$$\mathbb{E}[W(X_i)] = \sum_{x\in\mathcal{X}} p_x \cdot W(x) = H(X)\,.$$

Consider the sequence of random variables $\big\{A_W^{(n)}\big\}_{n\in\mathbb{N}\setminus\{0\}}$ defined by $A_W^{(n)} := \frac{\sum_{i=1}^n W(X_i)}{n}$. By the weak law of the large numbers [50], it holds that

$$\lim_{n\to+\infty} A_W^{(n)} = \mathbb{E}[W(X_i)] = H(X)\,,$$

where convergence is in probability. This limit can be equivalently expressed in the following way: let

$$T_\epsilon^{(n)} := \Big\{(x_1,\dots,x_n)\in\mathcal{X}^n \,\Big|\, \big|\tfrac{1}{n}\textstyle\sum_{i=1}^n W(x_i) - H(X)\big| < \epsilon\Big\} =$$

$$= \Big\{(x_1,\dots,x_n)\in\mathcal{X}^n \,\Big|\, e^{-n\left(H(X)+\epsilon\right)} < \mathbf{p}_{X^{(n)}}\big((x_1,\dots,x_n)\big) < e^{-n\left(H(X)-\epsilon\right)}\Big\}, \quad (7.1)$$

then $\forall\,\epsilon>0$ and $\forall\,\delta>0$, $\exists\,\hat{n}$ such that $\forall\, n\geq\hat{n}$,

$$\mathbf{p}_{X^{(n)}}\big(T_\epsilon^{(n)}\big) > 1-\delta\,.$$

Thus, for sufficiently large n, most of the sequences of length n drawn by the source belong to the set $T_\epsilon^{(n)}$: these sequences are referred to as *typical*. Moreover, equation (7.1) expresses the fact that typical sequences are approximately equiprobable, say $\forall\,(x_1,\dots,x_n)\in T_\epsilon^{(n)}$, $\mathbf{p}_{X^{(n)}}\big((x_1,\dots,x_n)\big)\simeq e^{-nH(X)}$. These two properties yield

$$\#\,T_\epsilon^{(n)} \simeq e^{nH(X)}\,. \quad (7.2)$$

This means that, for sufficiently large n, the behavior of the discrete random source made of the blocks of length n of the original process can be approximated by a sequence of independent and uniformly distributed random variables taking values in a set made of $e^{nH(X)}$ symbols (this is a formulation of the "asymptotic equipartition property"). Taking the n–th root in both sides of equation (7.2), we have

$$\bar{N}_n := \sqrt[n]{\#\,T_\epsilon^{(n)}} \simeq e^{H(X)}\,,$$

where $\bar{N}_n$ represents the average number of codewords needed to encode each variable X_i of the source.

Notice that, if the random variables X_i are not uniformly distributed, then $H(X) < \log(\#\,\mathcal{X})$ (see [26]) and

$$\frac{\#\,T_\epsilon^{(n)}}{\#\,\mathcal{X}^n} \simeq \frac{e^{nH(X)}}{e^{n\log(\#\,\mathcal{X})}} \to 0 \quad \text{for } n\to+\infty\,.$$

That is, for large n, typical sequences are a negligible portion of the total number of sequences drawn by the source but they account for most of the probability: this gives a further motivation to the choice of counting the codewords needed to encode a discrete source through $\#\,T_\epsilon^{(n)}$ instead of $\#\,\mathcal{X}^n$.

To recap, the entropy $H(X)$ of the source is a measure of the minimal average length of the codewords among decodable codes of the source. Moreover, $e^{H(X)}$ represents, in the sense of the asymptotic equipartition property, the average number of codewords needed to encode each variable X_i.

Remark 32 *We have considered the case of a source made of independent and identically distributed random variables. Actually, analogous results can be proved for more general processes as, for instance, Markovian sources (see [52]). In particular, the asymptotic equipartition property is based on the weak law of the large numbers, therefore it holds for a wide variety of stochastic processes to which that law can be applied, such as ergodic processes.*

7.2 The measure of performance and complexity

Let us introduce the class of quantized systems considered in this chapter, then, for these systems, we define the measures of performance and the notion of complexity of a quantizer. With reference to the framework introduced in Section 2.4, when analyzing the complexity of a quantizer, what matters is not the knowledge of the function, but only to know the induced partition. Therefore, in this chapter, quantizations keep up with locally finite partitions. We limit ourselves to consider partitions of $\mathbb{R}$ made of intervals:

Definition 35 *A partition of* $\mathbb{R}$ *of the type* $\mathbb{R} = \bigcup_{k\in\mathbb{Z}} I_k$ *which is locally finite and such that* $\forall\, k \in \mathbb{Z}$, I_k *is an interval of nonzero length, is called a* quantization of $\mathbb{R}$.

A quantization of $\mathbb{R}$ is denoted by $\mathcal{I} = \{I_k\}_{k\in\mathbb{Z}}$. In compliance with Definition 35, all along this chapter, by a *quantizer* we mean any function $q : \mathbb{R} \to \mathcal{U}$ such that the induced partition $\mathbb{R} = \bigcup_{u\in\mathcal{U}}\{q^{-1}(u)\}$ is a quantization of[1] $\mathbb{R}$.
We consider a discrete time scalar linear system interconnected with a static feedback quantizer $u : \mathbb{R} \to \mathcal{U} \subset \mathbb{R}$, namely:

$$x^+ = \varphi(x) := ax + u(x)\,, \tag{7.3}$$

where $|a| > 1$. Let $X_0 : \Omega \to \mathbb{R}$ be a random variable (defined on some probability space Ω) representing the initial condition: $X_t = \varphi^t(X_0)$ represents the state of the process at time t. The distribution of X_t is denoted by μ_t while its mean–square value (or energy) is denoted by $\mathcal{E}(\mu_t) := \int_{\mathbb{R}} x^2 d\mu_t$ (when μ_t is clear from the context, we also denote it by $\mathcal{E}_t$). The initial distribution μ_0 is supposed to belong to some class of probability distributions $\mathcal{P}$. We assume that $\mathcal{P}$ is closed under the dynamics φ (namely, $\mu_0 \in \mathcal{P} \Rightarrow \mu_1 \in \mathcal{P}$) and $\forall\, \mu \in \mathcal{P}$, $\mathcal{E}(\mu) < +\infty$. A class of distributions $\mathcal{P}$ with these properties is referred to as *admissible*. Examples of admissible classes $\mathcal{P}$ are

$$\mathcal{P}_{\rm all} := \{\mu \in \mathcal{P}r(\mathbb{R}) \,|\, \mathcal{E}(\mu) < +\infty\}\,, \tag{7.4}$$

or the space of probabilities that are absolutely continuous with respect to the Lebesgue measure and having finite energy.

Definition 36 *A quantized control law* $u : \mathbb{R} \to \mathcal{U}$ *is said to be* $\mathcal{E}_\infty$–converging *iff the closed loop dynamics (7.3) is such that* $\forall\, \mu_0 \in \mathcal{P}$, $\limsup_{t\to+\infty} \mathcal{E}_t \le \mathcal{E}_\infty$. *We say that* u *is* mean–square practically converging *if it is* $\mathcal{E}_\infty$*–converging for some* $\mathcal{E}_\infty$.

The quantity $\mathcal{E}_\infty$ represents a *steady–state* performance measure of the closed loop system. We are interested in analyzing also the transient behavior of this kind of dynamics. To this

[1] The terminology and the notation used in this chapter is slightly different from that introduced in Chapter 2. In particular: since the elements of the considered partitions are intervals, they are more conveniently denoted by I_k, instead of using the symbol $\mathcal{C}$. Also the term quantizer is referred to a different notion from that introduced in Definition 4: in fact, $\mathcal{U}$ is not supposed to be a quantized set but the induced partition is required to be locally finite (and made of intervals).

end we consider a parameter related with the decaying rate of the energy. More precisely, let $J_0 := [-r_0\,,\, r_0] \subset \mathbb{R}$, for some $r_0 > 0$, and $J_e := \mathbb{R} \setminus J_0$. We assume that φ is such that $\varphi(J_0) \subseteq J_0$. In this framework we introduce the following

Definition 37 *Let* $\mu \in \mathcal{P}r(\mathbb{R})$, *the* external energy *of* μ *is defined by* $\mathcal{E}_e(\mu) := \int_{J_e} x^2 d\mu$.

Clearly, if it happens that

$$\forall\, \mu_0 \in \mathcal{P}\,,\ \lim_{t\to+\infty} \mathcal{E}_e(\mu_t) = 0\,,$$

then the control law is $\mathcal{E}_\infty$–converging for every $\mathcal{E}_\infty \geq r_0^2$. In this context a way to measure the *transient* behavior is through

$$\mathcal{T}_e := -\sup_{\mu_0\in\mathcal{P}} \limsup_{t\to+\infty} \frac{\log \mathcal{E}_e(\mu_t)}{t} \tag{7.5}$$

which represents the worst case decaying rate of the external energy. Increasing positive values of $\mathcal{T}_e$ correspond to faster convergence towards zero of the external energy.

Before introducing a measure of *complexity* for a quantized controller, the following preliminary remark is needed. Since the system was supposed to be open loop unstable, if the class $\mathcal{P}$ contains distributions having unbounded support, a necessary condition for the mean–square practical convergence is that the quantizer u takes infinite values. In fact, if $\mathcal{U}$ is a finite set, then the induced quantization contains unbounded elements $\big($namely, the controller $u(x)$ saturates$\big)$ and the energy diverges [89]. Therefore, we tacitly assume that any quantization $\mathcal{I} = \{I_k\}_{k\in\mathbb{Z}}$ considered in this chapter is such that $\forall\, k \in \mathbb{Z}$, I_k is bounded.
The definition of a complexity function for a controller has been motivated in Section 2.4 with reference to the coding problem. In Section 7.1, the coding problem has been also related to the minimization of transmission delays which is a crucial issue in control. Let us follow the approach taken in Section 2.4 and let us show how the theory presented in Section 7.1 allows us to fit it to the probabilistic setting considered in this chapter.
In Section 2.4, the complexity function was defined by counting the number $N(r)$ of control values needed to deal with a system whose state belongs to a bounded set $\mathcal{B}_r$. Here, r may be replaced by an energy value $\mathcal{E}$ and $\mathcal{B}_r$ by the family of probability distributions whose energy is not larger than $\mathcal{E}$. In this case, however, the natural choice of replacing N with the worst case number of control values needed to deal with these distributions is meaningless because the considered family contains distributions having unbounded support, thus requiring infinite control values. As we have illustrated in Section 7.1, the count of the needed control values may be more suitably done by taking the statistics into account as suggested by the "source coding theorem" and by the "asymptotic equipartition property". Therefore, we define a coding complexity function $\mathcal{H}$ for system (7.3), as well as the corresponding coding cardinality function $\mathcal{N}$, as entropy–like functions associated to the quantization induced by u.

Definition 38 (Coding complexity function) *Let $\mathcal{I} = \{I_k\}_{k\in\mathbb{Z}}$ be a quantization of $\mathbb{R}$ such that $\forall\, k \in \mathbb{Z}$, I_k is bounded. Given a probability measure $\mu \in \mathcal{P}r(\mathbb{R})$, let $\mathbf{p}_{\mu,\mathcal{I}} \in \mathcal{P}_r(\mathbb{Z})$ be defined by $\mathbf{p}_{\mu,\mathcal{I}} := \{\mu(I_k)\}_{k\in\mathbb{Z}}$. The* coding complexity function *$\mathcal{H}$ associated to $\mathcal{I}$ is defined by*

$$\begin{array}{rccl} \mathcal{H}: & \mathbb{R}^+ & \to & \mathbb{R}^+ \\ & \mathcal{E} & \mapsto & \sup\limits_{\substack{\mu\in\mathcal{P}r(\mathbb{R}):\\ \int_{\mathbb{R}} x^2 d\mu = \mathcal{E}}} H(\mathbf{p}_{\mu,\mathcal{I}})\,. \end{array} \tag{7.6}$$

Accordingly, the function $\mathcal{N}(\mathcal{E}) := e^{\mathcal{H}(\mathcal{E})}$ represents the worst case average number of control symbols necessary to encode a distribution having energy $\mathcal{E}$ and will be referred to as the *coding cardinality function*.
The coding complexity function is well–defined under suitable assumptions on the quantization $\mathcal{I}$ which are quite general and will be specified in next Section 7.3.1.

Remark 33 *The definition of a coding complexity function has been related to the effect of transmission delays on the dynamics of the system caused by the presence of arbitrarily long coding sequences. In this chapter, however, we do not consider delays and the presented analysis is quite theoretical. Namely, the coding and delay issue are the motivation to introduce the considered measures of complexity but, at this stage, there is no code implementation. Anyhow, the study of the coding complexity function has a central role in our analysis. Indeed, not only this study is the first step towards the analysis in a more general case including the effects of delays, but also it allows us to point out fundamental relations between the quantization structure and the achievable performance in terms of the convergence rate $\mathcal{T}_e$ and of the minimal asymptotic value of the energy $\mathcal{E}_\infty$. These issues are illustrated in Section 7.4.*

7.3 The coding complexity function

7.3.1 Analysis

It is not apparent from Definition 38 how to analyze the properties of the coding complexity function. In particular, the value of $\mathcal{H}(\mathcal{E})$ results from a maximization problem over a space of probability distributions on $\mathbb{R}$. We first show that the function $\mathcal{H}$ has an equivalent definition involving a maximization problem over discrete probability measures only. For such a problem an implicit solution can be found. Moreover, we introduce an easily computable way to characterize the geometric structure of the quantization and we show that the relation between the geometric structure of the quantization and the behavior of the coding complexity function can be expressed in terms of a Laplace transformation. This is the fundamental step which allows us to analyze the behavior of the coding complexity function associated to a wide range of quantizers. A synthesis of the main results on the Laplace transformation is given in Appendix A.6.1.

For the sake of simplicity, we consider *symmetric quantizations*[2], that is: let $\mathcal{D}_{\mathcal{I}} = \{d_k\}_{k\in\mathbb{Z}}$,

[2]This assumption can be removed, see next Remark 35.

where $d_k := \operatorname*{argmin}_{x \in \bar{I}_k} |x|$ (namely, $\mathcal{D}_{\mathcal{I}}$ is the set containing 0 and the extremes of the intervals I_k's), we assume that $\mathcal{D}_{\mathcal{I}} = -\mathcal{D}_{\mathcal{I}}$.
Since a quantization $\mathcal{I}$ is a locally finite partition, we can suppose that the elements of $\mathcal{D}_{\mathcal{I}}$ are indexed so that $\forall\, k \in \mathbb{Z}$, $d_k < d_{k+1}$ and $d_0 = 0$. In particular, it holds that $d_{-k} = -d_k$.
The geometric structure of the quantization $\mathcal{I}$ is described by the following function:

Definition 39 *Given a symmetric quantization $\mathcal{I}$, the function*

$$g(x) := \#\left(\mathcal{D}_{\mathcal{I}} \cap \,]\,0\,,\,\sqrt{x}\,]\right)$$

is referred to as the g–function *associated to $\mathcal{I}$.*

It holds that $\forall\, k \in \mathbb{N}$ and $\forall\, x \in [d_k^2\,,\, d_{k+1}^2[$, $g(x) = k$. Also, as the I_k's are bounded,

$$\lim_{x \to +\infty} g(x) = +\infty\,.$$

In order to ensure the existence of the coding complexity function, we restrict ourselves to consider quantizers u whose induced quantization $\mathcal{I}$ is such that the growth of the d_k's is at least monomial, namely

A1) $\exists\, \lambda > 0$ such that

$$\lim_{k \to +\infty} \frac{d_k}{k^{\lambda}} = +\infty\,. \tag{7.7}$$

Assumption **A1** can be equivalently formulated in terms of the g–function whose growth must be at most monomial (see Lemma 32 in Appendix A.6.2):

A1) $\exists\, \gamma > 0$ such that

$$\lim_{x \to +\infty} \frac{g(x)}{x^{\gamma}} = 0\,. \tag{7.8}$$

Such an assumption is a mild restriction, in fact all the typically encountered quantizers (such as uniform and logarithmic ones) are included in our analysis.
Consider the following discrete version of the coding complexity function:

$$\begin{array}{rccc} \mathbb{H}: & \mathbb{R}^+ & \to & \mathbb{R}^+ \\ & \mathcal{E} & \mapsto & \displaystyle\sup_{\substack{\mathbf{p} \in \mathcal{P}r(\mathbb{Z}) \\ \sum_k d_k^2 p_k = \mathcal{E}}} H(\mathbf{p})\,. \end{array} \tag{7.9}$$

The main result on the analysis of the coding complexity function is provided by the following

Theorem 14 (Properties of $\mathcal{H}$) *For a symmetric quantization $\mathcal{I}$ satisfying* **A1** *the following facts hold:*

i) $\mathcal{H}(\mathcal{E}) = \mathbb{H}(\mathcal{E})$ *;*

ii) The function $\mathbb{H}(\mathcal{E})$ is implicitly defined by the following system

$$\begin{cases} \mathbb{H}\big(\mathcal{E}(\beta)\big) = \log\big(1 + 2\beta G(\beta)\big) + \beta\,\mathcal{E}(\beta) & \text{(7.10a)} \\ \mathcal{E}(\beta) = -\dfrac{d}{d\beta}\log\big(1 + 2\beta G(\beta)\big) \\ \qquad\quad = -\dfrac{2\big(G(\beta) + \beta G'(\beta)\big)}{1 + 2\beta G(\beta)} & \text{(7.10b)} \\ \beta > 0\,. \end{cases}$$

where $G(s)$ is the Laplace transform of the g–function associated to $\mathcal{I}$.

iii) $\mathcal{E}(\beta)$ is a decreasing analytic function with

$$\begin{cases} \lim_{\beta\to 0^+}\mathcal{E}(\beta) = +\infty \\ \lim_{\beta\to+\infty}\mathcal{E}(\beta) = 0\,. \end{cases}$$

iv) $\mathbb{H}$ is analytic and $\frac{d\mathbb{H}}{d\mathcal{E}} = \beta$. In particular, $\mathbb{H}(\mathcal{E})$ is an increasing function. Moreover,

$$\begin{cases} \lim_{\mathcal{E}\to 0^+}\mathbb{H}(\mathcal{E}) = 0 \\ \lim_{\mathcal{E}\to+\infty}\mathbb{H}(\mathcal{E}) = +\infty\,. \end{cases}$$

Proof. The proof is given below, after one preliminary result. ■

Remark 34 *By Theorem 14.iii, $\mathcal{E}(\beta)$ is an analytic and invertible function having non zero derivative, therefore its inverse $\beta(\mathcal{E})$ is analytic. Parts ii and iii of the theorem can be hence rephrased by saying that $\mathcal{E}(\beta)$ is an analytic diffeomorphism from $\mathbb{R}^+$ to itself not preserving the orientation and that, in the new coordinates, $\mathbb{H}$ is represented by the function*

$$\widetilde{\mathbb{H}}(\beta) := \mathbb{H}\big(\mathcal{E}(\beta)\big) = \log\big(1 + 2\beta G(\beta)\big) - \frac{2\big(\beta G(\beta) + \beta^2 G'(\beta)\big)}{1 + 2\beta G(\beta)}\,. \tag{7.11}$$

In order to prove the theorem we shall make use of some concepts borrowed from the statistical mechanics:

Definition 40 *Let $\mathcal{I} = \{I_k\}_{k\in\mathbb{Z}}$ be a quantization of $\mathbb{R}$. Should the series $\sum_{k\in\mathbb{Z}} e^{-\beta d_k^2}$ converge for some $\beta \in \mathbb{R}^+$, then it defines a function $\mathcal{Z}(\beta)$ which is called the* partition function *associated to $\mathcal{I}$.*

Proposition 21 (Representation of $\mathcal{Z}$ as a Laplace integral) *Consider a symmetric quantization $\mathcal{I}$ satisfying assumption* **A1**. *Then the partition function associated to $\mathcal{I}$ is defined $\forall\,\beta > 0$. Moreover, the right half–plane of convergence of the Laplace transform $G(s)$ of the g–function associated to $\mathcal{I}$ contains $\{s \in \mathbb{C}\,|\,\mathcal{R}e(s) > 0\}$ and*

$$\mathcal{Z}(\beta) = 1 + 2\beta G(\beta)\,.$$

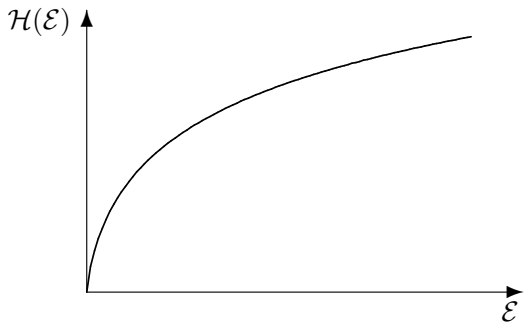

Figure 7.2: Typical behavior of the function $\mathcal{H}(\mathcal{E})$.

Proof. Let $\Phi(\beta) := \sum_{k=1}^{+\infty} e^{-\beta d_k^2}$, by the symmetry assumption

$$\mathcal{Z}(\beta) = 1 + 2\Phi(\beta)\,.$$

Assumption **A1** in the form of equation (7.7) implies that $\exists\,\lambda > 0$, $M > 0$ and $k_M > 0$ such that $\forall\, k \geq k_M$, $d_k \geq M k^{\lambda}$. In particular, $e^{-\beta d_k^2} \leq e^{-\beta M^2 k^{2\lambda}}$. Since $\forall\,\beta > 0$, $\exists\,\hat{k} > 0$ such that $\forall\, k \geq \hat{k}$, $e^{-\beta M^2 k^{2\lambda}} \leq \frac{1}{k^2}$ and $\sum_{k=1}^{+\infty} \frac{1}{k^2}$ is convergent, then $\Phi(\beta)$, and hence the partition function, is defined on $\mathbb{R}^+$.
By assumption **A1** in the form of equation (7.8), $\exists\,\gamma > 0$, $\epsilon > 0$ and $x_\epsilon > 0$ such that $\forall\, x \geq x_\epsilon$, $g(x) \leq \epsilon x^{\gamma}$. Hence, $\forall\,\beta > 0$, $G(\beta) = \int_0^{+\infty} g(x)e^{-\beta x}dx \leq \int_0^{x_\epsilon} g(x)e^{-\beta x}dx + \epsilon \int_{x_\epsilon}^{+\infty} x^{\gamma} e^{-\beta x}dx < +\infty$.
Finally, let us show that $\Phi(\beta) = \beta G(\beta)$. Indeed,

$$\begin{aligned}\beta G(\beta) = &\ \beta \int_0^{+\infty} g(x)e^{-\beta x}dx = \beta \sum_{k=0}^{+\infty} \int_{d_k^2}^{d_{k+1}^2} g(x)e^{-\beta x}dx = \\ = &\ \beta \sum_{k=0}^{+\infty} k \int_{d_k^2}^{d_{k+1}^2} e^{-\beta x}dx = \sum_{k=0}^{+\infty} k(e^{-\beta d_k^2} - e^{-\beta d_{k+1}^2})\,.\end{aligned}$$

Let us compute the partial sum of the latter series: $S_N := \sum_{k=0}^{N-1} k(e^{-\beta d_k^2} - e^{-\beta d_{k+1}^2}) = \sum_{k=1}^{N-1} e^{-\beta d_k^2} - (N-1)e^{-\beta d_N^2}$. Hence, $\Phi(\beta) - S_N = \sum_{k=N}^{+\infty} e^{-\beta d_k^2} + (N-1)e^{-\beta d_N^2}$ which converges to 0 as $N \to +\infty$ thanks to the assumption on the growth of the d_k's. ■

Remark 35 *A representation of $\mathcal{Z}(\beta)$ as a Laplace integral can be obtained also when the set $\mathcal{D}_{\mathcal{I}}$ is not symmetric. To this aim call $g^+(x)$ the function introduced in Definition 39 and let $g^-(x) := \#\left(-\mathcal{D}_{\mathcal{I}} \cap \left]\,0\,,\,\sqrt{x}\,\right]\right)$: it is easy to see that $\mathcal{Z}(\beta) = 1 + \beta\big(G^-(\beta) + G^+(\beta)\big)$.*

We are ready for the
Proof of Theorem 14. We first prove the properties of the function $\mathbb{H}$ (i.e., parts *ii*–*iv*) as some of them are instrumental for the proof of part *i*.

Proof of part ii: the optimization problem in (7.9) can be solved via the use of Lagrange multipliers. This yields that, for $\mathcal{E} > 0$, the maximizing probability measure results from

the solution of the following system:

$$\begin{cases} \log p_k + 1 + \alpha + \beta d_k^2 = 0\,, \quad k \in \mathbb{Z} & \text{(7.12a)} \\ \sum_k p_k = 1 & \text{(7.12b)} \\ \sum_k d_k^2 p_k = \mathcal{E}\,. & \text{(7.12c)} \end{cases}$$

Solving equation (7.12a) for p_k, one gets $p_k = ce^{-\beta d_k^2}$, with c not depending on k. Such a constant is determined by imposing the normalization condition in equation (7.12b). In this way, an implicit solution for system (7.12) is

$$p_k = \frac{1}{\mathcal{Z}(\beta)}\, e^{-\beta d_k^2} \tag{7.13}$$

(where the partition function is convergent thanks to Proposition 21). The multiplier β, which is necessarily positive, is determined by equation (7.12c), namely

$$\frac{\sum_{k\in\mathbb{Z}} d_k^2 e^{-\beta d_k^2}}{\sum_{k\in\mathbb{Z}} e^{-\beta d_k^2}} = \mathcal{E}\,. \tag{7.14}$$

This equation can be written in the form

$$\mathcal{E}(\beta) = -\frac{d}{d\beta}\log \mathcal{Z}(\beta) = -\frac{\mathcal{Z}'(\beta)}{\mathcal{Z}(\beta)}\,. \tag{7.15}$$

Let $\mathbf{p}_{\max}(\mathcal{E}) = \{p_k\}_{k\in\mathbb{Z}}$ be the maximizing probability measure. By direct computation $\big($i.e., plug the expression of the p_k's given in equation (7.13) in the definition of the entropy and take advantage of equation (7.14)$\big)$, the corresponding value of the entropy is given by the implicit expression

$$H\big(\mathbf{p}_{\max}(\mathcal{E})\big) = \mathbb{H}\big(\mathcal{E}(\beta)\big) = \log \mathcal{Z}(\beta) + \beta\, \mathcal{E}(\beta)\,. \tag{7.16}$$

Using the representation of $\mathcal{Z}$ as a Laplace integral, equation (7.16) can be written as

$$\mathbb{H}\big(\mathcal{E}(\beta)\big) = \log\big(1 + 2\beta G(\beta)\big) + \beta\, \mathcal{E}(\beta)$$

and equation (7.15) as

$$\mathcal{E}(\beta) = -\frac{d}{d\beta}\log\big(1 + 2\beta G(\beta)\big)\ :$$

this concludes the proof of part *ii*.

Proof of part iii: by equation (7.10b), the analyticity of $\mathcal{E}(\beta)$ is a consequence of the analyticity of $G(\beta)$ which follows by Theorem 18 in Appendix A.6.1. Let us show that $\mathcal{E}(\beta)$ is decreasing: by equation (7.15) we have $\frac{d}{d\beta}\mathcal{E}(\beta) = -\frac{d}{d\beta}\frac{\mathcal{Z}'}{\mathcal{Z}} = -\big(\frac{\mathcal{Z}''}{\mathcal{Z}} - (\frac{\mathcal{Z}'}{\mathcal{Z}})^2\big) = -\big(\frac{\mathcal{Z}''}{\mathcal{Z}} - \mathcal{E}^2\big)$. Consider $\mathbf{p}_{\max}(\mathcal{E}) = \{p_k\}_{k\in\mathbb{Z}}$ and let $\mathbf{p}_{\mathcal{D}_\mathcal{I}}$ be the probability measure on $\mathcal{D}_\mathcal{I}$ defined by $\mathbf{p}_{\mathcal{D}_\mathcal{I}}(d_k) := p_k$. Let D be a random variable taking values in $\mathcal{D}_\mathcal{I}$ and distributed according to $\mathbf{p}_{\mathcal{D}_\mathcal{I}}$: by equation (7.14), $\mathcal{E} = \mathbb{E}[D^2]$ and $\frac{\mathcal{Z}''}{\mathcal{Z}} = \frac{\sum_{i\in\mathbb{Z}} d_i^4 e^{-\beta d_i^2}}{\mathcal{Z}} = \mathbb{E}[D^4]$. Hence, $\frac{d}{d\beta}\mathcal{E}(\beta) =$

$-\big(\mathbb{E}[D^4] - (\mathbb{E}[D^2])^2\big) = -\mathbb{V}\mathrm{ar}[D^2] < 0$.
Let us prove that $\lim_{\beta\to 0^+} \mathcal{E}(\beta) = +\infty$. According to Proposition 21, equation (7.15) can be written as

$$\mathcal{E}(\beta) = -\frac{2\Phi'(\beta)}{1+2\Phi(\beta)}\,,$$

where $\Phi(\beta) = \sum_{k=1}^{+\infty} e^{-\beta d_k^2}$. Also, $\Phi(\beta) = \beta G(\beta)$ and, by the "Final value theorem" (see Theorem 20 in Appendix A.6.1), $\lim_{\beta\to 0^+} \Phi(\beta) = +\infty$. We have hence to prove that $\lim_{\beta\to 0^+} -\frac{\Phi'(\beta)}{\Phi(\beta)} = +\infty$. Indeed, by assumption **A1** in the form of equation (7.7), $\exists\,\lambda > 0$ such that $\lim_{k\to+\infty} \frac{d_k}{k^\lambda} = +\infty$. Hence, $\forall\, M > 0$, $\exists\, k_M$ such that $\forall\, k \geq k_M$, $d_k \geq Mk^\lambda$. Let $S(\beta) := \sum_{k=1}^{k_M-1} d_k^2 e^{-\beta d_k^2}$, then

$$\begin{aligned}
-\tfrac{\Phi'(\beta)}{\Phi(\beta)} &= \tfrac{S(\beta)+\sum_{k=k_M}^{+\infty} d_k^2 e^{-\beta d_k^2}}{\Phi(\beta)} \geq \\
&\geq \tfrac{S(\beta)+M^2\sum_{k=k_M}^{+\infty} k^{2\lambda} e^{-\beta d_k^2}}{\Phi(\beta)} \geq \\
&\geq \tfrac{S(\beta)+M^2k_M^{2\lambda}\sum_{k=k_M}^{+\infty} e^{-\beta d_k^2}}{\Phi(\beta)} = \\
&= \tfrac{\left(S(\beta)-M^2k_M^{2\lambda}\sum_{k=1}^{k_M-1} e^{-\beta d_k^2}\right)+M^2k_M^{2\lambda}\Phi(\beta)}{\Phi(\beta)} = \\
&= \tfrac{\mathscr{S}(\beta)+M^2k_M^{2\lambda}\Phi(\beta)}{\Phi(\beta)}\,,
\end{aligned}$$

where $\mathscr{S}(\beta) := S(\beta) - M^2k_M^{2\lambda}\sum_{k=1}^{k_M-1} e^{-\beta d_k^2}$. $\mathscr{S}(\beta)$ is defined by a finite sum and has a finite limit as $\beta \to 0^+$, therefore

$$\liminf_{\beta\to 0^+} -\frac{\Phi'(\beta)}{\Phi(\beta)} \geq \lim_{\beta\to 0^+} \frac{\mathscr{S}(\beta)+M^2k_M^{2\lambda}\Phi(\beta)}{\Phi(\beta)} = M^2k_M^{2\lambda}\,.$$

Since M can be chosen so as to make $M^2k_M^{2\lambda}$ arbitrarily large, the thesis follows.
Finally, let us show that $\lim_{\beta\to+\infty} \mathcal{E}(\beta) = 0$. Consider the expression for $\mathcal{E}(\beta)$ given in equation (7.10b). By the "Initial value theorem" (see Theorem 19 in Appendix A.6.1),

$$\lim_{\beta\to+\infty} \beta G(\beta) = \lim_{x\to 0^+} g(x) = 0\,, \tag{7.17}$$

therefore $\lim_{\beta\to+\infty} G(\beta) = 0$. Also,

$$\lim_{\beta\to+\infty} \beta G'(\beta) = \lim_{\beta\to+\infty} \beta\mathcal{L}[-xg(x)](\beta) = -\lim_{x\to 0^+} xg(x) = 0\,,$$

where the first equality holds by Theorem 18 and the second one by Theorem 19. These limits, applied to equation (7.10b), yield the result.

Proof of part iv: we have already noticed in Remark 34 that part *iii* implies the analyticity of $\beta(\mathcal{E})$. It then follows that $\mathbb{H}(\mathcal{E})$ is analytic because $\mathbb{H}(\mathcal{E}) = \widetilde{\mathbb{H}}\big(\beta(\mathcal{E})\big)$ and $\widetilde{\mathbb{H}}(\beta)$ (see

equation (7.11)) is analytic by Theorem 18 .
Using the expression for $\widetilde{\mathbb{H}}(\beta)$ given in equation (7.16) ,

$$\frac{d\,\mathbb{H}(\mathcal{E})}{d\mathcal{E}} = \frac{d\,\widetilde{\mathbb{H}}\big(\beta(\mathcal{E})\big)}{d\mathcal{E}} = \frac{d\,\widetilde{\mathbb{H}}(\beta)}{d\beta}\cdot\frac{d\beta}{d\mathcal{E}} = \frac{\mathcal{Z}'}{\mathcal{Z}}\cdot\frac{d\beta}{d\mathcal{E}} + \mathcal{E}\frac{d\beta}{d\mathcal{E}} + \beta\frac{d\mathcal{E}}{d\beta}\cdot\frac{d\beta}{d\mathcal{E}} = \beta$$

because $\mathcal{E} = -\frac{\mathcal{Z}'}{\mathcal{Z}}$ (see equation (7.15)) .
By part iii , $\lim_{\mathcal{E}\to 0^+}\mathbb{H}(\mathcal{E}) = \lim_{\beta\to+\infty}\mathbb{H}\big(\mathcal{E}(\beta)\big)$ and $\lim_{\mathcal{E}\to+\infty}\mathbb{H}(\mathcal{E}) = \lim_{\beta\to 0^+}\mathbb{H}\big(\mathcal{E}(\beta)\big)$.
Let us compute these two limits.
By equations (7.11) and (7.17) ,

$$\lim_{\beta\to+\infty}\mathbb{H}\big(\mathcal{E}(\beta)\big) = -2\lim_{\beta\to+\infty}\beta^2 G'(\beta)\,.$$

Since $G(\beta) = \frac{\Phi(\beta)}{\beta}$, then

$$\beta^2 G'(\beta) = \beta\Phi'(\beta) - \Phi(\beta)$$

and, again by equation (7.17) , $\lim_{\beta\to+\infty}\Phi(\beta) = 0$. It is hence sufficient to show that $\lim_{\beta\to+\infty}\beta\Phi'(\beta) = 0$. Indeed, $\beta\Phi'(\beta) = -\sum_{k=1}^{+\infty}\beta d_k^2 e^{-\beta d_k^2}$: this series is uniformly convergent on $[1\,,\,+\infty[$, thus[3]

$$\lim_{\beta\to+\infty}\sum_{k=1}^{+\infty}\beta d_k^2 e^{-\beta d_k^2} = \sum_{k=1}^{+\infty}\lim_{\beta\to+\infty}\beta d_k^2 e^{-\beta d_k^2} = 0\,.$$

To compute the last limit, let us use the expression for $\mathbb{H}\big(\mathcal{E}(\beta)\big)$ given in equation (7.10a) .
Since $\forall\,\beta > 0$, $\beta\mathcal{E}(\beta) > 0$, it is sufficient to show that $\lim_{\beta\to 0^+}\log\big(1+2\beta G(\beta)\big) = +\infty$. This holds true because we have already shown in the proof of part iii that $\lim_{\beta\to 0^+}\beta G(\beta) = +\infty$.

Proof of part i : we first prove that $\mathbb{H}(\mathcal{E}) \geq \mathcal{H}(\mathcal{E})$ by showing that $\forall\,\mu \in \mathcal{P}r(\mathbb{R})$ such that $\int_{\mathbb{R}} x^2 d\,\mu = \mathcal{E}$, it holds that $H(\mathbf{p}_{\mu,\mathcal{I}}) \leq \mathbb{H}(\mathcal{E})$. Indeed, given such a μ , consider $\{p_k\}_{k\in\mathbb{Z}} = \mathbf{p}_{\mu,\mathcal{I}}$. Since $\mathcal{E} = \int_{\mathbb{R}} x^2 d\,\mu = \sum_{k\in\mathbb{Z}}\int_{I_k} x^2 d\,\mu \geq \sum_{k\in\mathbb{Z}} d_k^2 p_k := \mathcal{E}'$ and $\mathbb{H}(\mathcal{E})$ is an increasing function (thanks to part iv) , then $H(\mathbf{p}_{\mu,\mathcal{I}}) \leq \mathbb{H}(\mathcal{E}') \leq \mathbb{H}(\mathcal{E})$.
On the other hand, let $\mathbf{p} = \{p_k\}_{k\in\mathbb{Z}}$ be such that $\sum_k d_k^2 p_k = \mathcal{E}$: it is easy to see that $\forall\,\epsilon > 0$, $\exists\,\mu_\epsilon \in \mathcal{P}r(\mathbb{R})$ such that $\mathbf{p}_{\mu_\epsilon,\mathcal{I}} = \mathbf{p}$ and $\int_{\mathbb{R}} x^2 d\,\mu_\epsilon = \mathcal{E} + \epsilon$. This implies that $\forall\,\epsilon > 0$, $\mathcal{H}(\mathcal{E}+\epsilon) \geq \mathbb{H}(\mathcal{E})$. Now, suppose by contradiction that $\mathcal{H}(\mathcal{E}) = \mathbb{H}(\mathcal{E}) - \delta$ with $\delta > 0$: by the continuity of $\mathbb{H}(\mathcal{E})$, $\exists\,\epsilon > 0$ such that $\mathbb{H}(\mathcal{E}-\epsilon) > \mathbb{H}(\mathcal{E}) - \delta$. Hence, $\mathcal{H}(\mathcal{E}) \geq \mathbb{H}(\mathcal{E}-\epsilon) > \mathbb{H}(\mathcal{E}) - \delta$ which is a contradiction. ■

Remark 36 *In the proof of part iv of Theorem 14 , we have shown that* $\lim_{\beta\to+\infty}\beta^2 G'(\beta) = 0$ *using tools from elementary analysis. An alternative proof of this equality can be obtained by resorting to the theory of the Laplace transformation as follows: let $g(x)$ be a function defined for $x > 0$, $Dg(x)$ is said to be a* generalized derivative *of $g(x)$ iff*

$$\textit{for}\quad x > 0\,,\quad g(x) = \lim_{x\to 0^+} g(x) + \int_0^x Dg(y)dy\,.$$

[3] An alternative way to show that $\lim_{\beta\to+\infty}\beta^2 G'(\beta) = 0$, based on the theory of the Laplace transformation, is provided in Remark 36 at the end of the proof of the theorem.

It holds that (see [35], Theorem 9.2, page 41) if $g(x)$ *has the generalized derivative* $Dg(x)$ *and* $\mathcal{L}[Dg](\beta)$ *converges for some* $\beta > 0$, *then* $G(\beta)$ *converges too for such a* β *and*

$$\mathcal{L}[Dg](\beta) = \beta G(\beta) - \lim_{x\to 0^+} g(x)\,. \tag{7.18}$$

Hence,

$$\begin{aligned}
-\lim_{\beta\to+\infty} \beta^2 G'(\beta) = \; & -\lim_{\beta\to+\infty} \beta\big(\beta G'(\beta)\big) = \\
\overset{\text{(a)}}{=} \; & \lim_{\beta\to+\infty} \beta\big(\beta\mathcal{L}[xg(x)](\beta)\big) = \\
\overset{\text{(b)}}{=} \; & \lim_{\beta\to+\infty} \beta\big(\mathcal{L}\big[D\big(xg(x)\big)\big](\beta) + \lim_{x\to 0^+} xg(x)\big) = \\
= \; & \lim_{\beta\to+\infty} \beta\big(\mathcal{L}\big[D\big(xg(x)\big)\big](\beta)\big) = \\
\overset{\text{(c)}}{=} \; & \lim_{x\to 0^+} D\big(xg(x)\big) = 0\,,
\end{aligned}$$

where equality (a) *holds by Theorem 18, equality* (b) *by equation* (7.18) *and equality* (c) *by the "Initial value theorem".*

The study of the function $\mathcal{H}(\mathcal{E})$ is aided by the tools provided by the Laplace transformation theory. As it is shown in next Section 7.3.2, this theory is particularly helpful to go into a thorough analysis of the asymptotic behavior of $\mathcal{H}(\mathcal{E})$ as $\mathcal{E} \to +\infty$.

Proposition 22 (Monotony with respect to the *g*–function) *Let* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *be two quantizations of* $\mathbb{R}$. *Denote by* g_1 *and* g_2 *the corresponding g–functions and by* $\mathbb{H}_{g_1}$ *and* $\mathbb{H}_{g_2}$ *the respective complexity measure functions. If* $g_1(x) \geq g_2(x)\ \forall x \geq 0$, *then* $\mathbb{H}_{g_1}(\mathcal{E}) \geq \mathbb{H}_{g_2}(\mathcal{E})\ \forall\, \mathcal{E} > 0$.

Proof. Denote by $d_{k,i}$ the elements of $\mathcal{D}_{\mathcal{I}_i}$. The assumption $g_1(x) \geq g_2(x)\ \forall x \geq 0$ is equivalent to assume that $d_{k,1} \leq d_{k,2}\ \forall\, k \geq 0$. Therefore, for $\mathbf{p} \in \mathcal{P}r(\mathbb{Z})$, $\sum_k d_{k,1}^2 p_k \leq \sum_k d_{k,2}^2 p_k$ and

$$\begin{aligned}
\mathbb{H}_{g_2}(\mathcal{E}) = \; & \sup_{\substack{\mathbf{p}\in\mathcal{P}r(\mathbb{Z}):\\ \sum_k d_{k,2}^2 p_k = \mathcal{E}}} H(\mathbf{p}) \\
\leq \; & \sup_{\substack{\mathbf{p}\in\mathcal{P}r(\mathbb{Z}):\\ \sum_k d_{k,1}^2 p_k \leq \mathcal{E}}} H(\mathbf{p}) \\
= \; & \mathbb{H}_{g_1}(\mathcal{E})\,,
\end{aligned}$$

where the last equality holds because, by Theorem 14.*iv*, $\mathbb{H}_{g_1}(\mathcal{E})$ is an increasing function. ∎

7.3.2 Asymptotic behavior of $\mathcal{H}(\mathcal{E})$: monomial and floating–point quantizations

In this section we introduce two main classes of quantizers that extend the uniform and the logarithmic quantizers defined in Section 2.1. By taking advantage of Theorem 14, we analyze the asymptotic behavior of the corresponding coding complexity function.

If $\mathcal{D}_{\mathcal{I}} = \{\pm\, d_1 k^\lambda \,:\, k \in \mathbb{N}\}$, $d_1 > 0$ and $\lambda > 0$, then

$$g(x) = \Big\lfloor \Big(\frac{x}{d_1^2} \Big)^{1/2\lambda} \Big\rfloor . \tag{7.19}$$

When $\lambda = 1$, $\mathcal{I}$ is a uniform partition of $\mathbb{R}$ with step size d_1. More in general,

Definition 41 *A quantization is referred to as* monomial with parameters (λ, l) *(with $\lambda > 0$ and $l > 0$) iff the corresponding g–function is such that*

$$\lim_{x \to +\infty} \frac{g(x)}{x^{1/2\lambda}} = l\,.$$

Proposition 23 (Monomial quantizations) *If the quantization $\mathcal{I}$ is monomial with parameters (λ, l), then*

$$\lim_{\mathcal{E} \to +\infty} \frac{\mathcal{N}(\mathcal{E})}{\mathcal{E}^{1/2\lambda}} = c(\lambda, l)\,, \tag{7.20}$$

where[4] $c(\lambda, l) := \Gamma\Big(\frac{1}{2\lambda}\Big) \frac{l}{\lambda} (2\lambda\, e)^{1/2\lambda}$. *In particular,*

$$\lim_{\mathcal{E} \to +\infty} \frac{\mathcal{H}(\mathcal{E})}{\log \mathcal{E}} = \frac{1}{2\lambda}\,.$$

Proof. Thanks to Theorem 14, it is sufficient to prove that

$$\lim_{\beta \to 0^+} \frac{e^{\mathbb{H}\left(\mathcal{E}(\beta)\right)}}{\mathcal{E}(\beta)^{1/2\lambda}} = c(\lambda, l)\,.$$

According to the expression for $\mathbb{H}\big(\mathcal{E}(\beta)\big)$ given in equation (7.10), we have to analyze the behavior of $G(\beta)$ and $\mathcal{E}(\beta)$ as $\beta \to 0^+$. By Corollary 11 in Appendix A.6.1,

$$\lim_{\beta \to 0^+} \frac{G(\beta)}{l\, \Gamma\big(\frac{1}{2\lambda} + 1\big)\, \beta^{-\left(\frac{1}{2\lambda}+1\right)}} = 1\,.$$

By Theorem 21.*iii* we also have that

$$\lim_{\beta \to 0^+} \frac{G'(\beta)}{-l\big(\frac{1}{2\lambda} + 1\big)\, \Gamma\big(\frac{1}{2\lambda} + 1\big)\, \beta^{-\left(\frac{1}{2\lambda}+2\right)}} = 1\,.$$

Thus $\big($see equation (7.10b)$\big)$,

$$\begin{aligned} \lim_{\beta \to 0^+} \beta \mathcal{E}(\beta) &= \lim_{\beta \to 0^+} -\frac{2\beta\big(G(\beta) + \beta G'(\beta)\big)}{1 + 2\beta G(\beta)} = \\ &= -\Big(1 + \lim_{\beta \to 0^+} \frac{\beta G'(\beta)}{G(\beta)}\Big) = \\ &= \frac{1}{2\lambda}\,. \end{aligned}$$

[4] See the definition of the function $\Gamma(x)$ in Section 1.5.

To sum up,

$$\begin{aligned}\lim_{\beta\to 0^+} \frac{e^{\mathbb{H}\left(\mathcal{E}(\beta)\right)}}{\mathcal{E}(\beta)^{1/2\lambda}} &= \lim_{\beta\to 0^+} \frac{\left(1+2\beta G(\beta)\right)e^{\beta\mathcal{E}(\beta)}}{\mathcal{E}(\beta)^{1/2\lambda}} = \\ &= \lim_{\beta\to 0^+} \frac{2\beta G(\beta)e^{1/2\lambda}}{(\frac{1}{2\lambda})^{1/2\lambda}\beta^{-1/2\lambda}} = \\ &= 2\,l\,\Gamma\Big(\tfrac{1}{2\lambda}+1\Big)(2\lambda\,e)^{1/2\lambda} = \\ &= c(\lambda,l)\,,\end{aligned}$$

where the last equality holds because $\Gamma(x+1) = x\,\Gamma(x)$. ■

In the particular case of a uniform partition $\mathcal{I}$ with step size d_1, it holds that

$$\text{for} \quad \mathcal{E}\to+\infty\,, \quad \mathcal{N}(\mathcal{E}) \sim \frac{\sqrt{2\pi e}}{d_1}\sqrt{\mathcal{E}}\,.$$

We now introduce the class of *floating–point* quantizers as a generalization of the logarithmic quantizers. Let $\theta > 1$, (n_0, M) be a pair of positive integers and $\mathbb{R} \ni r_0 > 0$. Consider a symmetric quantization $\mathcal{I}$ such that $[0\,,\,r_0]$ is partitioned into n_0 intervals and $\forall\, h \in \mathbb{N}$, $[r_0\theta^h\,,\,r_0\theta^{h+1}]$ is partitioned into M intervals. For $x \geq r_0^2$ it holds that $r_0\theta^{\hat{h}-1} < \sqrt{x} \leq r_0\theta^{\hat{h}}$, where $\hat{h} = \lceil \log_\theta \frac{\sqrt{x}}{r_0} \rceil$. Hence, $\forall\, x \geq r_0^2$,

$$n_0 + M\Big(\lceil \log_\theta \tfrac{\sqrt{x}}{r_0}\rceil - 1\Big) \leq g(x) \leq n_0 + M\lceil \log_\theta \tfrac{\sqrt{x}}{r_0}\rceil$$

and

$$\lim_{x\to+\infty} \frac{g(x)}{(M/2)\log_\theta x} = 1\,.$$

Namely, the growth of the g–function is logarithmic and its asymptotic behavior only depends on M and θ.
The case $M = 1$ corresponds to a logarithmic quantization. When $\theta \in \mathbb{N}$ and $r_0 = \theta^{-m}$ for some $m \in \mathbb{N}$, notice the analogy between the described quantization and the one induced by a floating–point representation of the real numbers in the basis θ.
More in general,

Definition 42 *A quantization is referred to as* floating–point with parameters (M,θ) *(with $M > 0$ and $\theta > 1$) iff the corresponding g–function is such that*

$$\lim_{x\to+\infty} \frac{g(x)}{(M/2)\log_\theta x} = 1\,.$$

Remark 37 (Floating–point quantizations are logarithmic quantizations) *Notice that a quantization is floating–point with parameters (M,θ) if and only if it is floating–point with parameters $(1,\theta^{1/M})$. In fact, $(M/2)\log_\theta x = (1/2)\log_{\theta^{1/M}} x$.*

Proposition 24 (Floating–point quantizations) *If the quantization $\mathcal{I}$ is floating–point with parameters (M, θ), then*

$$\lim_{\mathcal{E}\to+\infty} \frac{\mathcal{N}(\mathcal{E})}{\log \mathcal{E}} = \frac{M}{\log \theta}\,. \tag{7.21}$$

In particular,

$$\lim_{\mathcal{E}\to+\infty} \frac{\mathcal{H}(\mathcal{E})}{\log\log \mathcal{E}} = 1\,.$$

Proof. The proof follows similar arguments to those we used to prove Proposition 23. The details are reported in Appendix A.6.3. ■

The theoretical interest of logarithmic quantizations has been repeatedly pointed out in various parts of this thesis and, with reference to the small–gain theorems, has been discussed in Example 3 of Section 2.3 as well as in Remark 24 of Section 5.3.2. Let us linger over logarithmic quantizations and let us analyze the controllers built on logarithmically quantized sets proposed in the previous chapters.

If $u(x) = q_\mu(Kx)$, where q_μ is a logarithmic quantization of $\mathbb{R}$ with parameters (u_0, θ), then the quantization induced by $u(x)$ is floating–point with parameters $(1, \theta)$. In fact, it is straightforward to see that $\mathcal{D}_\mathcal{I} = \{0\,, \pm\frac{u_0}{2|K|}\} \cup \{\pm \frac{u_0(\theta+1)}{2|K|}\theta^h \,|\, h \in \mathbb{N}\}$. In particular, the asymptotic behavior of the corresponding coding complexity function does not depend on u_0 and K. When $K = -a$ (i.e., $u(x)$ is a qdb–controller), the closed loop dynamics has the following properties (see case 2 of Example 4 in Section 2.3.1 and Fig. 2.7): for $|x| \geq \frac{u_0(\theta+1)}{4\theta} := x_0$, $|\varphi(x)| \leq \sigma|x|$, where $\sigma = |a|\frac{\theta-1}{\theta+1}$; for $|x| > x_0$, $|\varphi(x)| = \sigma|x|$ in correspondence of the discontinuity points of φ. It is useful to summarize these properties in the following

Definition 43 *Consider system (7.3), for $\sigma \in]0\,,\,1]$ and $x_0 > 0$, let $\theta := \frac{|a|+\sigma}{|a|-\sigma}$ and $x_h := x_0\theta^h$, $h \in \mathbb{N}$.*

ı) The quantized control law defined by

$$u(x) = \begin{cases} 0 & \text{if} \quad x \in [0\,, x_0[\\ -\text{sign}(a)\big(|a| + \sigma\big)x_h & \text{if} \quad x \in [x_h\,,\, x_{h+1}[\\ -u(-x) & \text{for} \quad x < 0 \end{cases} \tag{7.22}$$

is referred to as a standard logarithmic quantizer of parameter σ.

ıı) A closed loop dynamics $x^+ = \varphi(x)$ as in equation (7.3) is said to be standard logarithmic of parameter σ *iff $\exists\, x_0 > 0$ such that $\forall\, x$ with $|x| \geq x_0$, it holds that $|\varphi(x)| \leq \sigma|x|$ and for $|x| > x_0$ the set of the discontinuity points of $\varphi(x)$ is $\{\pm x_0\theta^h \,|\, h \in \mathbb{N} \setminus \{0\}\}$.*

By the definition it follows that, if φ is standard logarithmic of parameter σ, then

$$\begin{cases} \forall\, h \in \mathbb{N}\,, & \lim_{x\to x_h^+} \varphi(x) = -\,\text{sign}(a)\sigma x_h \\ \forall\, h \geq 1\,, & \lim_{x\to x_h^-} \varphi(x) = \text{sign}(a)\sigma x_h \\ \forall\, h \geq 1\,, & |\varphi(x_h)| = |\varphi(-x_h)| = \sigma x_h\,. \end{cases} \tag{7.23}$$

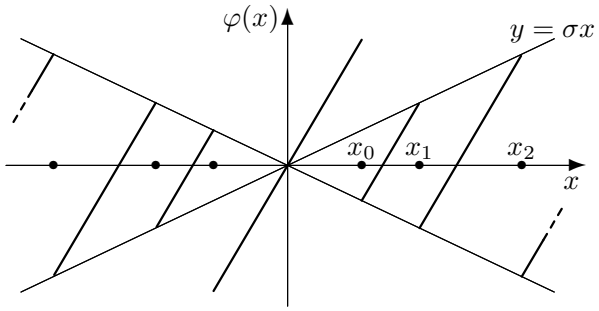

Figure 7.3: Closed loop dynamics generated by a standard logarithmic quantizer $u(x)$ of parameter $\sigma = 7/10$ ($a = 5/2$ and, accordingly, $\theta \simeq 1.78$).

The closed loop dynamics induced the control law in equation (7.22) is standard logarithmic of parameter σ (see also Fig. 7.3) . More in general, any control law $u(x)$ realizing a closed loop dynamics which is standard logarithmic of parameter σ induces a floating–point quantization of parameters $(1, \theta)$, where $\theta = \frac{|a|+\sigma}{|a|-\sigma}$. Accordingly, the corresponding coding cardinality function $\mathcal{N}$ is such that,

$$\text{for} \quad \mathcal{E} \to +\infty\,, \quad \mathcal{N}(\mathcal{E}) \sim \frac{1}{\log \theta} \log \mathcal{E}\,. \tag{7.24}$$

The decrease of σ assures a faster convergence rate (i.e., a better transient behavior) , but it produces the increase of the controller complexity (in fact, $\lim_{\sigma \to 0^+} \frac{1}{\log \theta} = +\infty$) : a detailed quantitative analysis of this trade off is offered in Section 7.4.2 .

Remark 38 *Notice the analogy between the notion of standard logarithmic dynamics of parameter σ and that of standard nonlinearity with natural external gain γ_e given in Definition 28 of Chapter 5.3 .*

7.4 Performance vs complexity

In this section, we aim at highlighting relations between the behavior of the energy and the complexity of the controller as measured by the function $\mathcal{N}$.

7.4.1 Lower bound for the minimal asymptotic energy

Suppose that a quantization of the state space is assigned and that the designer is only allowed to select the control values taken by the quantizer u within each element of the quantization. In a mean–square practical convergence problem, the goal is to design u so as to minimize $\mathcal{E}_\infty$ (see Definition 36) . Because of quantization, it is not possible to obtain arbitrarily small values for $\mathcal{E}_\infty$. In this section, a lower bound for $\mathcal{E}_\infty$ is provided which depends on the dynamics of the system through a, and on the assigned quantization through the corresponding coding complexity function $\mathcal{H}$.

Definition 44 *Let X be a random variable taking values in $\mathbb{R}$: if the probability distribution of X is absolutely continuous with respect to the Lebesgue measure and f is its density, the* differential entropy *of X is defined by*

$$h(X) := -\int_{\mathbb{R}} f(x) \log f(x) dx$$

(provided that the integral makes sense).

The differential entropy of a random variable X is a measure of the dispersion of its density (see [26]). The increase of the dispersion caused by the unstable dynamics can be counteracted by control. However, the amount of contraction which can be obtained by a quantized controller is bounded: this phenomenon can be quantified in terms of the following entropy inequality

Lemma 21 *Let $Y = X + U$ where X is a continuous random variable having differential entropy $h(X) \in \mathbb{R}$ and U is a discrete random variable. Then,*

$$h(Y) \geq h(X) - H(U).$$

Proof. See [46]. ■

This inequality, together with an Information theory inequality relating the differential entropy to the energy of the distribution, yields a lower bound on the attainable asymptotic value of the energy. Indeed we have the following

Theorem 15 (Lower bound for $\mathcal{E}_\infty$) *Consider the closed loop system* (7.3) *and let $\mathcal{H}$ be the coding complexity function associated to the quantization induced by u (which is supposed to satisfy* **A1***). If the distribution μ_0 of X_0 is absolutely continuous with respect to the Lebesgue measure and $h(X_0) \in \mathbb{R}$, then*

$$\limsup_{t \to +\infty} \mathcal{E}(\mu_t) \geq \mathcal{H}^{-1}(\log |a|).$$

In particular, if $\mathcal{P}$ contains any probability as μ_0 above, then the closed loop system (7.3) can not be $\mathcal{E}_\infty$–stable for any $\mathcal{E}_\infty < \mathcal{H}^{-1}(\log |a|)$.

Proof. Let $U_t := u(X_t)$, then

$$H(U_t) \leq \mathcal{H}\big(\mathcal{E}(\mu_t)\big), \tag{7.25}$$

in fact: $u(x)$ takes a constant value, say u_k, on each interval I_k, hence $H(U_t) \leq H(\mathbf{p}_{\mu_t,\mathcal{I}})$ $\big($with equality if $k_1 \neq k_2 \Rightarrow u_{k_1} \neq u_{k_2}\big)$ and $H(\mathbf{p}_{\mu_t,\mathcal{I}}) \leq \mathcal{H}\big(\mathcal{E}(\mu_t)\big)$ by definition of $\mathcal{H}$.
By $X_t = a^t X_0 + \sum_{i=0}^{t-1} a^{t-i-1} U_i$, it follows that

$$h(X_t) \geq h(X_0) + \log |a|^t - \sum_{i=0}^{t-1} \mathcal{H}\big(\mathcal{E}(\mu_i)\big), \tag{7.26}$$

in fact:

$$\begin{aligned} h(X_t) &\overset{(a)}{\geq} h(a^t X_0) - H\big(\textstyle\sum_{i=0}^{t-1} a^{t-i-1} U_i\big) \geq \\ &\overset{(b)}{\geq} h(X_0) + \log|a|^t - H(U_0, \dots, U_{t-1}) \geq \\ &\overset{(c)}{\geq} h(X_0) + \log|a|^t - \textstyle\sum_{i=0}^{t-1} H(U_i) \geq \\ &\overset{(d)}{\geq} h(X_0) + \log|a|^t - \textstyle\sum_{i=0}^{t-1} \mathcal{H}\big(\mathcal{E}(\mu_i)\big)\,, \end{aligned}$$

where inequality (a) follows by Lemma 21 , inequalities (b–c) follow by standard properties of the differential and of the discrete entropy (see [26]) , and inequality (d) follows by inequality (7.25) .

The variance of X_t is $\mathcal{E}(\mu_t) - m_t^2$, where $m_t := \mathbb{E}[X_t]$. Since the Gaussian distribution maximizes the differential entropy among the distributions having the same variance (see [26]) , then $h(X_t) \leq \frac{1}{2}\log\big((2\pi e)(\mathcal{E}(\mu_t) - m_t^2)\big)$ which, together with inequality (7.26) , yields

$$\mathcal{E}(\mu_t) \geq \frac{e^{2h(X_0)}}{2\pi e} \frac{|a|^{2t}}{e^{2\sum_{i=0}^{t-1}\mathcal{H}\left(\mathcal{E}(\mu_i)\right)}} + m_t^2\,.$$

If $\limsup_{t\to+\infty} \mathcal{E}(\mu_t) < +\infty$, then $\exists\, \bar{\mathcal{E}}$ such that $\forall\, t \in \mathbb{N}$, $\mathcal{E}(\mu_t) \leq \bar{\mathcal{E}}$. Because $\mathcal{H}$ is an increasing function, $\forall\, t \in \mathbb{N}$, $\mathcal{H}\big(\mathcal{E}(\mu_t)\big) \leq \mathcal{H}(\bar{\mathcal{E}})$, hence

$$\mathcal{E}(\mu_t) \geq \frac{e^{2h(X_0)}}{2\pi e} \Big(\frac{|a|^2}{e^{2\mathcal{H}(\bar{\mathcal{E}})}}\Big)^t + m_t^2\,.$$

Since the sequence $\{\mathcal{E}(\mu_t)\}_{t\in\mathbb{N}}$ is bounded, then $\frac{|a|^2}{e^{2\mathcal{H}(\bar{\mathcal{E}})}} \leq 1$: by the monotonicity of $\mathcal{H}$ and the fact that $\lim_{\mathcal{E}\to 0^+} \mathcal{H}(\mathcal{E}) = 0$ (see Theorem 14) , this is equivalent to $\bar{\mathcal{E}} \geq \mathcal{H}^{-1}(\log|a|)$. As a consequence, $\forall\, \epsilon > 0$, the sequence $\{\mathcal{E}(\mu_t)\}_{t\in\mathbb{N}}$ cannot be definitively upper bounded by $\mathcal{H}^{-1}(\log|a|) - \epsilon$: the thesis follows. ■

Remark 39 *Because of quantization, even if $\mathcal{E}(\mu_0) > 0$ is arbitrarily small (e.g., μ_0 is Gaussian with zero mean and arbitrarily small variance), the closed loop dynamics is so that such a density spreads over and the energy increases at least up to $\mathcal{H}^{-1}(\log|a|)$.*

7.4.2 The transient behavior and its relations with the controller complexity: the logarithmic regime

According to the framework in which the performance parameter $\mathcal{T}_e$ was defined, let us suppose that the closed loop dynamics $x^+ = \varphi(x)$ in equation (7.3) is so that $\varphi(J_0) \subseteq J_0$ for some $J_0 = [-r_0\,,\, r_0]$ and that, with $J_e := \mathbb{R} \setminus J_0$,

$$\forall\, \mu_0 \in \mathcal{P}\,,\ \lim_{t\to+\infty} \mathcal{E}_e(\mu_t) = 0\,.$$

The monotonic convergence to zero of the external energy is a desired property as it guarantees practical stability properties (rather than mere convergence) . We hence give the following

Definition 45 *Consider system* (7.3), *an external energy value* $\eta > 0$ *is said to be* φ–invariant *iff* $\forall\, \mu_0 \in \mathcal{P}r(\mathbb{R})$ *such that* $\mathcal{E}_e(\mu_0) \leq \eta$, *it holds that* $\mathcal{E}_e(\mu_1) \leq \eta$. *The external energy is said to be* monotonically decreasing *iff* $\forall\, \mu_0 \in \mathcal{P}r(\mathbb{R})$ *such that* $0 < \mathcal{E}_e(\mu_0) < +\infty$, *it holds that* $\mathcal{E}_e(\mu_1) < \mathcal{E}_e(\mu_0)$.

Clearly, the monotonic decrease of the external energy implies the φ–invariance of any $\eta > 0$.

Lemma 22 *Assume that* $\varphi(x)$ *in equation* (7.3) *is such that* $\varphi(J_0) \subseteq J_0$.

i) A necessary condition for the φ*–invariance of a value* $\eta > 0$ *of the external energy is*

$$\forall\, x \in J_e \quad \textit{such that} \quad |x| \geq \sqrt{\eta}\,,\ |\varphi(x)| \leq |x|\,.$$

ii) A necessary condition for the monotonic decrease of the external energy is

$$\forall\, x \in J_e\,,\ |\varphi(x)| \leq |x|\,.$$

iii) A sufficient condition for the monotonic decrease of the external energy is

$$\exists\, \sigma < 1 \quad \textit{such that} \quad \forall\, x \in J_e\,,\ |\varphi(x)| \leq \sigma |x|\,,$$

in this case $\mathcal{E}_e(\mu_1) \leq \sigma^2 \mathcal{E}_e(\mu_0)$.

Proof. See in Appendix A.6.4. ■

Remark 40 *The external energy, and hence the performance parameter* $\mathcal{T}_e$, *has been defined for closed loop dynamics ensuring that* $\varphi(J_0) \subseteq J_0$. *This requirement is not restrictive. In fact, similarly to Definition 45, one may introduce the more general concept of* φ*–invariance of the total energy as follows: an energy value* $\eta > 0$ *is said to be* φ–invariant *iff* $\forall\, \mu_0 \in \mathcal{P}r(\mathbb{R})$ *such that* $\mathcal{E}(\mu_0) \leq \eta$, *it holds that* $\mathcal{E}(\mu_1) \leq \eta$. *Then, following similar arguments to those used to prove Lemma 22, it is easy to show that a necessary condition for a total energy value* $\eta > 0$ *to be* φ*–invariant is that* $\forall\, x$ *such that* $|x| \leq \sqrt{\eta}$, $|\varphi(x)| \leq \sqrt{\eta}$ *and* $\forall\, x$ *such that* $|x| \geq \sqrt{\eta}$, $|\varphi(x)| \leq |x|$. *Namely, the interval* $[-\sqrt{\eta}\,, \sqrt{\eta}]$ *plays the role of* J_0. *In particular, the property* $\varphi\big([-\sqrt{\eta}\,, \sqrt{\eta}]\big) \subseteq [-\sqrt{\eta}\,, \sqrt{\eta}]$ *has to be guaranteed also in this (apparently) more general case.*

The main consequence of Lemma 22, is that requiring the φ–invariance of a single value of the external energy (and, a fortiori, requiring the monotonic decrease of the external energy) entails that the controller complexity is at least that of a logarithmic controller. Indeed,

Theorem 16 (φ–invariance implies logarithmic complexity) *Consider system (7.3), assume that* $u(x)$ *is such that* $\varphi(J_0) \subseteq J_0$ *for some* $J_0 = [-r_0\,, r_0]$ *and that there exists a value* $\eta > 0$ *of the external energy which is* φ*–invariant. Let* $\mathcal{N}(\mathcal{E})$ *be the coding cardinality function associated to* φ. *Then*

$$\forall\, \mathcal{E} > 0\,,\ \mathcal{N}(\mathcal{E}) \geq \mathcal{N}_{\rm sl}(\mathcal{E})\,, \tag{7.27}$$

where $\mathcal{N}_{\rm sl}(\mathcal{E})$ *is the coding cardinality function associated to a closed loop dynamics which is standard logarithmic of parameter* $\sigma = 1$. *In particular,* $\exists\, \mathcal{E}^* > 0$ *and* $C > 0$ *such that*

$$\forall\, \mathcal{E} > \mathcal{E}^*\,,\ \mathcal{N}(\mathcal{E}) \geq C \log \mathcal{E}\,. \tag{7.28}$$

Proof. Let $x_0 := \max\{r_0, \sqrt{\eta}\}$: by Lemma 22.*i* it holds that

$$\text{if} \quad |x| \geq x_0\,, \quad \text{then} \quad |\varphi(x)| \leq |x|\,. \tag{7.29}$$

For such an x_0 and $\sigma = 1$, let $u_{\text{sl}}(x)$ be the control law in equation (7.22). The resulting closed loop dynamics is standard logarithmic of parameter $\sigma = 1$. Denote by $g_{\text{sl}}(x)$ the g–function associated to the quantization induced by $u_{\text{sl}}(x)$ and by $\mathcal{N}_{\text{sl}}(\mathcal{E})$ the corresponding coding cardinality function. Also, let $g(x)$ be the g–function associated to the quantization induced by $u(x)$: because $\varphi(x)$ satisfies the property in equation (7.29), it is easy to recognize that $\forall\, x \geq 0$, $g_{\text{sl}}(x) \leq g(x)$. Therefore, inequality (7.27) follows by the monotonicity of $\mathcal{H}(\mathcal{E})$ with respect to the g–function (see Proposition 23).
The existence of $\mathcal{E}^* > 0$ and $C > 0$ such that inequality (7.28) holds is a consequence of inequality (7.27) and of the fact that, by equation (7.24), for $\mathcal{E} \to +\infty$, $\mathcal{N}_{\text{sl}}(\mathcal{E}) \sim \frac{1}{\log\theta} \log \mathcal{E}$. ■

The above theorem says that there is no control law $u(x)$ which induces a coarser quantization than a floating–point one and produces closed loop dynamics with monotonic decrease of the energy. Actually, floating–point quantizations do indeed permit to obtain such behaviors. This is the case, for instance, when $\varphi(x)$ is standard logarithmic of parameter $\sigma < 1$, in fact the monotonic decrease is guaranteed by Lemma 22.*iii*.

For standard logarithmic dynamics, a thorough analysis of the measure $\mathcal{T}_e$ of performance in the transient behavior is possible: this, together with the study of the relation between $\mathcal{T}_e$ and the behavior of $\mathcal{N}(\mathcal{E})$, is the subject of the remaining part of this section.
First, recall that $\mathcal{T}_e$ depends on the considered class of distributions $\mathcal{P}$ $\big($see equation (7.5)$\big)$. Furthermore, $\mathcal{T}_e$ depends on the parameter σ associated to the standard logarithmic dynamics: next Lemma 23 and Lemma 24 provide bounds for $\mathcal{T}_e$ that allow us to catch the main properties of such a dependence.

Lemma 23 (Lower bound for $\mathcal{T}_e$) *Consider system (7.3), assume that $\varphi(J_0) \subseteq J_0$ for some $J_0 = [-r_0\,, r_0]$ and that $\exists\, \sigma < 1$ such that $\forall\, x \in J_e$, $|\varphi(x)| \leq \sigma|x|$. Let $\mathcal{P}$ be any class of admissible distributions (see Section 7.2), then*

$$\mathcal{T}_e \geq -2\log\sigma\,. \tag{7.30}$$

Proof. By Lemma 22.*iii*, $\mathcal{E}_e(\mu_t) \leq \sigma^2 \mathcal{E}_e(\mu_{t-1})$. Therefore, $\forall\, \mu_0 \in \mathcal{P}$, $\mathcal{E}_e(\mu_t) \leq \sigma^{2t} \mathcal{E}_e(\mu_0)$ so that $\frac{\log \mathcal{E}_e(\mu_t)}{t} \leq 2\log\sigma + \frac{\log \mathcal{E}_e(\mu_0)}{t}$: the thesis follows. ■

In general, the lower bound provided by the lemma is conservative. For instance, in the presence of a uniform partition of step size d_1, it is possible to design a control law $u(x)$ such that $\forall\, x \in \mathbb{R}$, $\varphi(x) \in J_0 := \big[-\frac{|a|d_1}{2}\,, \frac{|a|d_1}{2}\big]$: in this case $\mathcal{T}_e = +\infty$. In the logarithmic regime, instead, the following upper bound can be proved:

Lemma 24 (Upper bound for $\mathcal{T}_e$) *Consider system (7.3), assume that $\varphi(J_0) \subseteq J_0$ for some $J_0 = [-r_0\,, r_0]$ and that $\varphi(x)$ is standard logarithmic of parameter $\sigma < 1$. Let*

$m(\sigma) := \left\lceil -\log_{\frac{|a|+\sigma}{|a|-\sigma}} \sigma \right\rceil$ *and* $\mathcal{P} = \mathcal{P}_{\text{all}}$ *(see equation* (7.4)*), then*

$$\mathcal{T}_e \leq 2m(\sigma) \log \frac{|a|+\sigma}{|a|-\sigma} \,. \tag{7.31}$$

Proof. The proof is much more involved than the one for the lower bound and it is given in Appendix A.6.4 . ■

Thus, under the assumptions of Lemma 24 ,

$$\begin{aligned} \mathcal{T}_e \leq \;& 2m(\sigma) \log \tfrac{|a|+\sigma}{|a|-\sigma} < 2\Big(-\log_{\frac{|a|+\sigma}{|a|-\sigma}} \sigma + 1 \Big) \log \tfrac{|a|+\sigma}{|a|-\sigma} = \\ =\;& -2\log\sigma + 2\log \tfrac{|a|+\sigma}{|a|-\sigma} := f_1(\sigma) \,, \end{aligned} \tag{7.32}$$

whereas, under the assumptions of Lemma 23 ,

$$\mathcal{T}_e \geq -2\log\sigma := f_2(\sigma) \,. \tag{7.33}$$

Both $f_1(\sigma)$ and $f_2(\sigma)$ are invertible functions on an interval $]0\,,\,\sigma_0[$, for sufficiently small $\sigma_0 > 0$. When both Lemma 23 and Lemma 24 hold, the behavior of $\mathcal{T}_e$ for $\sigma \to 0^+$ is well described by the function f_2 and it is possible to establish a relation between $\mathcal{T}_e$ and the behavior of the coding cardinality function.

Theorem 17 (Asymptotic trade off between complexity and performance) *Consider a family of systems*

$$\big\{x^+ = \varphi_\sigma(x) := ax + u_\sigma(x)\big\}_{\sigma \in]0\,,\,1[}$$

of the type in equation (7.3) *such that the following properties hold:*

a) $\forall\, \sigma \in]0\,,\,1[\,,\ \varphi_\sigma$ *is standard logarithmic of parameter* σ *;*
b) $\exists\, J_0 = [-r_0\,,\,r_0]$ *such that* $\forall\, \sigma \in]0\,,\,1[\,,\ \varphi_\sigma(J_0) \subseteq J_0$ *;*
c) $\forall\, x \in J_e$ *and* $\forall\, \sigma \in]0\,,\,1[\,,\ |\varphi_\sigma(x)| \leq \sigma |x|$ *.*

Consider the functions $f_1(\sigma)$ *and* $f_2(\sigma)$ *defined in equations* (7.32) *and* (7.33) *, respectively. Let* $\mathcal{P} = \mathcal{P}_{\text{all}}$ *(see equation* (7.4)*), denote by* $\mathcal{T}_e(\sigma)$ *the measure of the transient behavior for system* $x^+ = \varphi_\sigma(x)$ *and by* $\mathcal{N}_\sigma(\mathcal{E})$ *the coding cardinality function associated to the quantization induced by* $u_\sigma(x)$ *. Then,*

ı) $f_2(\sigma) \leq \mathcal{T}_e(\sigma) < f_1(\sigma)$ *, with* $\lim_{\sigma\to 0^+}(f_2 - f_1)(\sigma) = 0$ *and* $\lim_{\sigma\to 0^+} \frac{f_1(\sigma)}{f_2(\sigma)} = 1$ *;*
ıı) for $\mathcal{E} \to +\infty$ *,* $\mathcal{N}_\sigma(\mathcal{E}) \sim C(\sigma) \log \mathcal{E}$ *, where*

$$C(\sigma) = \frac{1}{\log \frac{|a|+\sigma}{|a|-\sigma}} \tag{7.34}$$

is such that, with $C_i(\mathcal{T}_e) := C\big(f_i^{-1}(\mathcal{T}_e)\big)$*,* $i = 1\,,2\,,$

$$C_1\big(\mathcal{T}_e(\sigma)\big) < C(\sigma) \leq C_2\big(\mathcal{T}_e(\sigma)\big) \tag{7.35}$$

and

$$\textit{for} \quad \mathcal{T}_e \to +\infty\,, \quad C_1(\mathcal{T}_e) \sim C_2(\mathcal{T}_e) \sim \frac{|a|}{2} e^{\mathcal{T}_e/2}\,.$$

In particular, if $\mathcal{T}_e(\sigma)$ *is an invertible function of* σ*, then*

$$\begin{cases} \textit{for} \quad \mathcal{E} \to +\infty\,, \quad \mathcal{N}(\mathcal{E}) \sim C(\mathcal{T}_e) \log \mathcal{E} \\ \textit{for} \quad \mathcal{T}_e \to +\infty\,, \quad C(\mathcal{T}_e) \sim \frac{|a|}{2} e^{\mathcal{T}_e/2}\,, \end{cases} \tag{7.36}$$

where, with slight abuse of notation, we let $C(\mathcal{T}_e) := C\big(\sigma(\mathcal{T}_e)\big)$*.*

Proof. It is a consequence of the fact that $\forall\, \sigma \in]0\,,\, 1[$, the hypotheses of both Lemma 23 and Lemma 24 hold for φ_σ. The details of the proof are reported in Appendix A.6.4. ■

In brief, system (7.3), when controlled so that the closed loop dynamics is standard logarithmic, shows an asymptotic trade off between complexity and performance of the type in equation (7.36).

7.5 Example: analysis of performance and complexity for nested quantized control laws

The class of closed loop dynamics generated by nest–structured quantized control laws has been studied in [45]. Let us analyze them in terms of the performance and complexity parameters introduced so far.

Let the scalar dynamical system

$$x^+ = ax + u$$

be such that $a \in \mathbb{Z}$, $|a| > 1$.
Consider a nest–structured quantized control law $u : \mathbb{R} \to \mathcal{U}$ defined as follows: let $\theta > 1$ be in the form $\theta = 1 + \frac{\nu}{|a|}$ for some $\nu \in \mathbb{N} \setminus \{0\}$ and

$$u(x) := \begin{cases} 2\big\lfloor \frac{1-ax}{2} \big\rfloor & \text{if} \quad x \in [0\,,\, \theta] \\ \theta^i \cdot u\big(\frac{x}{\theta^i}\big) & \text{if} \quad x \in]\theta^i\,,\, \theta^{i+1}]\,,\ i \in \mathbb{N} \setminus \{0\} \\ -u(-x) & \text{if} \quad x < 0\,. \end{cases}$$

As usual, denote by $\varphi(x) := ax + u(x)$ the corresponding closed loop dynamics. Let $J_i := [-\theta^i\,,\, \theta^i]$, $i \in \mathbb{N}$, and $J_{-1} := \emptyset$: $\varphi(x)$ is such that $\varphi(J_0) \subseteq J_0$ and

$$\forall\, i \geq 1\,, \quad \varphi(J_i) \subseteq J_{i-1} \tag{7.37}$$

(see Fig. 7.4).
With reference to the terminology introduced in Section 7.2, we consider two classes of initial distributions.

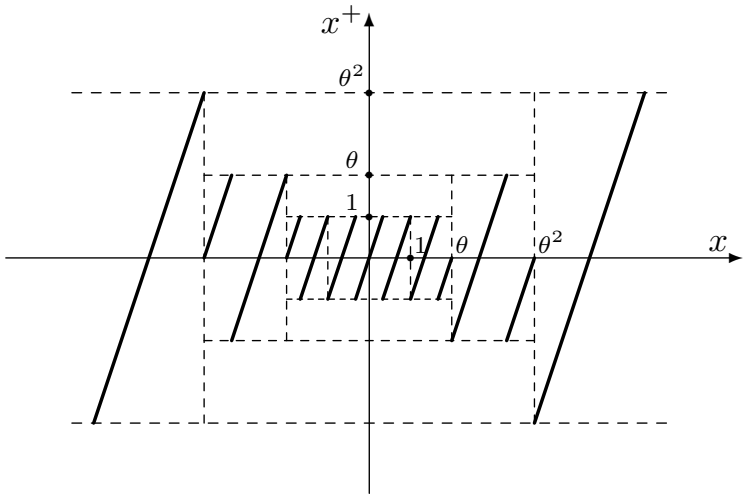

Figure 7.4: The closed loop dynamics generated by $u(x)$ when $a = 3$ and $\theta = 2$.

Case 1: $\mathcal{P} = \mathcal{P}_{\text{pwc}}$, where

$$\mathcal{P}_{\text{pwc}} := \left\{ \mu \in \mathcal{P}r(\mathbb{R}) \,|\, \mathcal{E}(\mu) < +\infty \quad \text{and} \quad \mu = f(x)dx \quad \text{with} \quad f(x) = \textstyle\sum_{i=0}^{+\infty} \alpha_i \cdot \chi_{J_i \setminus J_{i-1}}(x) \right\},$$

namely, $\mathcal{P}_{\text{pwc}}$ is the set of probability distributions having finite energy and constant density on $J_i \setminus J_{i-1}$, $\forall\, i \in \mathbb{N}$.

Case 2: $\mathcal{P} = \mathcal{P}_{\text{all}}$ $\big($see equation (7.4)$\big)$.

Both classes are closed under the dynamics φ, therefore admissible.

7.5.1 Complexity analysis: $\mathcal{N}(\mathcal{E})$

The controller complexity is represented by the coding cardinality function $\mathcal{N}(\mathcal{E}) = e^{\mathcal{H}(\mathcal{E})}$ associated to the quantization induced by $u(x)$. Such a quantization is floating–point with parameters $\big(M\,,\,\theta\big)$, where, by definition of θ,

$$M = \left\lceil \frac{\nu}{2} \right\rceil = \left\lceil \frac{|a|(\theta - 1)}{2} \right\rceil . \tag{7.38}$$

Therefore, by Proposition 24,

$$\text{for} \quad \mathcal{E} \to +\infty\,, \quad \mathcal{N}(\mathcal{E}) \sim \frac{\left\lceil \frac{|a|(\theta-1)}{2} \right\rceil}{\log \theta} \log \mathcal{E}\,. \tag{7.39}$$

7.5.2 Performance analysis: $\mathcal{E}_\infty$ and $\mathcal{T}_e$

In both cases of $\mathcal{P} = \mathcal{P}_{\text{pwc}}$ and $\mathcal{P} = \mathcal{P}_{\text{all}}$, we compute the transient performance parameter $\mathcal{T}_e$ and analyze the steady–state performance by finding the minimal value for $\mathcal{E}_\infty$ such that the dynamics φ is $\mathcal{E}_\infty$–converging. As for $\mathcal{T}_e$, with $J_e = \mathbb{R} \setminus [-1\,,\,1]$ we are in the right framework to define and analyze the behavior of the external energy.

Case 1: $\mu \in \mathcal{P}_{\rm pwc}$

For $\mu \in \mathcal{P}_{\rm pwc}$ we have

Proposition 25 *For any initial distribution $\mu_0 \in \mathcal{P}_{\rm pwc}$ and $\forall\, t \geq 0$ it holds that $\mathcal{E}(\mu_t) \geq \frac{1}{3}$ and*

$$\lim_{t \to +\infty} \mathcal{E}(\mu_t) = \frac{1}{3}\,.$$

In particular, the closed loop dynamics $x^+ = \varphi(x)$ is $\mathcal{E}_\infty$–converging if and only if $\mathcal{E}_\infty \geq \frac{1}{3}$. Moreover,

$$\mathcal{T}_e = \log(\theta^2 + \theta + 1)\,. \tag{7.40}$$

In order to prove the proposition, it is useful to give an explicit representation of the *Perron–Frobenius* operator associated to φ and describing the dynamics of the probability distributions $\mu \in \mathcal{P}_{\rm pwc}$. Namely, to represent the linear operator $\mathcal{F}$ such that $\mu_{t+1} = \mathcal{F}(\mu_t)$. Suppose that at time t the state of the process is distributed according to $\mu_t \in \mathcal{P}_{\rm pwc}$: we use the infinite vector $\mathbf{p}(t) := \{p_i(t)\}_{i \in \mathbb{N}}$, where $p_i(t) := \mu_t(J_i \backslash J_{i-1})$, to identify μ_t. In this case, the Perron–Frobenius operator can be represented as an infinite matrix $\mathcal{F} := \{\mathcal{F}_{i,j}\}_{(i,j) \in \mathbb{N}^2}$, that is:

$$\mathbf{p}(t+1) = \mathcal{F}\mathbf{p}(t)\,,$$

where this notation means that $\forall\, i \in \mathbb{N},\ p_i(t+1) = \sum_{j=0}^{+\infty} \mathcal{F}_{i,j}\, p_j(t)$. Simple calculations show that

$$\begin{cases} \mathcal{F}_{0,0} = 1 & \\ \mathcal{F}_{0,j} = \theta^{-j+1} & \text{if} \quad j \geq 1 \\ \mathcal{F}_{i,j} = (\theta - 1)\,\theta^{i-j} & \text{if} \quad 1 \leq i < j \\ \mathcal{F}_{i,j} = 0 & \text{otherwise}\,. \end{cases} \tag{7.41}$$

Let $\vec{\zeta} := \{\zeta_i\}_{i \in \mathbb{N}}$, where

$$\begin{cases} \zeta_0 = \frac{1}{3} & \\ \zeta_i = \frac{\theta^2 + \theta + 1}{3}\,\theta^{2(i-1)} & \text{for} \quad i \geq 1\,. \end{cases} \tag{7.42}$$

For a distribution $\mu \in \mathcal{P}_{\rm pwc}$ represented by the probability vector $\mathbf{p} = \{p_i\}_{i \in \mathbb{N}}$, it is straightforward to see that

$$\mathcal{E}(\mu) = \vec{\zeta} \cdot \mathbf{p} := \sum_{i=0}^{+\infty} \zeta_i p_i\,.$$

Given $\mu_0 \in \mathcal{P}_{\rm pwc}$, the energy $\mathcal{E}(\mu_t)$ is shortly denoted by $\mathcal{E}_t$. The energy at time $t+1$ is given by the expression

$$\mathcal{E}_{t+1} = \vec{\zeta} \cdot \mathcal{F}\mathbf{p}(t)\,.$$

Proof of Proposition 25. Let us start by analyzing the dependence of $\mathcal{E}_{t+1}$ from $\mathcal{E}_t$ and $\mathbf{p}(t)$. Denote by $\vec{e}^{(i)}$ the infinite vector such that[5] $e^{(i)}_j = \delta_{ij}$, then $\mathcal{E}_{t+1} =$

[5] Where $\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise.} \end{cases}$

$\vec{\zeta}\cdot\mathcal{F}\sum_{i=0}^{+\infty}p_i(t)\vec{e}^{(i)}$. Using equations (7.41) and (7.42), a direct computation allows one to show that

$$\vec{\zeta}\cdot\mathcal{F}\vec{e}^{(i)}=\begin{cases}\frac{1}{3} & \text{for}\quad i=0\\ \frac{1}{3}\theta^{2(i-1)} & \text{for}\quad i\geq 1\,.\end{cases}$$

We can hence explicitly write $\mathcal{E}_t$ and $\mathcal{E}_{t+1}$ in the form

$$\mathcal{E}_t=\frac{1}{3}\,p_0(t)+\frac{\theta^2+\theta+1}{3}\sum_{i=1}^{+\infty}\theta^{2(i-1)}p_i(t) \tag{7.43}$$

and

$$\mathcal{E}_{t+1}=\frac{1}{3}\,p_0(t)+\frac{1}{3}\sum_{i=1}^{+\infty}\theta^{2(i-1)}p_i(t)\,. \tag{7.44}$$

Therefore, $\mathcal{E}_{t+1}-\frac{1}{3}\,p_0(t)=\frac{1}{\theta^2+\theta+1}\big(\mathcal{E}_t-\frac{1}{3}p_0(t)\big)$ and

$$\mathcal{E}_{t+1}=\frac{1}{\theta^2+\theta+1}\mathcal{E}_t+\frac{(\theta^2+\theta)p_0(t)}{3(\theta^2+\theta+1)}\,. \tag{7.45}$$

Since $\theta>1$, then $\forall\, i\geq 1$, $(\theta^2+\theta+1)\,\theta^{2(i-1)}>1$. Therefore, by equation (7.43), $\mathcal{E}_t\geq\frac{1}{3}\sum_{i=0}^{+\infty}p_i(t)=\frac{1}{3}$.
By equation (7.45) we immediately get

$$\mathcal{E}_{t+1}-\frac{1}{3}=\frac{\mathcal{E}_t-\frac{1}{3}}{\theta^2+\theta+1}+\frac{(\theta^2+\theta)\big(p_0(t)-1\big)}{3(\theta^2+\theta+1)}\,.$$

As $p_0(t)-1\leq 0$, then $\mathcal{E}_{t+1}-\frac{1}{3}\leq\frac{1}{\theta^2+\theta+1}\left(\mathcal{E}_t-\frac{1}{3}\right)$ and, because $\mathcal{E}_t-\frac{1}{3}\geq 0$, it holds that $\lim_{t\to+\infty}\mathcal{E}_t=\frac{1}{3}$.

Let us show that $\mathcal{T}_e=\log(\theta^2+\theta+1)$. Given $\mu_0\in\mathcal{P}_{\text{pwc}}$, the external energy $\mathcal{E}_e(\mu_t)$ is shortly denoted by $\mathcal{E}_{e\,t}$. Notice that

$$\mathcal{E}_{e\,t}=\mathcal{E}_t-\zeta_0p_0(t)=\sum_{i=1}^{+\infty}\zeta_ip_i(0) \tag{7.46}$$

thus $\big($see equation (7.43)$\big)$,

$$\mathcal{E}_{e\,t}=\frac{\theta^2+\theta+1}{3}\sum_{i=1}^{+\infty}\theta^{2(i-1)}p_i(t)\,. \tag{7.47}$$

At time $t+1$,

$$\mathcal{E}_{e\,t+1}\overset{\text{(a)}}{=}\mathcal{E}_{t+1}-\zeta_0p_0(t+1)\overset{\text{(b)}}{=}\mathcal{E}_{t+1}-\frac{1}{3}\big(p_0(t)+\sum_{i=1}^{+\infty}\theta^{-i+1}p_i(t)\big)\overset{\text{(c)}}{=}\frac{1}{3}\sum_{i=1}^{+\infty}(\theta^{2(i-1)}-\theta^{-i+1})p_i(t)\,,$$

where equality (a) follows by equation (7.46), in equality (b) we used $\zeta_0 = 1/3$ (see equation (7.42)) and the expression for $\mathcal{F}$ given in equation (7.41), and equality (c) follows by equation (7.44). Thus (see equation (7.47)),

$$\mathcal{E}_{e\,t+1} = \frac{\mathcal{E}_{e\,t}}{\theta^2+\theta+1} - \frac{1}{3}\sum_{i=1}^{+\infty} \theta^{-i+1} p_i(t)$$

and $\mathcal{E}_{e\,t+1} \leq \frac{1}{\theta^2+\theta+1}\mathcal{E}_{e\,t}$ from which it immediately follows that

$$\mathcal{T}_e \geq \log(\theta^2+\theta+1)\,.$$

To prove the equality, let us restrict to consider initial probability distributions $\mu_0 \in \mathcal{P}_{\mathrm{pwc}}$ having exponentially decaying masses. More precisely, let $\mathbf{p}(0)$ be such that $p_i(0) = (1-q)q^i$ for some $q \in]0\,,\,1[$. The energy of this type of distribution is

$$\mathcal{E}_0 = \vec{\zeta}\cdot\mathbf{p}(0) = \tfrac{(1-q)(\theta^2+\theta+1)}{3}\big(\tfrac{1}{\theta^2+\theta+1} + \textstyle\sum_{i=1}^{+\infty} q^i\theta^{2(i-1)}\big)$$

and it is finite if and only if $q < \frac{1}{\theta^2}$.
Given such a probability distribution μ_0, the external energy at time t, which depends on q, is denoted by $\mathcal{E}_{e\,t}(q)$. The thesis is achieved by showing that

$$-\sup_{q\in\,]0\,,\,\frac{1}{\theta^2}[}\ \limsup_{t\to+\infty}\frac{\log\mathcal{E}_{e\,t}(q)}{t} = \log(\theta^2+\theta+1)\,.$$

To this aim, it is useful to introduce the following block decomposition: let $\mathbf{p}^\star := \{p_i\}_{i\in\mathbb{N}\setminus\{0\}}$ and $\vec{\zeta}^\star := \{\zeta_i\}_{i\in\mathbb{N}\setminus\{0\}}$. For $q \in]0\,,\,\frac{1}{\theta^2}[$,

$$\mathbf{p}(1) = \mathcal{F}\mathbf{p}(0) = \mathcal{F}\begin{pmatrix} p_0(0) \\ \mathbf{p}^\star(0)\end{pmatrix} = \begin{pmatrix} p_0(1) \\ \lambda(q)\mathbf{p}^\star(0)\end{pmatrix}\,,$$

where

$$\lambda(q) = \frac{q(\theta-1)}{\theta-q}\,, \tag{7.48}$$

and

$$\mathbf{p}(t) = \mathcal{F}^t\mathbf{p}(0) = \begin{pmatrix} p_0(t) \\ \lambda(q)^t\mathbf{p}^\star(0)\end{pmatrix}\,.$$

Therefore, $\mathcal{E}_{e\,t} = \vec{\zeta}^\star\cdot\mathbf{p}^\star(t) = \lambda(q)^t\vec{\zeta}^\star\cdot\mathbf{p}^\star(0)$. Since $\vec{\zeta}^\star\cdot\mathbf{p}^\star(0)$ does not depend on t we get

$$\limsup_{t\to+\infty}\frac{\log\mathcal{E}_{e\,t}}{t} = \log\lambda(q)\,.$$

As

$$\lim_{q\to\frac{1}{\theta^2}}\lambda(q) = \frac{1}{\theta^2+\theta+1}$$

(see equation (7.48)), the thesis follows. ■

Case 2: $\mu \in \mathcal{P}_{\rm all}$

For $\mu \in \mathcal{P}_{\rm all}$ we have

Proposition 26 *It holds that*

$$\sup_{\mu_0 \in \mathcal{P}_{\rm all}} \limsup_{t \to +\infty} \mathcal{E}(\mu_t) = 1 \,,$$

in particular, the closed loop dynamics $x^+ = \varphi(x)$ *is* $\mathcal{E}_\infty$*–converging if and only if* $\mathcal{E}_\infty \geq 1$*. Moreover,*

$$\mathcal{T}_e \geq \log \theta^2$$

and if $1 - a - \nu$ *is an even number,*

$$\mathcal{T}_e = \log \theta^2 \,. \tag{7.49}$$

Proof. Let us start by proving the results on $\mathcal{T}_e$. For $k \in \mathbb{N}$, let $L_k := J_k \setminus J_{k-1}$. Namely, $L_0 = J_0$ and, for $k \geq 1$, $L_k = [-\theta^k, -\theta^{k-1}[\cup]\theta^{k-1}, \theta^k]$. Clearly, $L_i \cap L_j = \emptyset$ for $i \neq j$, and $J_e = \bigcup_{k \geq 1} L_k$. Also, by equation (7.37),

$$\begin{cases} \text{if} \quad k \leq t \,, & \varphi^t(L_k) \subseteq J_0 \\ \text{if} \quad k \geq t+1 \,, & \varphi^t(L_k) \subseteq J_{k-t} \,. \end{cases} \tag{7.50}$$

For any given $\mu_0 \in \mathcal{P}_{\rm all}$, let

$$\mu_0^\delta := \sum_{k \in \mathbb{N}} \mu_0(L_k) \delta_{\theta^k} \,.$$

Let us show that $\mu_0^\delta \in \mathcal{P}_{\rm all}$, in fact:

$$\mathcal{E}(\mu_0) = \textstyle\int_{\mathbb{R}} x^2 d\mu_0 = \sum_{k \in \mathbb{N}} \int_{L_k} x^2 d\mu_0 \geq 0 + \sum_{k \geq 1} \theta^{2(k-1)} \mu_0(L_k) = \frac{1}{\theta^2} \big(\mathcal{E}(\mu_0^\delta) - \mu_0(L_0) \big) \,,$$

thus $\mathcal{E}(\mu_0^\delta) \leq \theta^2 \mathcal{E}(\mu_0) + \mu_0(L_0) < +\infty$.
Let $\mu_0 \in \mathcal{P}_{\rm all}$, it holds that

$$\mathcal{E}_e(\mu_t) \leq \frac{\mathcal{E}_e(\mu_0^\delta)}{\theta^{2t}} \,, \tag{7.51}$$

in fact:

$$\begin{aligned} \mathcal{E}_e(\mu_t) = & \int_{J_e} x^2 d\mu_t \\ = & \int_{\varphi^{-t}(J_e)} \big(\varphi^t(x)\big)^2 d\mu_0 \\ = & \sum_{k \geq 1} \int_{\varphi^{-t}(J_e) \cap L_k} \big(\varphi^t(x)\big)^2 d\mu_0 \\ \overset{\text{(a)}}{=} & \sum_{k \geq t+1} \int_{\varphi^{-t}(J_e) \cap L_k} \big(\varphi^t(x)\big)^2 d\mu_0 \\ \overset{\text{(b)}}{\leq} & \sum_{k \geq t+1} \theta^{2(k-t)} \mu_0(L_k) \\ = & \frac{1}{\theta^{2t}} \big(\mathcal{E}_e(\mu_0^\delta) - \textstyle\sum_{k=1}^t \theta^{2k} \mu_0(L_k) \big) \\ \leq & \frac{\mathcal{E}_e(\mu_0^\delta)}{\theta^{2t}} \,, \end{aligned}$$

where both equality (a) and inequality (b) follow by equation (7.50). By inequality (7.51),

$$\forall\, \mu_0 \in \mathcal{P}_{\rm all}\,, \quad \limsup_{t\to+\infty} \frac{\log \mathcal{E}_e(\mu_t)}{t} \leq \log \frac{1}{\theta^2}$$

and hence $\mathcal{T}_e \geq \log \theta^2$.

Let us show that, when $1-a-\nu$ is an even number, also the converse inequality holds. In this case it holds that

$$\forall\, k \geq 1\,, \quad \varphi(\theta^k) = \theta^{k-1}\,. \tag{7.52}$$

Similarly to the case $\mathcal{P} = \mathcal{P}_{\rm pwc}$, let us restrict to consider initial probability distributions in the form $\mu_0 = \frac{q}{1-q}\sum_{k\geq 1} q^k \delta_{\theta^k}$ for some $q \in]0\,,\,1[$. The energy of this type of distribution is $\mathcal{E}(\mu_0) = \frac{q}{1-q}\sum_{k\geq 1}(q\theta^2)^k$ and it is finite if and only if $q < \frac{1}{\theta^2}$. Given such a probability distribution μ_0,

$$\begin{aligned} \mathcal{E}_e(\mu_t) = & \ \ \tfrac{q}{1-q}\textstyle\sum_{k\geq t+1} q^k \theta^{2(k-t)} \\ = & \ \ \tfrac{(q\theta)^2}{(1-q)(1-q\theta^2)} q^t\,, \end{aligned}$$

where in the first equality we take advantage of equation (7.52). Therefore,

$$\limsup\nolimits_{t\to+\infty} \tfrac{\log \mathcal{E}_e(\mu_t)}{t} = \log q$$

and

$$\mathcal{T}_e \leq - \sup_{q\in\,]\,0\,,\,\frac{1}{\theta^2}\,[} \limsup_{t\to+\infty} \frac{\log \mathcal{E}_e(\mu_t)}{t} = \log \theta^2\,.$$

Finally, let us prove that $\sup\limits_{\mu_0\in\mathcal{P}_{\rm all}} \limsup\limits_{t\to+\infty} \mathcal{E}(\mu_t) = 1$. For any $\mu_0 \in \mathcal{P}_{\rm all}$,

$$\mathcal{E}(\mu_t) = \int_{-1}^{1} x^2 d\mu_t + \mathcal{E}_e(\mu_t) \leq \mu_t(J_0) + \mathcal{E}_e(\mu_t) \leq 1 + \mathcal{E}_e(\mu_t)\,.$$

By inequality (7.51), $\lim_{t\to+\infty} \mathcal{E}_e(\mu_t) = 0$, therefore $\sup\limits_{\mu_0\in\mathcal{P}_{\rm all}} \limsup\limits_{t\to+\infty} \mathcal{E}(\mu_t) \leq 1$.

On the other hand let μ be the uniform distribution on J_0 with unitary total mass. The measure μ is ergodic for the closed loop dynamics $x^+ = \varphi(x)$ restricted to J_0 (see [71]), this means that

$$\forall\, \epsilon > 0 \quad \text{and for } \mu\text{–almost all } x_0 \in J_0\,, \quad \lim_{t\to+\infty} \frac{\sum_{i=0}^{t-1} \chi_{[1-\epsilon\,,\,1]}\big(\varphi^i(x_0)\big)}{t} = \epsilon\,.$$

That is, the frequency with which a trajectory starting from x_0 visits the interval $[1-\epsilon\,,\,1]$ converges to ϵ. Hence, for such an x_0, $\mu_0 = \delta_{x_0}$ is such that $\limsup_{t\to+\infty} \mathcal{E}_t \geq (1-\epsilon)^2$. This yields the result. ■

As it is expected, the worst case decaying rate of the external energy is smaller in the general case (i.e., there are distributions, not belonging to $\mathcal{P}_{\rm pwc}$, whose external energy has a decaying rate which is slower than the decaying rate of any $\mu_0 \in \mathcal{P}_{\rm pwc}$). Also the steady–state performance are worst in the general case.

Remark 41 *In case* $1-a-\nu$ *is odd, because* $\mathcal{P}_{\rm pwc} \subset \mathcal{P}_{\rm all}$ *, we can conclude that* $\log\theta^2 \leq \mathcal{T}_e \leq \log(\theta^2+\theta+1)$ *. As it will be clear in next Section 7.5.3, this guarantees that the asymptotic trade off between complexity and performance does not depend on the parity of* $1-a-\nu$ *.*

7.5.3 Performance vs complexity

Relation (7.39) compared with either (7.40) or (7.49) shows an asymptotic trade off between complexity and performance of type

$$\begin{cases} \text{for} \quad \mathcal{E} \to +\infty\,, \quad \mathcal{N}(\mathcal{E}) \sim C(\mathcal{T}_e)\log\mathcal{E} \\ \text{for} \quad \mathcal{T}_e \to +\infty\,, \quad C(\mathcal{T}_e) \sim |a|\frac{e^{\mathcal{T}_e/2}}{\mathcal{T}_e}\,. \end{cases}$$

Notice that, in this case, $\mathcal{T}_e \to +\infty$ corresponds to $\theta \to +\infty$: thus, in order to improve performance it is necessary to change both the parameters of the floating–point quantization (see equation (7.38)).

Conclusion

We have presented various results on the practical stabilization problem for linear systems under arbitrarily assigned input and/or output quantization. Novel analysis techniques to study controlled invariance have been proposed and particular attention has been turned to optimality of the steady–state performance (i.e., in finding minimal invariant sets within which the state of the system can be ultimately bounded). The small–gain approach has offered systematic tools for the control synthesis, those based on ℓ_1 theory are particularly suitable to deal with quantized controls. We have then analyzed how the closed loop performance changes as the complexity of the quantized controller varies. Information theory provided us with suitable tools to carry out fundamental relations between performance and complexity. Some directions of research which may be further explored are the following:

- For multi–input systems only input quantization has been considered in this thesis. The extension of the small–gain approach to include quantized multi–input systems under arbitrarily assigned quantized measurements is certainly a matter of primary importance. More in general, in analogy with the framework proposed in [122] for hybrid systems, we expect that the small–gain approach may offer the tools enabling one to work out a systematic theory for the synthesis of practically stabilizing dynamic controllers under assigned input and output quantization.

- In the case of multi–input systems, the control synthesis results based on ℓ_1 theory need to be extended so that the design of the controller can be done with a pure ℓ_1 approach (whereas, in the present version, the control synthesis still relies on the H_∞ theory). Moreover, the proposed formulation of the control synthesis design in terms of a mixed H_∞/ℓ_1 control problem has been investigated through a simple example only: we think that this is an interesting issue which deserves further investigations especially for the class of positive systems where a special relation between the H_∞ norm and the ℓ_∞–gain of the system holds.

- The proposed analysis of performance and complexity does not take transmission delays into account. Under this simplifying assumption, controllers of increasingly high complexity offer better and better performance. In practice, however, the deleterious effects of delays increase with the complexity: there is hence a more involved trade off between complexity and performance which is worth studying. To this end, a more

in–depth analysis of the coding issue is needed and the proposed theoretical framework has to be declined in a more practical context.

- Motivated by the spreading of technological applications involving complex and distributed systems, the most important research areas on quantized control have been those related with the control under communication constraints. This appears to be the predominant trend also for the close future: topics like networked and decentralized control of distributed systems are the hub of the most recent developments (see, e.g., [93, 20]) . In this context, there is still much insight to gain on the relation between quantization and control/communication protocols. The integration of these issues with those from the control under assigned quantization, such as the analysis presented in Section 3.1.3 , opens up for interesting design problem to be further investigated.

Appendix A

Appendix

A.1 Appendix to Chapter 2

A.1.1 Appendix to Section 2.1

Proof of Lemma 1. *ı*) Proof of 1 $\Rightarrow$ 2: if $\bar{u}$ was an accumulation point for $\mathcal{U}$, then $\bar{u} \in \mathcal{U}$ because $\mathcal{U}$ is closed, but $\bar{u}$ is not isolated.
Proof of 2 $\Rightarrow$ 3: let $\mathcal{S} \subset \mathbb{R}^m$ be a bounded set and consider $\bar{\mathcal{S}}$, it is sufficient to prove that the set $L := \bar{\mathcal{S}} \cap \mathcal{U}$ is finite. The set L is closed and bounded, therefore it is a compact set. By contradiction, if L was made of infinite points, then an injective sequence $l : \mathbb{N} \to L$ could be defined and, because L is compact, it can be assumed that l is convergent. This is a contradiction because, $\lim_{n\to+\infty} l(n) = \ell$ and the injectivity of l imply that ℓ is an accumulation point for L, and hence for $\mathcal{U}$.
Proof of 3 $\Rightarrow$ 1: this implication is trivial.
ıı) For $u \in \mathcal{U}$, let $d(u) := \inf_{v\in\mathcal{U}\setminus\{u\}} \|u - v\|_2$. Because $\mathcal{U}$ is discrete, then $\forall\, u \in \mathcal{U}$, $d(u) > 0$. By construction, the family $\{\mathcal{B}(u)\}_{u\in\mathcal{U}}$, where $\mathcal{B}(u) := \{x \in \mathbb{R}^m \,|\, \|x - u\|_2 < d(u)/2\}$, is made of disjoint elements. Thus, it can be defined an injective function $Q : \mathcal{U} \to \mathbb{Q}^m$ such that $Q(u) \in \mathcal{B}(u)$: this concludes the proof as it is well known that there exists a bijection between $\mathbb{Q}^m$ and $\mathbb{N}$. ■

With reference to the discussion on the relations between locally finite partitions and the partitions induced by a quantizer, consider the following case:

Example 30 *Let us construct an example of quantizer $q_{\mathcal{U}} : \mathbb{R}^2 \to \mathcal{U} \subset \mathbb{R}^2$ such that $\forall\, u \in \mathcal{U}$, $u \in q_{\mathcal{U}}^{-1}(u)$ but the induced partition is not locally finite. Consider the following partition of $\mathbb{R}^2$: let*

$$\begin{cases} \mathcal{C}_0 := \{(y_1, y_2) \in \mathbb{R}^2 \,|\, y_2 = 0\} \\ \mathcal{C}_1 := \{(y_1, y_2) \in \mathbb{R}^2 \,|\, y_2 > 1\} \\ \mathcal{C}_n := \{(y_1, y_2) \in \mathbb{R}^2 \,|\, \frac{1}{n} < y_2 \le \frac{1}{n-1}\}, \ \mathbb{N} \ni n \ge 2 \\ \mathcal{C}_{-n} := -\mathcal{C}_n, \ \mathbb{N} \ni n \ge 1. \end{cases}$$

The family $\{\mathcal{C}_i\}_{i\in\mathbb{Z}}$ *defines a partition of* $\mathbb{R}^2$ *which is not locally finite because any neighborhood of* 0 *intersects infinite elements of the partition. On the other hand, the map* $q_{\mathcal{U}}$ *defined by*

$$q_{\mathcal{U}}(y) := \begin{cases} (0,0) & \text{if} \quad y \in \mathcal{C}_0 \\ (1,2) & \text{if} \quad y \in \mathcal{C}_1 \\ \left(n, \frac{1}{n-1}\right) & \text{if} \quad y \in \mathcal{C}_n\,,\ \mathbb{N} \ni n \geq 2 \\ -q_{\mathcal{U}}(-y) & \text{if} \quad y \in \mathcal{C}_{-n}\,,\ \mathbb{N} \ni n \geq 1 \end{cases}$$

is a quantizer inducing the partition $\{\mathcal{C}_i\}_{i\in\mathbb{Z}}$ *and such that* $\forall\, u \in \mathcal{U}\,,\ u \in q_{\mathcal{U}}^{-1}(u)\,.$ ♣

Proof of Lemma 3. Without loss of generality, we assume that $y > 0$. For $h \in \mathbb{Z}$, let $u_h := u_0\theta^h \in \mathcal{U}$. Let us consider first the case in which y is equidistant from two control values, say u_h and u_{h+1}. In this case, $y = y_h := \frac{u_h+u_{h+1}}{2} = \frac{u_0(\theta+1)}{2}\theta^h$ and, no matter whether $q_{\mathcal{U}}(y_h) = u_h$ or $q_{\mathcal{U}}(y_h) = u_{h+1}$, it holds that

$$\frac{|q_e(y_h)|}{|y_h|} = \frac{(u_{h+1}-u_h)/2}{(u_h+u_{h+1})/2} = \frac{\theta-1}{\theta+1}\,,$$

irrespective of h. If instead y belongs to the interior part of the interval $q_{\mathcal{U}}^{-1}(u_h)$, that is $y \in]y_{h-1}\,, y_h[$, then $q_{\mathcal{U}}(y) = u_h$ and $\frac{|q_e(y)|}{|y|} = \frac{|u_h-y|}{|y|} = \left|\frac{u_h}{y} - 1\right|$. This function approaches its supremum for y approaching the extremes of the interval $q_{\mathcal{U}}^{-1}(u_h)$ where $\frac{|q_e(y_{h-1})|}{|y_{h-1}|} = \frac{|q_e(y_h)|}{|y_h|} = \frac{\theta-1}{\theta+1}$, irrespective of h. ■

A.1.2 Appendix to Section 2.3

With reference to Example 3, let us prove the following

Lemma 25 *Consider the scalar system*

$$x(t+1) = ax(t) + q_{\mathcal{U}}\big(Kx(t)\big) = (a+K)x(t) + q_e\big(Kx(t)\big)\,, \tag{A.1}$$

where $|a| > 1$, $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *is a nearest neighbor quantizer and* $\mathcal{U}$ *is a logarithmically quantized set in the generalized sense with parameters* (u_0, θ). *The function* $V(x) = x^2$ *is a Lyapunov function for system* (A.1)*, for any choice of the nearest neighbor quantizer* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$*, if and only if* $K \in \mathbb{R}$ *is such that inequality* (2.5) *is satisfied.*

Proof. For $x \neq 0$, it holds that

$$|x^+| < |x| \ \Leftrightarrow\ \big|(a+K)x + q_e(Kx)\big| < |x| \ \Leftrightarrow\ \left|(a+K) + \frac{q_e(Kx)}{x}\right| < 1\,.$$

By Lemma 3, $\forall\, x \neq 0$, it holds that $\frac{|q_e(Kx)|}{|Kx|} \leq \frac{\theta-1}{\theta+1}$: this implies the sufficiency of condition (2.5), indeed $\left|(a+K) + \frac{q_e(Kx)}{x}\right| \leq |a+K| + \frac{|q_e(Kx)|}{|Kx|}|K| \leq |a+K| + \frac{\theta-1}{\theta+1}|K|$.
Let us prove the necessity: let x be such that Kx is a discontinuity point for q_e, namely

$Kx = \frac{u_0(\theta+1)}{2}\theta^h$, then $q_e(Kx) = \pm\frac{u_0(\theta-1)}{2}\theta^h$ (see also the proof of Lemma 3). The sign of $q_e(Kx)$ depends on the particular choice made in the definition of the nearest neighbor quantizer for $y = Kx$ (Kx is equidistant from $u_0\theta^h$ and $u_0\theta^{h+1}$). Thus, for such an x, $\left|(a+K) + \frac{q_e(Kx)}{x}\right| = \left|(a+K) + \frac{q_e(Kx)}{Kx}K\right| = \left|(a+K) \pm \frac{\theta-1}{\theta+1}K\right| \overset{\text{(a)}}{=} |a+K| + \frac{\theta-1}{\theta+1}|K|$, where equality (a) holds if $q_{\mathcal{U}}(Kx)$ has been defined so that $\text{sign}\,(a+K) = \text{sign}\big(\frac{q_e(Kx)}{Kx}K\big)$, the thesis follows. ■

A.2 Appendix to Chapter 3

A.2.1 Appendix to Section 3.1.1

Example 8: logarithmically quantized controls.

Let us prove that the one in equation (3.15) is the expression for $\rho(\Delta)$ when $\mathcal{U}$ is a logarithmically quantized set with parameters (u_0, θ). In fact: with the notation introduced in Remark 5, let $u_k := u_0\theta^k$, $k \in \mathbb{N}$. For $k \geq 1$, it holds that

$$u_k - u_{k-1} = u_0(\theta-1)\theta^{k-1}\,.$$

In particular, $\forall\, k \geq 2$, $u_k - u_{k-1} > u_{k-1} - u_{k-2}$. This means that $\mathbb{K}_c = \{0\,, k_1\,, k_1+1\,, k_1+2\,\cdots\}$. By definition, k_1 is the smallest value of $k \geq 1$ such that $u_k - u_{k-1} > u_0$, that is

$$k_1 = \big\lfloor \log_\theta \tfrac{\theta}{\theta-1}\big\rfloor + 1 = \big\lfloor \log_\theta \tfrac{\theta}{\theta-1} + 1\big\rfloor = \big\lfloor \log_\theta \tfrac{\theta^2}{\theta-1}\big\rfloor\,.$$

Thus, for $\Delta \in \big[\,\frac{2u_0}{\alpha+1}, \frac{2u_0\theta^{k_1}}{\alpha+1}\,\big[$, $\rho(\Delta) = u_0$; instead, for $\Delta \geq \frac{2u_0\theta^{k_1}}{\alpha+1} = \frac{2u_{k_1}}{\alpha+1}$, $\rho(\Delta) = u_k - u_{k-1} = u_0\frac{\theta-1}{\theta}\theta^k$, with k such that $\Delta \in \big[\,\frac{2u_k}{\alpha+1}, \frac{2u_{k+1}}{\alpha+1}\,\big[$, that is $k = \big\lfloor \log_\theta\big(\frac{(\alpha+1)\,\Delta}{2u_0}\big)\big\rfloor$.

A.2.2 Appendix to Section 3.1.3

Proof of Proposition 4. Let $\tilde{x}_m(\tau)$ be the solution of

$$\begin{cases} \dot{\tilde{x}}(\tau) = a\tilde{x}(\tau) + u - \frac{\mathrm{w}}{2} \\ \tilde{x}(0) = x^0 \end{cases}$$

and $\tilde{x}_M(\tau)$ be the solution of the same system with $+\frac{\mathrm{w}}{2}$ in place of $-\frac{\mathrm{w}}{2}$. For any integrable function $\tilde{w} : [\,0\,,\,T\,] \to I(\mathrm{w})$, the solution $\tilde{x}(\tau)$ of system (3.29) is such that $\tilde{x}_m(\tau) \leq \tilde{x}(\tau) \leq \tilde{x}_M(\tau)$, in fact:

$$\tilde{x}(\tau) - \tilde{x}_m(\tau) = \int_0^\tau e^{a(\tau-s)}\Big(\tilde{w}(s) + \frac{\mathrm{w}}{2}\Big)ds \geq 0$$

because the integrand is positive. The other inequality is analogous. It is then sufficient to show that $\forall\,\tau \in [\,0\,,\,T\,]$, $\tilde{x}_m(\tau) \in I(\Delta)$ and $\tilde{x}_M(\tau) \in I(\Delta)$. Since $\tilde{x}_m(\tau)$ is the solution of the differential equation $\dot{\tilde{x}} = a\tilde{x} + u - \frac{\mathrm{w}}{2}$ and the right hand–side is not explicitly depending on τ, then $\tilde{x}_m(\tau)$ is a monotonic function. Hence, the desired property holds because, by

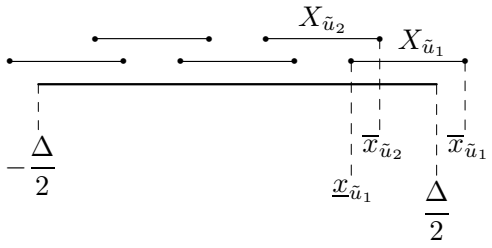

Figure A.1: Construction of $\tilde{\mathcal{U}}$ in the proof of Proposition 7.

assumption, both $\tilde{x}_m(0) \in I(\Delta)$ and $\tilde{x}_m(T) \in I(\Delta)$.
Similar arguments can be applied to $\tilde{x}_M(\tau)$. ■

Proof of Proposition 6. *i*) It holds that

$$R \overset{\text{(a)}}{\geq} \frac{1}{T}\log_2 \ell(\Delta, T) \overset{\text{(b)}}{\geq} \frac{1}{T}\log_2 \alpha + \frac{1}{T}\log_2 \frac{\Delta}{\Delta - \beta(T)\cdot \mathrm{w}} = \frac{a}{\log 2} + \frac{1}{T}\log_2 \frac{\Delta}{\Delta - \beta(T)\cdot \mathrm{w}},$$

where inequality (a) follows by condition (3.33) and inequality (b) by inequality (3.38).

ii) Since $T > 0$ and $a > 0$, then $2 \leq \lceil e^{aT} \rceil \overset{\text{(c)}}{\leq} \ell(\Delta, T) \leq 2^{RT}$ $\big($inequality (c) follows by inequality (3.39)$\big)$, that is $RT \geq 1$.
iii) Apply Proposition 6. *ii* to the condition provided by Proposition 5. *ii*. ■

To prove Proposition 7, we first need the following result:

Lemma 26 $\forall\, x \in \mathbb{R}$ *and* $\forall\, n \in \mathbb{N} \setminus \{0\}$, $\left\lceil \frac{x}{n} \right\rceil = \left\lceil \frac{\lceil x \rceil}{n} \right\rceil$.

Proof. Any $x \in \mathbb{R}$ can be written as $x = \lceil x \rceil - \theta_x$, with $0 \leq \theta_x < 1$. Then,

$$\frac{x}{n} = \frac{\lceil x \rceil}{n} - \frac{\theta_x}{n}.$$

If $\frac{\lceil x \rceil}{n} \in \mathbb{Z}$, then $\left\lceil \frac{x}{n} \right\rceil = \left\lceil \frac{\lceil x \rceil}{n} - \frac{\theta_x}{n} \right\rceil = \frac{\lceil x \rceil}{n} + \left\lceil -\frac{\theta_x}{n} \right\rceil = \frac{\lceil x \rceil}{n} + 0$.
If instead $\frac{\lceil x \rceil}{n} \notin \mathbb{Z}$, since $\frac{x}{n} < \frac{\lceil x \rceil}{n}$, then the thesis follows by the fact that $\lfloor \frac{\lceil x \rceil}{n} \rfloor < \frac{x}{n}$. Let us show that this inequality holds true: according to the Euclidean division, $\lceil x \rceil = qn + r$, with $1 \leq r < n$, hence $\lfloor \frac{\lceil x \rceil}{n} \rfloor = \lfloor q + \frac{r}{n} \rfloor = q = \frac{\lceil x \rceil}{n} - \frac{r}{n} < \frac{\lceil x \rceil}{n} - \frac{\theta_x}{n} = \frac{x}{n}$. ■

Proof of Proposition 7. We show that the control set $\mathcal{U}$ defined by equation (3.42) ensures the w–controlled invariance of $I(\Delta)$ and that a control set $\mathcal{U}_{\min} \subset \epsilon\mathbb{Z}$ of minimal cardinality between those ensuring the invariance of $I(\Delta)$ is such that $\#\mathcal{U}_{\min} = \#\mathcal{U}$. The statement on the invariance of $I(\Delta)$ under a control law of the type in equation (3.43) is a consequence of the definition of X_u $\big($see equation (3.35)$\big)$.

According to the notation introduced in equation (3.36), consider the control set $\tilde{\mathcal{U}}$ defined by the following algorithm (see also Fig. A.1):

- Let $\tilde{\mathcal{U}} := \{\tilde{u}_1\}$, where $\tilde{u}_1$ is such that $\overline{x}_{\tilde{u}_1} = \min\ \overline{x}_u$; $\begin{cases} u \in \epsilon\mathbb{Z} \\ \overline{x}_u \geq \frac{\Delta}{2} \end{cases}$
- While $\underline{x}_{\tilde{u}_k} > -\frac{\Delta}{2}$, let $\tilde{\mathcal{U}} := \tilde{\mathcal{U}} \cup \{\tilde{u}_{k+1}\}$, where $\tilde{u}_{k+1}$ is such that $\overline{x}_{\tilde{u}_{k+1}} = \min\ \overline{x}_u$. $\begin{cases} u \in \epsilon\mathbb{Z} \\ \overline{x}_u \geq \underline{x}_{\tilde{u}_k} \end{cases}$ (A.2)

Let us show that $\tilde{\mathcal{U}} = \mathcal{U}$ and that it ensures the w–controlled invariance of $I(\Delta)$.
With $u = -z\epsilon$ and $z \in \mathbb{Z}$,

$$\overline{x}_u \geq \tfrac{\Delta}{2} \Leftrightarrow \tfrac{1}{\alpha}\big(\tfrac{\Delta}{2} + \beta z\epsilon - \tfrac{\beta \mathrm{w}}{2}\big) \geq \tfrac{\Delta}{2} \Leftrightarrow z \geq \tfrac{1}{2\epsilon}\left(\tfrac{\Delta(\alpha-1)}{\beta} + \mathrm{w}\right) \Leftrightarrow z \geq \left\lceil \tfrac{1}{2\epsilon}\left(\tfrac{\Delta(\alpha-1)}{\beta} + \mathrm{w}\right)\right\rceil ,$$

therefore, $\tilde{u}_1 = u_1$. Moreover, if $\underline{x}_{\tilde{u}_k} > -\frac{\Delta}{2}$, then

$$\tilde{u}_{k+1} = \tilde{u}_k + \left\lfloor \frac{\Delta - \beta \mathrm{w}}{\beta\epsilon} \right\rfloor \cdot \epsilon ,$$

in fact: by equation (3.36), it holds that $\overline{x}_{u+z\epsilon} = \overline{x}_u - \frac{\beta}{\alpha} z\epsilon$. Therefore, $\tilde{u}_{k+1} = \tilde{u}_k + h\epsilon$, where $h \in \mathbb{N}$ is the largest integer such that $\frac{\beta}{\alpha} h\epsilon \leq \mu(X_{\tilde{u}_k}) = \frac{\Delta - \beta \mathrm{w}}{\alpha}$ (μ stands for the Lebesgue measure). That is,

$$h = \left\lfloor \frac{\Delta - \beta \mathrm{w}}{\beta\epsilon} \right\rfloor .$$

Notice that, since $T \in \mathcal{T}(\Delta)$ and $a \geq 0$, then $\Delta \geq \beta(\epsilon + \mathrm{w})$ (see Proposition 5.*ii*). Therefore, $h \geq 1$ and this guarantees that the algorithm (A.2) terminates in a finite number of steps (in fact, $\tilde{u}_{k+1} > \tilde{u}_k$ so that $\underline{x}_{\tilde{u}_{k+1}} < \underline{x}_{\tilde{u}_k}$).
To prove that $\tilde{\mathcal{U}} = \mathcal{U}$, it remains to show that the cardinality of $\tilde{\mathcal{U}}$ is equal to the right hand–side of equation (3.41). By equation (3.36), it holds that $\underline{x}_{u+z\epsilon} = \underline{x}_u - \frac{\beta}{\alpha} z\epsilon$. From this equation and the construction of $\tilde{\mathcal{U}}$ it follows that $\#\tilde{\mathcal{U}} = l$, where $l \in \mathbb{N}$ is the smallest integer such that

$$\underline{x}_{\tilde{u}_1} - (l-1)\tfrac{\beta}{\alpha} h\epsilon \leq -\tfrac{\Delta}{2} ,$$

that is,

$$l = 1 + \left\lceil \frac{\frac{\alpha}{2}(2\underline{x}_{\tilde{u}_1} + \Delta)}{h\beta\epsilon} \right\rceil .$$

To show the equality of l with the right hand–side of equation (3.41), consider $y := \frac{1}{2\epsilon}\left(\frac{\Delta(\alpha-1)}{\beta} + \mathrm{w}\right)$, then

$$\tilde{u}_1 = u_1 = -\lceil y \rceil \epsilon .$$

Thus,

$$
\begin{aligned}
l = \; & 1 + \left\lceil \tfrac{\frac{\alpha}{2}(2\underline{x}_{\tilde{u}_1} + \Delta)}{h\beta\epsilon} \right\rceil \overset{\text{(a)}}{=} 1 + \left\lceil \frac{\frac{\alpha}{2}\left(2\left(-\frac{\Delta}{2\alpha} + \frac{\beta}{\alpha}\lceil y \rceil \epsilon + \frac{\beta \mathrm{w}}{2\alpha}\right) + \Delta\right)}{h\beta\epsilon} \right\rceil = \\
= \; & 1 + \left\lceil \frac{y + \lceil y \rceil}{h} \right\rceil \overset{\text{(b)}}{=} 1 + \left\lceil \frac{\left\lceil y + \lceil y \rceil \right\rceil}{h} \right\rceil = 1 + \left\lceil \frac{2\lceil y \rceil}{h} \right\rceil = \\
= \; & 1 + \left\lceil \frac{2\left\lceil \frac{1}{2\epsilon}\left(\frac{\Delta(\alpha-1)}{\beta} + \mathrm{w}\right) \right\rceil}{\left\lfloor \frac{\Delta - \beta \mathrm{w}}{\beta\epsilon} \right\rfloor} \right\rceil ,
\end{aligned}
$$

where equality (a) is obtained substituting $\underline{x}_{\tilde{u}_1}$ with the expression given in equation (3.36) with $\tilde{u}_1 = -\lceil y \rceil \epsilon$ and equality (b) follows by Lemma 26 .
By construction, $\tilde{\mathcal{U}}$ is such that the invariance condition (3.37) is satisfied (see also Fig. A.1) .

Finally, let us show that $\ell(\Delta, T)$ coincides with the expression in equation (3.41) . Assume that $\mathcal{U}_{\min} \subset \epsilon\mathbb{Z}$ is a control set of minimal cardinality within the family of the control sets $\mathcal{U} \subseteq \epsilon\mathbb{Z}$ ensuring the w–controlled invariance of $I(\Delta)$: the thesis is achieved by showing that $\#\mathcal{U}_{\min} = \#\tilde{\mathcal{U}}$. Suppose that the elements of $\mathcal{U}_{\min} = \{u_1^{(\min)}, \ldots, u_m^{(\min)}\}$ are ordered so that

$$
\overline{x}_{u_1^{(\min)}} > \overline{x}_{u_2^{(\min)}} > \cdots > \overline{x}_{u_m^{(\min)}} .
$$

By the invariance criterion (3.37) , it holds that $\overline{x}_{u_1^{(\min)}} \geq \frac{\Delta}{2}$. Therefore, by definition of $\tilde{u}_1$, $\overline{x}_{u_1^{(\min)}} \geq \overline{x}_{\tilde{u}_1}$. This implies that $\overline{x}_{u_2^{(\min)}} \geq \overline{x}_{\tilde{u}_2}$, in fact: by definition of $\overline{x}_{\tilde{u}_2}$, it is sufficient to show that $\overline{x}_{u_2^{(\min)}} \geq \underline{x}_{\tilde{u}_1}$. This inequality holds because $\overline{x}_{u_2^{(\min)}} \overset{\text{(c)}}{\geq} \underline{x}_{u_1^{(\min)}} \geq \underline{x}_{\tilde{u}_1}$, where inequality (c) holds by the invariance criterion (3.37) . Iterating the same argument, one can show that $\overline{x}_{u_m^{(\min)}} \geq \overline{x}_{\tilde{u}_m}$ and hence $\underline{x}_{\tilde{u}_m} \leq \underline{x}_{u_m^{(\min)}} \leq -\frac{\Delta}{2}$: since by construction $\underline{x}_{\tilde{u}_i} \leq -\frac{\Delta}{2}$ only for $i = \#\tilde{\mathcal{U}}$, then $\#\tilde{\mathcal{U}} = m = \#\mathcal{U}_{\min}$. ■

A.2.3 Appendix to Section 3.1.4

Proof of Proposition 10. *i*) The semi–axes of the ellipse $\mathcal{E}_{P(S), r_{\mathrm{i}}(S)}$ are $r_{\mathrm{i}}(S)/\sqrt{\lambda_{\min}\big(P(S)\big)}$ and $r_{\mathrm{i}}(S)/\sqrt{\lambda_{\max}\big(P(S)\big)}$, then

$$
\mathrm{Area}\left(\mathcal{E}_{P(S), r_{\mathrm{i}}}\right) = \frac{\pi\, r_{\mathrm{i}}^2(S)}{\sqrt{\det P(S)}} . \tag{A.3}
$$

Let us show that $\forall\, S > 0\,,\ \text{Area}\,(\mathcal{E}_{P(S),r_{\rm i}}) \geq \frac{\pi}{\sqrt{2}}$ and that the equality holds if and only if $S = s_1 I\,$. Indeed,

$$\begin{aligned}
\text{Area}\,(\mathcal{E}_{P(S),r_{\rm i}}) &\overset{\text{(a)}}{=} \frac{\pi}{4}\frac{\Big(|s_3| + \sqrt{s_3^2 + \lambda_{\min}(S)(s_1+s_2)}\Big)^2\big(s_1 + \lambda_{\min}(S)\big)}{\lambda_{\min}^2(S)\sqrt{s_1^2 + s_1 s_2 - s_3^2}} \geq \\
&\overset{\text{(b)}}{\geq} \frac{\pi}{4}\frac{\lambda_{\min}(S)(s_1+s_2)\big(s_1 + \lambda_{\min}(S)\big)}{\lambda_{\min}^2(S)\sqrt{s_1^2 + s_1 s_2}} = \\
&= \frac{\pi}{4}\left(1 + \frac{s_1}{\lambda_{\min}(S)}\right)\sqrt{1 + \frac{s_2}{s_1}} \geq \\
&\overset{\text{(c)}}{\geq} \frac{\pi}{4}\left(1 + \frac{s_1}{\min\{s_1\,,\,s_2\}}\right)\sqrt{1 + \frac{s_2}{s_1}} \geq \\
&\overset{\text{(d)}}{\geq} \frac{\pi}{\sqrt{2}}\,,
\end{aligned} \tag{A.4}$$

where equality (a) results from plugging in equation (A.3) the expressions for $\det P(S)\,$, $r_{\rm i}$ and R which can be obtained by equations (3.61), (3.63) and (3.62a), respectively; inequality (b) follows by replacing the s_3's explicitly appearing in the right hand–side of equality (a) with $0\,$; inequality (c) holds because $\frac{1}{\lambda_{\min}(S)}$ is an increasing function of s_3^2 $\big($see equation (3.60)$\big)$ and, for $s_3 = 0\,,\ \lambda_{\min}(S) = \min\{s_1\,,\,s_2\}\,$; finally, inequality (d) follows by the fact that in both cases $s_1 \geq s_2$ or $s_2 \geq s_1\,$, the minimum of the expression in the right hand–side of inequality (c) is achieved for $s_1 = s_2$ and it is equal to $\frac{\pi}{\sqrt{2}}\,$.
Since inequality (b) is strict for $s_3 \neq 0$ and the chain of inequalities (A.4) becomes a chain of equalities when $s_3 = 0$ and $s_1 = s_2\,$, the thesis follows.

ii) $Q_2(1) \subset \mathcal{E}_{P(S),r_{\rm i}(S)}$ if and only if

$$\forall\, x \in Q_2(1)\,, \quad s_1 x_1^2 + 2 s_3 x_1 x_2 + (s_1 + s_2) x_2^2 \leq r_{\rm i}^2(S)\,.$$

For $|x_1| \leq \frac{1}{2}$ and $|x_2| \leq \frac{1}{2}$ it holds that $s_1 x_1^2 + 2 s_3 x_1 x_2 + (s_1 + s_2) x_2^2 \leq \frac{2s_1 + 2|s_3| + s_2}{4}\,$. We have hence to show that

$$\frac{2s_1 + 2|s_3| + s_2}{4} \leq r_{\rm i}^2(S)\,. \tag{A.5}$$

Indeed, inequality (A.5) holds if and only if

$$2s_1 + 2|s_3| + s_2 \leq \frac{\Big(|s_3| + \sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big)^2 \big(s_1 + \lambda_{\min}(S)\big)}{\lambda_{\min}^2(S)} \tag{A.6a}$$

$$\Leftrightarrow\ \lambda_{\min}^2(S)\big(2s_1 + 2|s_3| + s_2\big) \leq s_1\Big(s_3^2 + \lambda_{\min}^2(S) + s_1 s_2 + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big) + \\ + \lambda_{\min}(S)\Big(2s_3^2 + \lambda_{\min}(S)(s_1 + s_2) + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big) \tag{A.6b}$$

$$\Leftrightarrow\ 2\lambda_{\min}^2(S)|s_3| \leq s_1\Big(s_3^2 + s_1 s_2 + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big) + \\ + \lambda_{\min}(S)\Big(2s_3^2 + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big)\,, \tag{A.6c}$$

where to write inequality (A.6a) we used the expression for R given in equation (3.62b) and to write the second addendum in the right hand side of inequality (A.6b) we took advantage of both the expressions for R given in equation (3.62) $\Big($in particular, the fact that $\sqrt{s_3^2 + \lambda_{\min}(S)(s_1 + s_2)} = \sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big)$.
To conclude the proof, let us show that inequality (A.6c) holds true: in fact,

$$\begin{aligned} s_1\Big(s_3^2 + s_1 s_2 + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big) + \lambda_{\min}(S)\Big(2s_3^2 + 2|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2}\Big) &\geq \\ 2\lambda_{\min}(S)|s_3|\sqrt{\lambda_{\min}^2(S) + s_1 s_2} &\geq \\ 2\lambda_{\min}^2(S)|s_3|\,. \end{aligned}$$

The Proposition is proved. ■

A.2.4 Appendix to Section 3.2.1

Proof of Lemma 8. *i*) It is enough to show that $q_{\mathcal{U}}$ maps the extremes of the interval $\mathrm{Pr}_n\big(AQ_n(\Delta)\big)$ to $\mathcal{U}(\Delta)$. It holds that $0 \leq q_{\mathcal{U}}\big(\frac{\Delta}{2}\,\alpha\big) \leq M(\Delta)$, in fact: if $u \in \mathcal{U}$ and $u > M(\Delta)$, then $u > \frac{\Delta}{2}(\alpha + 1)$, so that $\big|\frac{\Delta}{2}\,\alpha - u\big| > \frac{\Delta}{2}$; but $\big|\frac{\Delta}{2}\,\alpha - M(\Delta)\big| \leq \frac{\Delta}{2}$ by definition of $M(\Delta)$ and inequality (3.69b). That is, $\frac{\Delta}{2}\,\alpha$ is closest to $M(\Delta)$ than to any other $u \in \mathcal{U}$ such that $u > M(\Delta)$ and therefore $q_{\mathcal{U}}\big(\frac{\Delta}{2}\,\alpha\big) \leq M(\Delta)$. The other inequality holds because $\frac{\Delta}{2}\,\alpha$ is not further from $M(\Delta)$ than from 0, in fact $\frac{\Delta}{2}\,\alpha \geq \frac{\Delta}{2}$.
Similarly, $q_{\mathcal{U}}\big(-\frac{\Delta}{2}\,\alpha\big) \geq m(\Delta)$.

ii) We have to show that $\forall\, z \in \mathcal{N}_\Delta\,,\ \exists\, u \in \mathcal{U}$ such that $|u - z| \leq \frac{\rho(\Delta)}{2}$: this is obvious for $z \in \big[m(\Delta) - \frac{\rho(\Delta)}{2}\,,\, m(\Delta)\big[\ \cup\ \big]M(\Delta)\,,\, M(\Delta) + \frac{\rho(\Delta)}{2}\big]$; whilst for $z \in \big[m(\Delta)\,,\, M(\Delta)\big]$ the property holds by definition of $\rho(\Delta)$ $\big($see equation (3.8) in Section 3.1.1$\big)$.

iii) It easily follows by the definition of $\mathcal{S}_{M(\Delta)}$ and $\mathcal{S}_{m(\Delta)}$. ■

A.3 Appendix to Chapter 4

A.3.1 Appendix to Section 4.1

Proof of Lemma 10.*ı*. By definition of k, $x_n^+ = (Ax)_n + q_{\mathcal{U}}\big(-(Ax)_n\big)$, where $q_{\mathcal{U}}$ is a nearest neighbor quantizer. With reference to the partition $\mathbb{R} = \mathcal{S}_{m(\Delta)} \cup \mathcal{N}_\Delta \cup \mathcal{S}_{M(\Delta)}$ defined in equation (3.70) of Section 3.2.1, two cases can occur:

I) Suppose that $-(Ax)_n \in \mathcal{N}_\Delta$, then, by Lemma 8.*ıı*,

$$|x_n^+| = \big|(Ax)_n + q_{\mathcal{U}}\big(-(Ax)_n\big)\big| \leq \tfrac{\rho(\Delta)}{2}\,.$$

II) Suppose that $-(Ax)_n \in \mathcal{S}_\Delta$. If $-(Ax)_n \in \mathcal{S}_{m(\Delta)}$, then $|x_n^+| \leq \|x\|_\infty - \varphi(\Delta)$. In fact: in this case, $k(x) = m(\Delta)$ as it follows by Lemma 8.*ıı* that can be applied because, by Lemma 8.*ı*, $k(x) = q_{\mathcal{U}}\big(-(Ax)_n\big) \in \mathcal{U}(\Delta)$. We then conclude as in part III of the proof of Lemma 9.

The case $-(Ax)_n \in \mathcal{S}_{M(\Delta)}$ is similar. ■

Proof of Lemma 11. Let us prove by induction that the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ is non–increasing: $\Delta_1 < \Delta_0$ by assumption (in fact, ϕ takes values in the semi–open interval $[a, \Delta_0[\,$). Let $m \geq 1$, suppose that $\Delta_m \leq \Delta_{m-1}$ and, by contradiction, that $\Delta_{m+1} > \Delta_m$. The latter inequality is equivalent to $\phi(\Delta_m) > \phi(\Delta_{m-1})$ which, by the monotonicity of ϕ, implies $\Delta_m \geq \Delta_{m-1}$. Therefore, because of the inductive hypothesis, $\Delta_m = \Delta_{m-1}$. As a consequence, $\forall\, k \geq m$, $\Delta_k = \Delta_m$ which contradicts the assumption $\Delta_{m+1} > \Delta_m$.

As the sequence is monotonic, $\exists\, \Delta_{\inf} := \lim_{k\to+\infty} \Delta_k$. To prove the equality $\Delta_{\inf} = \max\big\{\Delta < \Delta_0 \,|\, \phi(\Delta) = \Delta\big\}$ it is sufficient to show that

$$\phi(\Delta_{\inf}) = \Delta_{\inf} \tag{A.7}$$

and

$$\forall\, \Delta \in]\Delta_{\inf}\,,\, \Delta_0]\,, \quad \phi(\Delta) < \Delta\,. \tag{A.8}$$

Equation (A.7) holds by the right continuity of ϕ; if instead ϕ is not right continuous but it takes a finite number of values for $\Delta \in [a, \Delta_0]$, then equation (A.7) holds because $\exists\, \hat{m} \in \mathbb{N}$ such that $\Delta_{\hat{m}} = \lim_{k\to+\infty} \Delta_k = \Delta_{\inf}$.

The property in (A.8) follows by showing that if $\Delta_{k+1} < \Delta_k$ (i.e., $\phi(\Delta_k) < \Delta_k$), then $\forall\, \Delta \in]\phi(\Delta_k)\,,\, \Delta_k]$ it holds that $\phi(\Delta) < \Delta$: indeed, if it was $\phi(\Delta) \geq \Delta$, then $\phi(\Delta) \geq \Delta > \phi(\Delta_k)$, therefore $\Delta > \Delta_k$ which is a contradiction. ■

Proof of Lemma 12. Let $\Delta' \in [\rho(\Delta), \Delta[\,$, $m(\Delta') < -\frac{\Delta'}{2}(\alpha-1) \Leftrightarrow \exists\, u \in \mathcal{U}$ such that $u \in I' := \big[-\frac{\Delta'}{2}(\alpha+1)\,,\, -\frac{\Delta'}{2}(\alpha-1)\big[\Leftrightarrow \mathcal{U}(\Delta') \cap I' \neq \emptyset \Leftrightarrow \mathcal{U}(\Delta) \cap I' \neq \emptyset$, because $\Delta' < \Delta$. It holds that $[m(\Delta)\,,\, M(\Delta)] \cap I' = \big[\max\big\{m(\Delta); -\frac{\Delta'}{2}(\alpha+1)\big\}\,,\, -\frac{\Delta'}{2}(\alpha-1)\big[\neq \emptyset$: this is an easy consequence of $m(\Delta) < -\frac{\Delta}{2}(\alpha-1) < -\frac{\Delta'}{2}(\alpha-1)$. By contradiction, if $\mathcal{U}(\Delta) \cap I' = \emptyset$, then $m(\Delta) < -\frac{\Delta'}{2}(\alpha+1)$ and $I' \subset [m(\Delta)\,,\, M(\Delta)]$: I' is a semi–open interval of length Δ', therefore $\rho(\Delta) > \Delta'$ (see the definition of $\rho(\Delta)$ in equation (3.8) of Section 3.1.1), but $\Delta' \in [\rho(\Delta), \Delta[\,$.

Similarly, $M(\Delta') > \frac{\Delta'}{2}(\alpha-1)$. ■

A.3.2 Appendix to Section 4.2.2

Let us represent the dynamic qdb–controller defined in equation (4.10) in the form of equation (2.9). Let

$$\mathcal{W} := \mathcal{Y}^n \times \mathcal{U}^{n-1},$$

the elements $w \in \mathcal{W}$ are denoted both by $w = (\vec{y}, \vec{u})$ and by $w = (w_1, \cdots, w_{2n-1})$. Consider

$$\tilde{k}: \ \mathcal{W} \times (\mathbb{N} \cup \{-1\}) \ \to \ \mathcal{U}$$

defined by

$$\tilde{k}\big((\vec{y}, \vec{u}), t\big) = \begin{cases} 0 & \text{if} \quad t \leq n-2 \\ (k \circ \psi)(\vec{y}, \vec{u}) & \text{if} \quad t \geq n-1, \end{cases}$$

and

$$\gamma: \ \mathcal{W} \times \mathcal{Y} \times \mathbb{N} \ \to \ \mathcal{W}$$

defined by

$$\gamma\big(w, y, t\big) = \big(w_2, \cdots, w_n, y, w_{n+2}, \cdots, w_{2n-1}, \tilde{k}(w, t-1)\big).$$

Finally,

$$\begin{array}{rccc} \bar{k}: & \mathcal{W} \times \mathcal{Y} \times \mathbb{N} & \to & \mathcal{U} \\ & \big(w, y, t\big) & \mapsto & \tilde{k}\big(\gamma(w, y, t), t\big). \end{array}$$

In practice, $w(t) = \big(\vec{y}\,(t-1), \vec{u}\,(t-1)\big)$ and $\tilde{k}\big(w(t), t-1\big) = u(t-1)$.

A.4 Appendix to Chapter 5

A.4.1 Appendix to Section 5.3.2

Let us provide the details of the computations leading to the properties of the quantizers that we have presented in Examples 19, 20 and 21 of Section 5.3.2.

Example 19: logarithmic quantization of $\mathbb{R}$.
Let us consider the function

$$\gamma(y) := \begin{cases} \frac{|q_e(y)|}{|y|} & \text{if } y \neq 0 \\ 1 & \text{if } y = 0. \end{cases}$$

(see Fig. A.2). We show that there exists an unbounded sequence $\{y_h\}_{h \in \mathbb{N}}$ such that $\forall\, h \in \mathbb{N}$, $\gamma(y_h) = \frac{\theta-1}{\theta+1}$ and if $|y| > \frac{u_0(\theta+1)}{2\theta}$, then $\gamma(y) \leq \frac{\theta-1}{\theta+1}$: this proves that q_e is standard with natural external gain $\frac{\theta-1}{\theta+1}$. We also show that if $|y| < \frac{u_0(\theta+1)}{2\theta}$, then $\gamma(y) > \frac{\theta-1}{\theta+1}$: this proves that $\frac{u_0(\theta+1)}{2\theta}$ is the smallest value of ϱ_0 ensuring that the corresponding external gain is the natural one.

Since $q_e(y) = q_{\mathcal{U}}(y) - y$ and $q_{\mathcal{U}}$ is a nearest neighbor quantizer, then the following properties hold for q_e (see also Fig. 5.3 in Section 5.3.2): the set $\mathcal{J}$ of the discontinuity points of q_e is

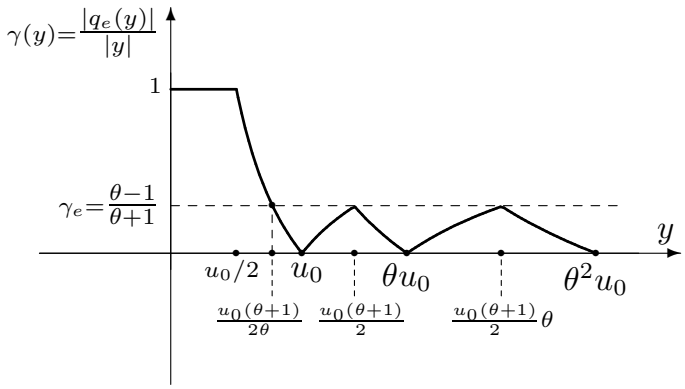

Figure A.2: Logarithmic quantization of $\mathbb{R}$, the behavior of $\gamma(y)$ for $\theta = 1.8$.

made of the middle points between consecutive elements of $\mathcal{U}$; the function $|q_e|$ is continuous and the set of its local maxima is exactly $\mathcal{J}$. It holds that

$$\mathcal{J} = \Big\{ \pm \frac{u_0}{2} \Big\} \cup \Big\{ \pm \frac{u_0(\theta+1)}{2}\theta^h \,|\, h \in \mathbb{N} \Big\}. \tag{A.9}$$

Again by the definition of $q_{\mathcal{U}}$, it is easy to verify that the following properties hold (see also the proof of Lemma 3 in Appendix A.1.1): γ is a continuous function; $\forall\, y \in \mathbb{R}$, $\gamma(y) \leq 1$; the set of the local maxima of γ is $\mathcal{J} \cup [-\frac{u_0}{2}, \frac{u_0}{2}]$. As γ is continuous, most of the information we need about γ can be obtained by evaluating this function in correspondence of the local maxima (the analysis can be restricted to $y \geq 0$ because $\mathcal{U}$ is symmetric with respect to the origin). For $y \in \big[0, \frac{u_0}{2}\big]$, $\gamma(y) = 1$; whereas for $y = y_h := \frac{u_0(\theta+1)}{2}\theta^h$, $|q_e(y)| = \frac{u_0(\theta-1)}{2}\theta^h$ and

$$\gamma(y_h) = \frac{|q_e(y_h)|}{y_h} = \frac{\theta-1}{\theta+1} \tag{A.10}$$

irrespective of h. On the interval $\big[\frac{u_0}{2}, u_0\big]$, $\gamma(y)$ is decreasing, in fact $\gamma(y) = \frac{u_0-y}{y}$. In particular, $\gamma\big(\frac{u_0}{2}\big) = 1$, $\gamma(u_0) = 0$ and $\gamma(y) = \frac{\theta-1}{\theta+1}$ if and only if $y = \frac{u_0(\theta+1)}{2\theta}$. For $y > u_0$, $\gamma(y) \leq \frac{\theta-1}{\theta+1}$ because γ is continuous and, in correspondence of the local maxima larger than u_0, $\gamma(y) = \frac{\theta-1}{\theta+1}$. All these facts yield the desired properties on q_e and ϱ_0.
As for E_0, since $\varrho_0 \in \big]\frac{u_0}{2}, u_0\big[$, it is immediate to check that $\max\limits_{|y|\leq\varrho_0} |q_e(y)| = \frac{u_0}{2}$.
For later use, it is useful to explicitly write the relation $\gamma(\varrho_0) = \gamma_e$ as

$$\frac{|q_e(\varrho_0)|}{\varrho_0} = \gamma_e\,. \tag{A.11}$$

Example 20: componentwise logarithmic quantization of $\mathbb{R}^2$.

With reference to the notation introduced in Example 20, let us first prove that, for $\varrho_0 \geq \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$, q_e has ϱ_0–external gain $\gamma_e(\varrho_0)$. Consider $y \in \mathbb{R}^2$ such that $\|y\|_2 > \varrho_0 \geq \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$. It holds that $|y_1| > \varrho_{01}$ or $|y_2| > \varrho_{02}$. Let us divide the analysis in three cases:

I) If $|y_1| > \varrho_{01}$ and $|y_2| \leq \varrho_{02}$, then $\varrho_0^2 < y_1^2 + y_2^2 \leq y_1^2 + \varrho_{02}^2$ and $|y_1| > \sqrt{\varrho_0^2 - \varrho_{02}^2}$. Therefore,

$$\frac{\|q_e(y)\|_2}{\|y\|_2} = \sqrt{\frac{\|q_e(y)\|_2^2}{\|y\|_2^2}} = \sqrt{\frac{|q_{e1}(y_1)|^2 + |q_{e2}(y_2)|^2}{|y_1|^2 + |y_2|^2}} \leq \sqrt{\frac{\gamma_{e1}^2|y_1|^2 + |y_2|^2}{|y_1|^2 + |y_2|^2}} := \Psi(y_1, y_2)\,,$$

where the inequality holds because, in the computations done for Example 19, we have shown that for $|y_1| > \varrho_{01}$, $|q_{e1}(y_1)| \leq \gamma_{e1}|y_1|$ and for $|y_2| \leq \varrho_{02}$, $|q_{e2}(y_2)| \leq |y_2|$. It is easy to see that, since $\gamma_{e1}^2 < 1$, then

$$\sup_{\begin{cases} |y_1| > \sqrt{\varrho_0^2 - \varrho_{02}^2} \\ |y_2| \leq \varrho_{02} \end{cases}} \Psi(y_1, y_2) = \Psi\Big(\sqrt{\varrho_0^2 - \varrho_{02}^2}\,, \varrho_{02}\Big)\,,$$

therefore

$$\sup_{\begin{cases} |y_1| > \sqrt{\varrho_0^2 - \varrho_{02}^2} \\ |y_2| \leq \varrho_{02} \end{cases}} \frac{\|q_e(y)\|_2}{\|y\|_2} \leq \Psi\Big(\sqrt{\varrho_0^2 - \varrho_{02}^2}\,, \varrho_{02}\Big) = \sqrt{\gamma_{e1}^2 + \Big(\frac{\varrho_{02}}{\varrho_0}\Big)^2(1 - \gamma_{e1}^2)}\,.$$

II) Similarly, if $|y_1| \leq \varrho_{01}$ and $|y_2| > \varrho_{02}$, then $|y_2| > \sqrt{\varrho_0^2 - \varrho_{01}^2}$ and

$$\sup_{\begin{cases} |y_1| \leq \varrho_{01} \\ |y_2| > \sqrt{\varrho_0^2 - \varrho_{01}^2} \end{cases}} \frac{\|q_e(y)\|_2}{\|y\|_2} \leq \sqrt{\gamma_{e2}^2 + \Big(\frac{\varrho_{01}}{\varrho_0}\Big)^2(1 - \gamma_{e2}^2)}\,.$$

III) Finally, if $|y_1| > \varrho_{01}$ and $|y_2| > \varrho_{02}$, then

$$\frac{\|q_e(y)\|_2}{\|y\|_2} \leq \sqrt{\frac{\gamma_{e1}^2|y_1|^2 + \gamma_{e2}^2|y_2|^2}{|y_1|^2 + |y_2|^2}} \leq \max\{\gamma_{e1}\,, \gamma_{e2}\}\,.$$

To some up, as for $i = 1, 2$, $\gamma_{ei} < 1$, then

$$\begin{cases} \sqrt{\gamma_{e1}^2 + \big(\frac{\varrho_{02}}{\varrho_0}\big)^2(1 - \gamma_{e1}^2)} > \gamma_{e1} \\ \sqrt{\gamma_{e2}^2 + \big(\frac{\varrho_{01}}{\varrho_0}\big)^2(1 - \gamma_{e2}^2)} > \gamma_{e2}\,. \end{cases} \tag{A.12}$$

Therefore, with $\gamma_e(\varrho_0)$ defined as in equation (5.34), it holds that $\forall\, y \in \mathbb{R}^2$ such that $\|y\|_2 > \varrho_0$, $\|q_e(y)\|_2 \leq \gamma_e(\varrho_0)\|y\|_2$. This proves that, for $\varrho_0 \geq \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$, q_e has ϱ_0–external gain $\gamma_e(\varrho_0)$.

By inequalities (A.12), it holds in particular that $\gamma_e(\varrho_0) > \max\{\gamma_{e1}\,, \gamma_{e2}\}$. To complete the proof of inequalities (5.36), let us show that $\gamma_e(\varrho_0) < 1$. This is again a consequence of the fact that $\gamma_{ei} < 1$, in fact:

$$\sqrt{\gamma_{e1}^2 + (\varrho_{02}/\varrho_0)^2(1 - \gamma_{e1}^2)} = \sqrt{\frac{\gamma_{e1}^2(\varrho_0^2 - \varrho_{02}^2) + \varrho_{02}^2}{\varrho_0^2}} \leq 1$$

and analogously, $\sqrt{\gamma_{e2}^2 + (\varrho_{01}/\varrho_0)^2(1-\gamma_{e2}^2)} < 1$.

Let us prove that relation (5.37) holds true. Namely, that for $\varrho_0 = \sqrt{\varrho_{01}^2 + \varrho_{02}^2}$, the infimum of the achievable values for $\gamma_e(\varrho_0)$ given in equation (5.34) is $\frac{\sqrt{2}}{2}$. In fact,

$$\inf_{\left\{\begin{array}{l} u_{01}>0 \\ u_{02}>0 \\ \theta_1>1 \\ \theta_2>1 \end{array}\right.} \gamma_e\Big(\sqrt{\varrho_{01}^2+\varrho_{02}^2}\Big) = \inf_{\left\{\begin{array}{l} 0<\gamma_{e1}^2<1 \\ 0<\gamma_{e2}^2<1 \\ \varrho_0>0 \\ \varrho_{01}^2+\varrho_{02}^2=\varrho_0^2 \end{array}\right.} \max\left\{\frac{\sqrt{\gamma_{e1}^2\varrho_{01}^2+\varrho_{02}^2}}{\varrho_0}\,,\,\frac{\sqrt{\varrho_{01}^2+\gamma_{e2}^2\varrho_{02}^2}}{\varrho_0}\right\} := \Theta\,.$$

Also,

$$\Theta^2 = \inf_{\left\{\begin{array}{l} 0<\gamma_{e1}^2<1 \\ 0<\gamma_{e2}^2<1 \\ \varrho_0>0 \\ \varrho_{01}^2+\varrho_{02}^2=\varrho_0^2 \end{array}\right.} \max\left\{\frac{\gamma_{e1}^2\varrho_{01}^2+\varrho_{02}^2}{\varrho_0^2}\,,\,\frac{\varrho_{01}^2+\gamma_{e2}^2\varrho_{02}^2}{\varrho_0^2}\right\}\,:$$

let us show that $\Theta^2 = \frac{1}{2}$. Since, $\max\{x_1\,,x_2\} \geq \frac{x_1+x_2}{2}$, then

$$\max\left\{\frac{\gamma_{e1}^2\varrho_{01}^2+\varrho_{02}^2}{\varrho_0^2}\,,\,\frac{\varrho_{01}^2+\gamma_{e2}^2\varrho_{02}^2}{\varrho_0^2}\right\} \geq \frac{\varrho_{01}^2(1+\gamma_{e1}^2)+\varrho_{02}^2(1+\gamma_{e2}^2)}{2\varrho_0^2} := \Upsilon^2\,.$$

Therefore,

$$\Theta^2 \geq \inf_{\left\{\begin{array}{l} 0<\gamma_{e1}^2<1 \\ 0<\gamma_{e2}^2<1 \\ \varrho_0>0 \\ \varrho_{01}^2+\varrho_{02}^2=\varrho_0^2 \end{array}\right.} \quad \Upsilon^2 \geq \inf_{\left\{\begin{array}{l} \gamma_{e1}^2>0 \\ \gamma_{e2}^2>0 \\ \varrho_0>0 \\ \varrho_{01}^2+\varrho_{02}^2=\varrho_0^2 \end{array}\right.} \quad \Upsilon^2 \geq \inf_{\left\{\begin{array}{l} \varrho_0>0 \\ \varrho_{01}^2+\varrho_{02}^2=\varrho_0^2 \end{array}\right.} \frac{\varrho_{01}^2+\varrho_{02}^2}{2\varrho_0^2} = \frac{1}{2}\,.$$

It is then easy to see that it is possible to have γ_e arbitrarily close to $\frac{\sqrt{2}}{2}$ by choosing $u_{01} = u_{02}$ and $\theta_1 = \theta_2$ with θ_1 sufficiently close to 1.

As far as the absolute quantization error is concerned, let us show that for $\|y\|_2 \leq \varrho_0$ it holds that $\|q_e(y)\|_2 \leq E_0(\varrho_0)$, where $E_0(\varrho_0)$ is defined by equations (5.34) and (5.35). Indeed, we have only to prove that equation (5.35) holds true, namely that

$$\mathtt{E}_{0i}(\varrho_0) := \max_{|y_i|\leq\varrho_0} |q_{ei}(y_i)| = \max\left\{\frac{u_{0i}}{2}\,,\,\gamma_{ei}\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)}\,,\,|u_{0i}\theta_i^{n_i(\varrho_0)+1}-\varrho_0|\right\}\,. \quad \text{(A.13)}$$

In fact, if $\|y\|_2 \leq \varrho_0$, then both $|y_1| \leq \varrho_0$ and $|y_2| \leq \varrho_0$, therefore

$$\|q_e(y)\|_2 = \sqrt{|q_{e1}(y_1)|^2 + |q_{e2}(y_2)|^2} \leq \sqrt{\mathtt{E}_{01}(\varrho_0)^2 + \mathtt{E}_{02}(\varrho_0)^2}\,.$$

Because of symmetry, we can restrict to $0 \leq y_i \leq \varrho_0$. We have seen in the treatment of Example 19 that $|q_{ei}|$ is a continuous function and, for $y_i \geq 0$, the set of its local maxima is

$$\left\{ \frac{u_{0i}}{2} \right\} \cup \left\{ \frac{u_{0i}(\theta_i + 1)}{2} \theta_i^h \,|\, h \in \mathbb{N} \right\}$$

(see equation (A.9)). Since $\varrho_0 > \frac{u_{0i}}{2}$ (in fact, $\varrho_0 > \varrho_{0i} > \frac{u_{0i}}{2}$) and, for $h_1 > h_2$, it holds that $\left| q_{ei}\left(\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{h_1}\right) \right| > \left| q_{ei}\left(\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{h_2}\right) \right|$, then

$$\max_{y_i \in [0\,,\,\varrho_0]} |q_{ei}(y_i)| = \max \left\{ \left| q_{ei}\left(\frac{u_{0i}}{2}\right) \right|, \left| q_{ei}\left(\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)}\right) \right|, |q_{ei}(\varrho_0)| \right\}, \tag{A.14}$$

where $n_i(\varrho_0)$ is the largest value of $h \in \mathbb{Z}$ such that $\frac{u_{0i}(\theta_i+1)}{2}\theta_i^h \leq \varrho_0$ (notice that for $h = -1$, $\frac{u_{0i}(\theta_i+1)}{2}\theta_i^h = \varrho_{0i} < \varrho_0$, hence $n_i(\varrho_0) \geq -1$: even if ϱ_{0i} is not a local maximum for $|q_{ei}|$, including also the case $h = -1$ allows us to give a unique formula for $\mathtt{E}_{0i}(\varrho_0)$). Let us show that the three quantities appearing in the maximization (A.14) are exactly those appearing in the maximization (A.13): $\left| q_{ei}\left(\frac{u_{0i}}{2}\right) \right| = \frac{u_{0i}}{2}$; it is straightforward to see that $n_i(\varrho_0) = \left\lfloor \log_{\theta_i} \frac{2\varrho_0}{u_{0i}(\theta_i+1)} \right\rfloor$, thus $\left| q_{ei}\left(\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)}\right) \right| = \gamma_{ei}\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)}$ (this follows by equation (A.10) if $n_i(\varrho_0) \geq 0$, and by equation (A.11) if $n_i(\varrho_0) = -1$). To conclude the proof we have only to show that $\left| q_{ei}(\varrho_0) \right| = |u_{0i}\theta_i^{n_i(\varrho_0)+1} - \varrho_0|$. Indeed, $\varrho_0 \in \left[\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)}, \frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)+1} \right]$ and on this interval $q_{ei}(y) = u_{0i}\theta_i^{n_i(\varrho_0)+1} - y$: the desired property follows.

Finally, let us show that both $\gamma_e(\varrho_0)$ and $E_0(\varrho_0)$ can be made arbitrarily small by properly choosing the parameters u_{0i} and θ_i defining the quantized set $\mathcal{U}$. Let us start by noting that, following the above computations, it is immediate to see that, for $i = 1, 2$, $\mathtt{E}_{0i}(\varrho_0) \leq \max \left\{ \frac{u_{0i}}{2}, \gamma_{ei}\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)+1} \right\}$. Because by definition $n_i(\varrho_0) \leq \log_{\theta_i} \frac{2\varrho_0}{u_{0i}(\theta_i+1)}$, then $\gamma_{ei}\frac{u_{0i}(\theta_i+1)}{2}\theta_i^{n_i(\varrho_0)+1} \leq \gamma_{ei}\varrho_0\theta_i$, therefore

$$\mathtt{E}_{0i}(\varrho_0) \leq \max \left\{ \frac{u_{0i}}{2}, \gamma_{ei}\varrho_0\theta_i \right\}. \tag{A.15}$$

Let us fix $\varrho_0 > 0$. One can first make γ_{ei} arbitrarily close to 0 by choosing θ_i sufficiently close to 1. Then, also $\frac{\varrho_{0i}}{\varrho_0}$ can be made arbitrarily small by picking u_{0i} sufficiently close to 0. With these choices, according to equation (5.34), $\gamma_e(\varrho_0)$ can be made arbitrarily small. The same holds for $E_0(\varrho_0)$ thanks to equation (A.15).

Example 21: the joint radial logarithmic quantization of $\mathbb{R}^2$.

• ***The input space partition:*** let us first give a quick description of the input space partition induced by $q_{\mathcal{U}}$ (see Fig. A.3). This is the Voronoi partition generated by $\mathcal{U}$ (see e.g., [94]). With the notation introduced in Example 21 and in Definition 9 in Section 2.1, $\forall\, k = 0, \ldots, N-1$ and $\forall\, h \in \mathbb{N}$, let $u_{kh} := \ell_k \cap c_h$ and V_{kh} be the Voronoi region containing u_{kh}, that is

$$V_{kh} = \overline{\{y \in \mathbb{R}^2 \,|\, q_{\mathcal{U}}(y) = u_{kh}\}}.$$

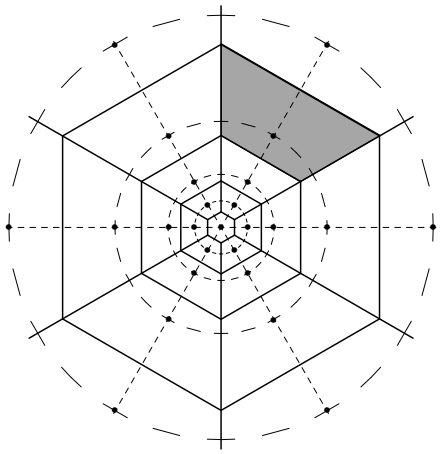

Figure A.3: Radial logarithmic quantization of $\mathbb{R}^2$ with $N = 6$ and $\theta = 2$. Full lines define the input space partition induced by $q_{\mathcal{U}}$.

Since $\mathcal{U}$ is invariant under rotations of an angle $\frac{2\pi}{N}$ around the origin, we can limit ourselves to consider the case $k = 0$. It is easy to see that

$$V_{0h} = \left\{ y \in \mathbb{R}^2 \mid -y_1 \tan(\pi/N) \leq y_2 \leq y_1 \tan(\pi/N) \quad \text{and} \quad \zeta_h \leq y_1 \leq \zeta_{h+1} \right\},$$

where

$$\zeta_h := \begin{cases} \frac{u_0}{2} & \text{if} \quad h = 0 \\ \frac{u_0(\theta+1)}{2} \theta^{h-1} & \text{if} \quad h \geq 1 \,. \end{cases}$$

Therefore,

$$V_{0h} = \left\{ y = (y_1, y_1 \tan\varphi) \in \mathbb{R}^2 \mid -\pi/N \leq \varphi \leq \pi/N \quad \text{and} \quad \zeta_h \leq y_1 \leq \zeta_{h+1} \right\} \tag{A.16}$$

(see Fig. A.4). By V_0 we denote the Voronoi region containing $u = 0$, that is

$$V_0 := \overline{\{y \in \mathbb{R}^2 \mid q_{\mathcal{U}}(y) = 0\}} = \overline{\mathbb{R}^2 \setminus \left(\bigcup_{k=0,\ldots,N-1\,;\, h \in \mathbb{N}} V_{kh} \right)}.$$

V_0 is a regular polyhedron centered in the origin, having N edges and the radius of the circle inscribed into it is $\frac{u_0}{2}$. Hence, denoted by r_{c} the radius of the circle circumscribed to V_0, we have

$$r_{\mathrm{c}} = \frac{u_0}{2\cos(\pi/N)} \,. \tag{A.17}$$

• ***Properties of q_e:*** let us prove that q_e is standard with the natural gain γ_e given in equation (5.39) and that $\frac{u_0(\theta+1)}{2\theta\cos(\pi/N)}$ is the smallest value of ϱ_0 ensuring that the corresponding external gain is the natural one. To this end, let us start by analyzing the behavior of

$\frac{\|q_e(y)\|_2}{\|y\|_2}$ within V_{kh}. By the symmetry of $\mathcal{U}$, there is no loss of generality in assuming $k = 0$. We hence consider $u_{0h} = (u_0\theta^h, 0) \in V_{0h}$, $h \in \mathbb{N}$. Thanks to equation (A.16),

$$\max_{y\in V_{0h}} \frac{\|q_e(y)\|_2}{\|y\|_2} = \max_{\varphi\in[-\frac{\pi}{N}\,,\,\frac{\pi}{N}]} \max_{y_1\in[\zeta_h\,,\,\zeta_{h+1}]} \frac{\|q_e(y_1, y_1\tan\varphi)\|_2}{\|(y_1, y_1\tan\varphi)\|_2}\,.$$

For $\varphi \in \big[-\frac{\pi}{N}\,,\,\frac{\pi}{N}\big]$, let $\Gamma^2(\varphi) := \max\limits_{y_1\in[\zeta_h\,,\,\zeta_{h+1}]} \frac{\|q_e(y_1,y_1\tan\varphi)\|_2^2}{\|(y_1,y_1\tan\varphi)\|_2^2}$, then

$$\begin{aligned}\Gamma^2(\varphi) = &\max_{y_1\in[\zeta_h\,,\,\zeta_{h+1}]} \frac{(u_0\theta^h-y_1)^2+y_1^2\tan^2\varphi}{y_1^2(1+\tan^2\varphi)} = \\ = &\max_{y_1\in[\zeta_h\,,\,\zeta_{h+1}]} \frac{y_1^2-2y_1u_0\theta^h\cos^2\varphi+u_0^2\theta^{2h}\cos^2\varphi}{y_1^2} = \\ = &\max_{\lambda\in[1/\zeta_{h+1}\,,\,1/\zeta_h]} \Psi_\varphi(\lambda)\,,\end{aligned}$$

where $\lambda := \frac{1}{y_1}$ and $\Psi_\varphi(\lambda) := (u_0^2\theta^{2h}\cos^2\varphi)\lambda^2 - 2(u_0\theta^h\cos^2\varphi)\lambda + 1$. For $h \geq 1$, it holds that $\Psi_\varphi(1/\zeta_{h+1}) = \Psi_\varphi(1/\zeta_h) = 1 - \frac{4\theta\cos^2\varphi}{(\theta+1)^2}$. Moreover, $u_0^2\theta^{2h}\cos^2\varphi > 0$ because $|\varphi| \leq \frac{\pi}{N} \leq \frac{\pi}{3}$, therefore

$$\Gamma^2(\varphi) = 1 - \frac{4\theta\cos^2\varphi}{(\theta+1)^2}\,.$$

Thus, $\forall\, h \geq 1$, we have

$$\max_{y\in V_{0h}} \frac{\|q_e(y)\|_2}{\|y\|_2} = \sqrt{\max_{\varphi\in[-\frac{\pi}{N}\,,\,\frac{\pi}{N}]} \Gamma^2(\varphi)} = \sqrt{1 - \frac{4\theta\cos^2(\pi/N)}{(\theta+1)^2}} := \gamma_e\,.$$

As for $h = 0$, let

$$\widetilde{V}_{00} := \big\{y \in \mathbb{R}^2 \,|\, -y_1\tan(\pi/N) \leq y_2 \leq y_1\tan(\pi/N) \quad \text{and} \quad \tilde{\zeta}_0 \leq y_1 \leq \zeta_1\big\} \subset V_{00}\,,$$

where $\tilde{\zeta}_0 := \frac{\zeta_1}{\theta} = \frac{u_0(\theta+1)}{2\theta}$ (see Fig. A.4). By repeating the above computations, we have

$$\max_{y\in\widetilde{V}_{00}} \frac{\|q_e(y)\|_2}{\|y\|_2} = \gamma_e\,. \tag{A.18}$$

To sum up, with

$$\mathcal{V} := \bigcup_{k=0,\ldots,N-1\,;\,h\geq 1} V_{kh} \bigcup \bigcup_{k=0,\ldots,N-1} \widetilde{V}_{k0}\,,$$

it holds that

$$\forall\, y \in \mathcal{V}\,, \quad \frac{\|q_e(y)\|_2}{\|y\|_2} \leq \gamma_e\,.$$

Thus, with ϱ_0 equal to the radius of the circle circumscribed to $\mathbb{R}^2 \setminus \mathcal{V}$, we have proved that q_e has ϱ_0–external gain γ_e. As $\mathbb{R}^2 \setminus \mathcal{V}$ is a regular polyhedron centered in the origin, having N edges and the radius of the circle inscribed into it is $\tilde{\zeta}_0$, then

$$\varrho_0 = \big\|\big(\tilde{\zeta}_0, \tilde{\zeta}_0\tan(\pi/N)\big)\big\|_2 = \frac{u_0(\theta+1)}{2\theta\cos(\pi/N)}\,.$$

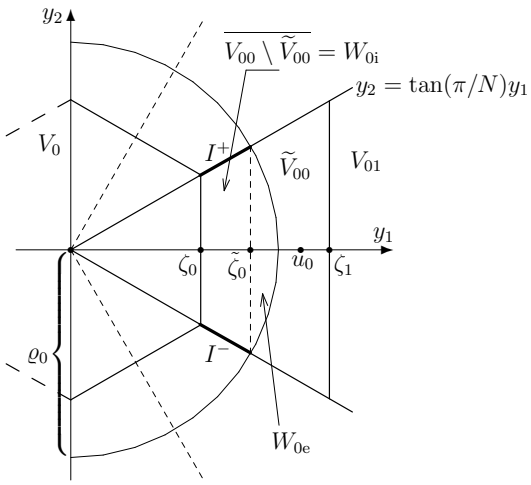

Figure A.4: Visualization of the symbols intervening in the computations to prove the facts concerned with Example 21 (the parameters are the same as in Fig. A.3) .

Indeed, we have also shown that $\{y^{(h)}\}_{h\in\mathbb{N}\setminus\{0\}}$, where $y^{(h)} := \big(\zeta_h, \zeta_h \tan(\pi/N)\big)$, is an unbounded sequence such that $\frac{\|q_e(y^{(h)})\|_2}{\|y^{(h)}\|_2} = \gamma_e$. This proves that q_e is standard with natural gain γ_e. Moreover, resuming the computations leading to equation (A.18), one sees that for $\lambda > \frac{1}{\tilde{\zeta}_0}$ and $\varphi = \frac{\pi}{N}$, $\Psi_\varphi(\lambda) > \Psi_\varphi(1/\tilde{\zeta}_0) = \gamma_e^2$: this means that if $y = \big(y_1, y_1 \tan(\pi/N)\big)$ with $\zeta_0 \le y_1 < \tilde{\zeta}_0$, then $\frac{\|q_e(y)\|_2}{\|y\|_2} > \gamma_e$. Therefore, if $\varrho_0 < \big\|\big(\tilde{\zeta}_0, \tilde{\zeta}_0 \tan(\pi/N)\big)\big\|_2$, then the ϱ_0–external gain of q_e is strictly larger than γ_e.

Now, let us show that $\max_{y\in\mathcal{B}_{\varrho_0}} \|q_e(y)\|_2 = E_0$, where E_0 is defined in equation (5.39). Consider the Voronoi region V_0 containing $u = 0$: by construction, $V_0 \subset \mathcal{B}_{\varrho_0}$. It holds that

$$\max_{y\in V_0} \|q_e(y)\|_2 = \frac{u_0}{2\cos(\pi/N)}\,, \tag{A.19}$$

in fact: for $y \in V_0$, $\|q_e(y)\|_2 = \|y\|_2$, therefore $\max_{y\in V_0} \|q_e(y)\|_2$ is equal to r_{c}, the radius of the circle circumscribed to V_0. Thus, equation (A.19) follows by equation (A.17).

Let us evaluate $\max_{y\in\mathcal{B}_{\varrho_0}\setminus V_0} \|q_e(y)\|_2$. Let $W_0 := \mathcal{B}_{\varrho_0} \cap (\cup_{h\in\mathbb{N}} V_{0h})$. As before, because of symmetry,

$$\max_{y\in\mathcal{B}_{\varrho_0}\setminus V_0} \|q_e(y)\|_2 = \max_{y\in W_0} \|q_e(y)\|_2\,.$$

Also, $W_0 = W_{0\mathrm{i}} \cup W_{0\mathrm{e}}$, where $W_{0\mathrm{i}} = \overline{V_{00} \setminus \widetilde{V}_{00}}$ and $W_{0\mathrm{e}} = W_0 \cap \mathcal{V}$ (see Fig. A.4). For $y \in W_{0\mathrm{e}}$ it holds that

$$\|q_e(y)\|_2 \overset{\text{(a)}}{\le} \gamma_e \|y\|_2 \overset{\text{(b)}}{\le} \gamma_e \varrho_0 = \frac{u_0}{2\theta}\sqrt{(\theta-1)^2 + (1+\theta)^2 \tan^2(\pi/N)}\,, \tag{A.20}$$

where inequality (a) holds because $W_{0\mathrm{e}} \subset \mathcal{V}$ and inequality (b) because $W_{0\mathrm{e}} \subset \mathcal{B}_{\varrho_0}$.
For $y \in W_{0\mathrm{i}}$, the maximum of $\|q_e\|_2$ is achieved on the boundary of $W_{0\mathrm{i}}$. In fact: by the definition of $q_{\mathcal{U}}$, the function $\|q_e\|_2$ is continuous and its local maxima lie on the boundary of the Voronoi regions; by construction, $W_{0\mathrm{i}}$ does not contain any of these boundary points into its interior. The only portion of the boundary of $W_{0\mathrm{i}}$ which is not contained into the already analyzed set $V_0 \cup W_{0\mathrm{e}}$ is made of the segments

$$I^+ := \left\{ y = \big(y_1, y_1 \tan(\pi/N)\big) \in \mathbb{R}^2 \,|\, \tfrac{u_0}{2} < y_1 < \tilde{\zeta}_0 \right\}$$
$$I^- := \left\{ y = \big(y_1, -y_1 \tan(\pi/N)\big) \in \mathbb{R}^2 \,|\, \tfrac{u_0}{2} < y_1 < \tilde{\zeta}_0 \right\}.$$

(see Fig. A.4). For $y \in I^{\pm}$, it holds that

$$\|q_e(y)\|_2 = \big\|\big(u_0 - y_1, -y_1 \tan(\pi/N)\big)\big\|_2 = \sqrt{\big(1 + \tan^2(\pi/N)\big) y_1^2 - 2u_0 y_1 + u_0^2} := f(y_1),$$

then

$$\begin{aligned} \sup_{y_1 \in \left]\frac{u_0}{2}, \tilde{\zeta}_0\right[} f(y_1) = & \quad \max\left\{ f(u_0/2)\,,\, f(\tilde{\zeta}_0) \right\} = \\ = & \quad \max\left\{ \tfrac{u_0}{2\cos(\pi/N)}\,,\, \tfrac{u_0}{2\theta}\sqrt{(\theta-1)^2 + (1+\theta)^2 \tan^2(\pi/N)} \right\}. \end{aligned}$$

This result, compared with those in equations (A.19) and (A.20), yields the expression for E_0 given in equation (5.39).
Finally, running again through the computations done to determine E_0, we immediately realize that if $\frac{u_0}{2\theta}\sqrt{(\theta-1)^2 + (1+\theta)^2 \tan^2(\pi/N)} > \frac{u_0}{2\cos(\pi/N)}$ and $\tilde{\varrho}_0 < \varrho_0$, then

$$\max_{y \in \mathcal{B}_{\tilde{\varrho}_0}} \|q_e(y)\|_2 < E_0 .$$

Proof of Lemma 4 of Section 2.1. In the discussion of Example 21, we have shown that the quantization error associated to a radial logarithmic quantization of $\mathbb{R}^2$ is such that, for $h \geq 1$,

$$\max_{y \in V_{0h}} \frac{\|q_e(y)\|_2}{\|y\|_2} = \sqrt{1 - \frac{4\theta \cos^2(\pi/N)}{(\theta+1)^2}}\,. \tag{A.21}$$

For a generalized radial logarithmic quantization of $\mathbb{R}^2$, to prove Lemma 4, it is sufficient to replace $h \geq 1$ with $h \in \mathbb{Z}$ both in the definition of V_{0h} and in the computations leading to equation (A.21). ■

A.5 Appendix to Chapter 6

A.5.1 Appendix to Section 6.1.1

Proof of Proposition 15. Let $\vec{u} = \{u(t)\}_{t \in \mathbb{N}} \in \ell_\infty(\mathbb{R}^m)$, then $\vec{y}_{\mathrm{f}}(\vec{u}) = (\vec{g} * \vec{u})$, that is $y_{\mathrm{f}}(t) = \sum_{\tau=0}^{t-1} g(t-\tau)u(\tau)$. According to equation (5.11), we have to show that

$$\sup_{\vec{u} \in \ell_\infty(\mathbb{R}^m) \setminus \{\vec{0}\}} \frac{\|\vec{y}_{\mathrm{f}}(\vec{u})\|_\infty}{\|\vec{u}\|_\infty} = \max_{i=1,\ldots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)|\,.$$

Let $\vec{u} \in \ell_\infty(\mathbb{R}^m) \setminus \{\vec{0}\}$, denote by $\vec{y}_{\rm f}$, for short, the corresponding output $\vec{y}_{\rm f}(\vec{u})$, then

$$\begin{aligned}
\|\vec{y}_{\rm f}\|_\infty = & \sup_{t\in\mathbb{N}} \|y_{\rm f}(t)\|_\infty = \\
= & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} |y_{{\rm f}\,i}(t)| = \\
= & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} \big|\textstyle\sum_{\tau=0}^{t-1} e_i' g(t-\tau)u(\tau)\big| \leq \\
\leq & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{t-1} |e_i' g(t-\tau)u(\tau)| = \\
= & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{t-1} \big|\sum_{j=1}^{m} g_{i,j}(t-\tau)u_j(\tau)\big| \leq \\
\leq & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{t-1} \sum_{j=1}^{m} |g_{i,j}(t-\tau)||u_j(\tau)| \leq \\
\leq & \sup_{t\in\mathbb{N}} \max_{i=1,\dots,q} \Big[\big(\max_{\tau=0,\dots,t-1} \max_{j=1,\dots,m} |u_j(\tau)|\big) \textstyle\sum_{\tau=0}^{t-1} \sum_{j=1}^{m} |g_{i,j}(\tau+1)|\Big] = \\
= & \sup_{t\in\mathbb{N}} \Big[\big(\max_{\tau=0,\dots,t-1} \|u(\tau)\|_\infty\big) \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{t} \sum_{j=1}^{m} |g_{i,j}(\tau)|\Big] \leq \\
\leq & \sup_{t\in\mathbb{N}} \Big[\big(\max_{\tau=0,\dots,t-1} \|u(\tau)\|_\infty\big) \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)|\Big] = \\
= & \|\vec{u}\|_\infty \cdot \max_{i=1,\dots,q} \textstyle\sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| \,.
\end{aligned}$$

Thus,

$$\sup_{\vec{u}\in\ell_\infty(\mathbb{R}^m)\setminus\{\vec{0}\}} \frac{\|\vec{y}_{\rm f}(\vec{u})\|_\infty}{\|\vec{u}\|_\infty} \leq \max_{i=1,\dots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| \,.$$

Vice versa, first notice that

$$\max_{i=1,\dots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| \leq \sum_{\tau=0}^{+\infty} \max_{i=1,\dots,q} \sum_{j=1}^{m} |g_{i,j}(\tau)| = \|\vec{g}\|_1 < +\infty$$

because the system is assumed to be BIBO–stable (see property 3 of Lemma 14). For any fixed $T > 0$, let

$$\hat{\imath} := \operatorname*{argmax}_{i=1,\dots,q} \sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{i,j}(\tau)| \tag{A.22}$$

and consider $\vec{u}^{\,(T)} \in \ell_\infty(\mathbb{R}^m)$ defined by

$$u_j^{(T)}(t) := \begin{cases} \operatorname{sign}\big(g_{\hat{\imath},j}(T-t)\big) & \text{if } 0 \leq t \leq T-1 \\ 0 & \text{if } t \geq T \,. \end{cases}$$

Denote by $\vec{y}_{\rm f}$, for short, the corresponding output signal $\vec{y}_{\rm f}(\vec{u}^{\,(T)})$, then $\forall\, i = 1, \dots, q$,

$$\begin{aligned}
|y_{{\rm f}\,i}(T)| = & \big|\textstyle\sum_{\tau=0}^{T-1} \sum_{j=1}^{m} g_{i,j}(T-\tau) u_j^{(T)}(\tau)\big| \leq \\
\leq & \textstyle\sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{i,j}(\tau)| \leq \\
\leq & \textstyle\sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{\hat{\imath},j}(\tau)| = \\
= & |y_{{\rm f}\,\hat{\imath}}(T)| \,,
\end{aligned}$$

therefore $\|\vec{y}_{\mathrm{f}}\|_\infty \geq \|y_{\mathrm{f}}(T)\|_\infty = |y_{\mathrm{f}\,\hat{\imath}}(T)| = \max\limits_{i=1,\dots,q} \sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{i,j}(\tau)|$ $\big($see equation (A.22)$\big)$. Since $\|\vec{u}^{\,(T)}\|_\infty = 1$, then

$$\forall\, T > 0\,, \qquad \sup_{\vec{u}\in\ell_\infty(\mathbb{R}^m)\setminus\{\vec{0}\}} \frac{\|\vec{y}_{\mathrm{f}}(\vec{u})\|_\infty}{\|\vec{u}\|_\infty} \geq \frac{\|\vec{y}_{\mathrm{f}}(\vec{u}^{\,(T)})\|_\infty}{\|\vec{u}^{\,(T)}\|_\infty} \geq \max_{i=1,\dots,q} \sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{i,j}(\tau)|\,.$$

In other words,

$$\forall\, T > 0\,, \qquad \sup_{\vec{u}\in\ell_\infty(\mathbb{R}^m)\setminus\{\vec{0}\}} \frac{\|\vec{y}_{\mathrm{f}}(\vec{u})\|_\infty}{\|\vec{u}\|_\infty} \geq \|g^{(T)}\|_\infty\,,$$

where $g^{(T)} := \sum_{\tau=0}^{T} \sum_{j=1}^{m} |g_{i,j}(\tau)| \in \mathbb{R}^q$. Since $\max\limits_{i=1,\dots,q} \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)| < +\infty$, then there exists $\lim_{T\to+\infty} g^{(T)} = \sum_{\tau=0}^{+\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)|$ and, by the continuity of the vector norm $\|\cdot\|_\infty$ in $\mathbb{R}^q$, $\lim_{T\to+\infty} \|g^{(T)}\|_\infty = \max\limits_{i=1,\dots,q} \sum_{\tau=0}^{\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)|$. Therefore,

$$\sup_{\vec{u}\in\ell_\infty(\mathbb{R}^m)\setminus\{\vec{0}\}} \frac{\|\vec{y}_{\mathrm{f}}(\vec{u})\|_\infty}{\|\vec{u}\|_\infty} \geq \max_{i=1,\dots,q} \sum_{\tau=0}^{\infty} \sum_{j=1}^{m} |g_{i,j}(\tau)|\ : \tag{A.23}$$

this concludes the proof. ■

Lemma 27 *Let ℓ be a Banach space, denote its norm by $\|\cdot\|_*$. Let $\Psi : \ell \to \ell$ be a linear operator and assume that $\|\Psi\|_* < 1$, where $\|\Psi\|_* := \sup\limits_{v\in\ell\setminus\{0\}} \frac{\|\Psi(v)\|_*}{\|v\|_*}$. Then it is defined the operator $(I+\Psi)^{-1}$, it holds that $(I+\Psi)^{-1} = \sum_{i=0}^{+\infty}(-\Psi)^i$ and $\|(I+\Psi)^{-1}\|_* \leq \frac{1}{1-\|\Psi\|_*}$.*

Proof. Since $\|\Psi\|_* < 1$, then $\sum_{i=0}^{+\infty} \|(-\Psi)^i\|_* \leq \sum_{i=0}^{+\infty} \|\Psi\|_*^i = \frac{1}{1-\|\Psi\|_*} < +\infty$. Because ℓ is a Banach space, this ensures the existence of a bounded linear operator represented by $\sum_{i=0}^{+\infty}(-\Psi)^i$ and that $\big\|\sum_{i=0}^{+\infty}(-\Psi)^i\big\|_* \leq \frac{1}{1-\|\Psi\|_*}$. To see that $\sum_{i=0}^{+\infty}(-\Psi)^i = (I+\Psi)^{-1}$, it is sufficient to notice that $\forall\, k \in \mathbb{N}$, one has $(I+\Psi)\sum_{i=0}^{k}(-\Psi)^i = I + (-1)^k\Psi^{k+1}$: letting $k \to +\infty$, since $\lim_{k\to+\infty} \|\Psi^k\|_* = 0$, then $(I+\Psi)\sum_{i=0}^{+\infty}(-\Psi)^i = I$ as desired. ■

Lemma 28 *Let $p(z) := z^n - \sum_{i=1}^{n} f_i z^{i-1}$. If $\sum_{i=1}^{n} |f_i| < 1$ and $z_* \in \mathbb{C}$ is such that $p(z_*) = 0$, then $|z_*| < 1$.*

Proof. If $p(z_*) = 0$ and $|z_*| \geq 1$, then

$$|z_*^n| = |\textstyle\sum_{i=1}^{n} f_i z_*^{i-1}| \leq \sum_{i=1}^{n} |f_i||z_*^{i-1}| \leq |z_*^{n-1}| \sum_{i=1}^{n} |f_i|\,.$$

Therefore, $\sum_{i=1}^{n} |f_i| \geq |z_*^n|/|z_*^{n-1}| = |z_*| \geq 1$, the thesis follows. ■

A.5.2 Appendix to Section 6.1.2

Lemma 29 *Let $L \in \mathbb{C}^{p\times q}$, $M \in \mathbb{C}^{q\times p}$ and $\lambda \in \mathbb{C}\setminus\{0\}$. Assume that both $\lambda I_p + LM \in \mathbb{C}^{p\times p}$ and $\lambda I_q + ML \in \mathbb{C}^{q\times q}$ are invertible, then*

$$M(\lambda I_p + LM)^{-1} = (\lambda I_q + ML)^{-1} M\,.$$

Proof. Indeed,

$$\begin{aligned} M = & \ (\lambda I_q + ML)^{-1}(\lambda I_q + ML)M = \\ = & \ (\lambda I_q + ML)^{-1}(\lambda M + MLM) = \\ = & \ (\lambda I_q + ML)^{-1}M(\lambda I_p + LM)\,. \end{aligned}$$

Multiplying the first and the last terms of the equality by $(\lambda I_p + LM)^{-1}$ on the right, the thesis follows. ■

A.5.3 Appendix to Section 6.2.1

Proof of Lemma 20. By induction on $\hat{m}$: for $\hat{m} = 1$ it is sufficient to take $\epsilon_1 \in]0, \epsilon'[$. Let us show that if the desired property holds for some $\hat{m} \geq 1$, then it holds for $\hat{m}+1$. Let $\epsilon' > 0$: by definition of $\mathscr{E}^+(\Delta_{\hat{m}})$, $\exists\, \tilde{\epsilon} > 0$ such that

$$\forall\, \epsilon \in]0, \tilde{\epsilon}[\,, \quad \mathscr{E}(\Delta_{\hat{m}} + \epsilon) < \mathscr{E}^+(\Delta_{\hat{m}}) + \frac{\epsilon'}{4\|\mathcal{G}^{(I)}\|_\infty}\,.$$

For such a $\tilde{\epsilon}$, by the inductive assumption, $\exists\, \{\epsilon_k\}_{k=1,\dots,\hat{m}}$ (with $\epsilon_k > 0 \ \ \forall\, k = 1 \dots, \hat{m}$) such that $\Delta^+_{\hat{m}} = \Delta_{\hat{m}} + \epsilon$ with $\epsilon \in]0, \tilde{\epsilon}[$: we claim that these ϵ_k's and $\epsilon_{\hat{m}+1} = \frac{\epsilon'}{2}$ ensure that $\Delta_{\hat{m}+1} < \Delta^+_{\hat{m}+1} < \Delta_{\hat{m}+1} + \epsilon'$. In fact:

$$\begin{aligned} \Delta^+_{\hat{m}+1} = & \ 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(\Delta^+_{\hat{m}}) + \epsilon_{\hat{m}+1} = \\ = & \ 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(\Delta_{\hat{m}} + \epsilon) + \tfrac{\epsilon'}{2} < \\ < & \ 2\|\mathcal{G}^{(I)}\|_\infty \big(\mathscr{E}^+(\Delta_{\hat{m}}) + \tfrac{\epsilon'}{4\|\mathcal{G}^{(I)}\|_\infty}\big) + \tfrac{\epsilon'}{2} = \\ = & \ 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}^+(\Delta_{\hat{m}}) + \epsilon' = \\ = & \ \Delta_{\hat{m}+1} + \epsilon'\,. \end{aligned}$$

On the other hand,

$$\begin{aligned} \Delta^+_{\hat{m}+1} = & \ 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}(\Delta_{\hat{m}} + \epsilon) + \tfrac{\epsilon'}{2} \geq \\ \geq & \ 2\|\mathcal{G}^{(I)}\|_\infty \mathscr{E}^+(\Delta_{\hat{m}}) + \tfrac{\epsilon'}{2} = \\ = & \ \Delta_{\hat{m}+1} + \tfrac{\epsilon'}{2} > \\ > & \ \Delta_{\hat{m}+1}\,. \end{aligned}$$

■

A.5.4 Appendix to Section 6.2.2

Let us prove Theorem 13. As in the proof of Theorem 6 in Section 4.1, the core argument is provided by a preliminary lemma yielding a partial practical stability result for the closed loop dynamics.

Lemma 30 (Main tool) *Consider system* (6.27), *assume* **A0** *and that* $\alpha \geq 1$. *Let* $K \in \mathbb{R}^{1\times n}$ *be such that* $F := A + BK$ *satisfies* $f := \sum_{i=1}^{n} |f_i| < 1$. *Consider the control law* $u(x) = q_{\mathcal{U}}(Kx)$, *where* $q_{\mathcal{U}} : \mathbb{R} \to \mathcal{U}$ *is a nearest neighbor quantizer. Let* $\Delta > 0$ *be such that* $\rho(\Delta) < +\infty$ *and inequalities (6.29a–b) hold.*

ı) If $x \in Q_n(\Delta)$ *and* $u(x) \in \mathcal{U}(\Delta)$, *then* $x^+ = Ax + Bk(x)$ *is such that*

$$|x_n^+| \leq \max\left\{\frac{\rho(\Delta)}{2}\,,\, \|x\|_\infty - \varphi(\Delta)\right\}, \tag{A.24}$$

where, as in equation (3.80),

$$\varphi(\Delta) := \min\left\{M(\Delta) - \tfrac{\Delta}{2}(\alpha-1)\,,\, -\tfrac{\Delta}{2}(\alpha-1) - m(\Delta)\right\}.$$

ıı) If $\rho(\Delta) < (1-f)\Delta$ *and* $\forall\, x \in Q_n(\Delta)$, $u(x) \in \mathcal{U}(\Delta)$, *then,* $\forall\, \Delta' \in [\, f\Delta + \rho(\Delta)\,, \Delta\,]$, $u(x)$ *is* $\big(Q_n(\Delta), Q_n(\Delta')\big)$*–stabilizing and* $u\big(Q_n(\Delta')\big) \subseteq \mathcal{U}(\Delta')$. *In particular,* $\forall\, \Delta' \in [\, f\Delta + \rho(\Delta)\,, \Delta\,]$, $Q_n(\Delta')$ *is positively invariant.*

Proof. The proof is similar to that of Lemma 9 in Section 3.2.2 (see also the proof of Lemma 10 in Section 4.1 and in Appendix A.3.1).

ı) Consider the quantization error $q_e = q_{\mathcal{U}} - I : \mathbb{R} \to \mathbb{R}$, then

$$x_n^+ = \sum_{i=1}^{n} a_i x_i + q_{\mathcal{U}}(Kx) = \sum_{i=1}^{n} f_i x_i + q_e(Kx).$$

With reference to the partition $\mathbb{R} = \mathcal{S}_{m(\Delta)} \cup \mathcal{N}_\Delta \cup \mathcal{S}_{M(\Delta)}$ defined in equation (3.70) of Section 3.2.1, two cases can occur:

I) Suppose that $Kx \in \mathcal{N}_\Delta$, then

$$\begin{aligned} |x_n^+| = &\ \big|\textstyle\sum_{i=1}^{n} f_i x_i + q_e(Kx)\big| \leq \\ \leq &\ \big|\textstyle\sum_{i=1}^{n} f_i x_i\big| + \big|q_e(Kx)\big| \leq \\ \leq &\ f\tfrac{\Delta}{2} + \tfrac{\rho(\Delta)}{2}\,, \end{aligned}$$

where the last inequality follows by the fact that $x \in Q_n(\Delta)$ and by Lemma 8.*ıı* in Section 3.2.1.

II) Assume instead that $Kx \in \mathcal{S}_\Delta$. If $Kx \in \mathcal{S}_{M(\Delta)}$, then $q_{\mathcal{U}}(Kx) = M(\Delta)$ thanks to Lemma 8.*ıı* that can be applied because we assumed that $u(x) = q_{\mathcal{U}}(Kx) \in \mathcal{U}(\Delta)$. Thus, $x_n^+ = \sum_{i=1}^{n} a_i x_i + M(\Delta)$. Since $-\sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n}(-a_i)x_i \leq \alpha\|x\|_\infty$, then $x_n^+ \geq -\alpha\|x\|_\infty + M(\Delta)$. Thanks to inequalities (6.29a–b) and by definition of $\varphi(\Delta)$ we can write $M(\Delta) = \frac{\Delta}{2}(\alpha-1) + \varphi(\Delta) + \theta$, with $\theta \geq 0$. Hence,

$$x_n^+ \geq -\alpha\|x\|_\infty + M(\Delta) = -\alpha\|x\|_\infty + \tfrac{\Delta}{2}(\alpha-1) + \varphi(\Delta) + \theta \geq -\|x\|_\infty + \varphi(\Delta)$$

where the last inequality holds because $(\alpha-1)\big(\frac{\Delta}{2} - \|x\|_\infty\big) + \theta \geq 0$. On the other hand, $x_n^+ = \sum_{i=1}^{n} f_i x_i + q_e(Kx) < \sum_{i=1}^{n} f_i x_i - \frac{\rho(\Delta)}{2}$ by Lemma 8.*ıı*. Thus $x_n^+ < \sum_{i=1}^{n} f_i x_i - \frac{\rho(\Delta)}{2} \leq \big|\sum_{i=1}^{n} f_i x_i - \frac{\rho(\Delta)}{2}\big| \leq f\frac{\Delta}{2} + \frac{\rho(\Delta)}{2}$. Namely,

$$-\|x\|_\infty + \varphi(\Delta) \leq x_n^+ < f\tfrac{\Delta}{2} + \tfrac{\rho(\Delta)}{2}$$

which implies $|x_n^+| \le \max\{f\frac{\Delta}{2} + \frac{\rho(\Delta)}{2}\,,\ \|x\|_\infty - \varphi(\Delta)\}$: in fact, if $-\|x\|_\infty + \varphi(\Delta) \ge 0$, then $|x_n^+| \le f\frac{\Delta}{2} + \frac{\rho(\Delta)}{2}$; if instead $-\|x\|_\infty + \varphi(\Delta) < 0$, then $|x_n^+| \le \max\Big\{f\frac{\Delta}{2} + \frac{\rho(\Delta)}{2}\,,\ \big|-\|x\|_\infty + \varphi(\Delta)\big|\Big\} = \max\{f\frac{\Delta}{2} + \frac{\rho(\Delta)}{2}\,,\ \|x\|_\infty - \varphi(\Delta)\}$.
The case $Kx \in \mathcal{S}_{m(\Delta)}$ is similar.

ii) It is a consequence of part *i* and of the controller form of the system. In fact: $\forall\, x \in Q_n(\Delta)$, inequality (A.24) holds; $\varphi(\Delta) > 0$ by inequalities (6.29a-b), therefore inequality (A.24) implies that $\forall\, \Delta' \in [\,f\Delta + \rho(\Delta)\,,\Delta\,]$, $Q_n(\Delta')$ is positively invariant and, by Lemma 6 in Section 3.1.1, $u\big(Q_n(\Delta')\big) \subseteq \mathcal{U}(\Delta')$. Moreover, because $x^+ = (x_2, \dots, x_n, x_n^+)$, inequality (A.24) also implies that

$$\forall\, x(0) \in Q_n(\Delta)\,, \quad \|x(n)\|_\infty \le \max\{f\tfrac{\Delta}{2} + \tfrac{\rho(\Delta)}{2}\,,\ \|x(0)\|_\infty - \varphi(\Delta)\}\ :$$

since $\varphi(\Delta) > 0$, the iteration of this argument yields the $\big(Q_n(\Delta), Q_n(\Delta')\big)$–stability. ■

Proof of Theorem 13. Let

$$\phi(\Delta) := f\Delta + \rho(\Delta)\ :$$

with $\bar\Delta := \frac{2u_0}{\alpha+1}$ (see equation (3.9) in Section 3.1.1), $\phi : [\bar\Delta, +\infty[\to \mathbb{R}$. Actually, $\phi : [\bar\Delta, +\infty[\to [\bar\Delta, +\infty[$ and it is a non–decreasing and right continuous function. In fact: ϕ is non–decreasing and right continuous because so is ρ and $f \ge 0$; moreover, $\forall\, \Delta \ge \bar\Delta$, $\phi(\Delta) \ge \rho(\Delta) \ge \bar\Delta$ (see equation (3.10)). Notice that, for $\Delta > \bar\Delta$,

$$\phi(\Delta) < \Delta \iff \rho(\Delta) < (1-f)\Delta\,. \tag{A.25}$$

Because inequality (6.29c) holds, then the restriction of ϕ to the interval $[\bar\Delta, \Delta_0]$ satisfies the hypotheses of Lemma 11 in Section 4.1. Consider the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ defined by $\Delta_{k+1} := \phi(\Delta_k)$: by Lemma 11, it is a non–increasing sequence and, as in the proofs of Theorem 12.*ii* and Proposition 19.*ii*, equation (4.4) and relation (A.25) imply that $\lim_{k\to+\infty} \Delta_k = \Delta_{\inf}(f)$, with $\Delta_{\inf}(f)$ as defined in equation (6.30).
Let $\Delta_\star \in\,]\Delta_{\inf}(f)\,,\, \Delta_0]$: by Lemma 11, $\exists\, \bar m \in \mathbb{N}$ such that $\Delta_{\inf} \le \Delta_{\bar m} < \Delta_\star$. We can hence define $\hat m := \min\{\bar m \in \mathbb{N} \,|\, \Delta_{\bar m} < \Delta_\star\}$. We claim that $\forall\, k = 0, \dots, \hat m - 1$, the hypotheses of Lemma 30.*ii* are satisfied with $\Delta = \Delta_k$.

The claim implies the thesis, in fact: by Lemma 30.*ii*, $\forall\, k = 0, \dots, \hat m - 1$ and $\forall\, \Delta' \in [\Delta_{k+1}, \Delta_k]$, $Q_n(\Delta')$ is positively invariant, in particular so are both $Q_n(\Delta_0)$ and $Q_n(\Delta_\star)$; moreover $\forall\, k = 0, \dots, \hat m - 1$, $u(x)$ is $\big(Q_n(\Delta_k), Q_n(\Delta_{k+1})\big)$–stabilizing, in particular $u(x)$ is $\big(Q_n(\Delta_0), Q_n(\Delta_{\hat m})\big)$–stabilizing and, because $\Delta_{\hat m} < \Delta_\star$, $\big(Q_n(\Delta_0), Q_n(\Delta_\star)\big)$–stabilizing.

Let us prove the claim: inequality (6.29c) is satisfied by Δ_k if and only if $\phi(\Delta_k) < \Delta_k$, namely $\Delta_{k+1} < \Delta_k$. Let us show that for $k \le \hat m - 1$, it holds that $\Delta_{k+1} < \Delta_k$. This is a consequence of the following two facts: first, by definition of $\hat m$, $\Delta_k \ge \Delta_\star > \Delta_{\hat m}$; secondly, if $\Delta_h = \Delta_{h+1}$ for some $h \in \mathbb{N}$, then $\forall\, k \ge h$, $\Delta_k = \Delta_h$. Inequalities (6.29a-b) are satisfied by Δ_k, $\forall\, k \in \mathbb{N}$: by induction, for $k = 0$ they hold by assumption while for $k \ge 1$, since $\Delta_k = f\Delta_{k-1} + \rho(\Delta_{k-1}) \ge \rho(\Delta_{k-1})$, it follows by Lemma 12 of Section 4.1. Finally,

$u\big(Q_n(\Delta_k)\big) \subseteq \mathcal{U}(\Delta_k)$: this holds for $k = 0$ as $\mathcal{U} = \mathcal{U}(\Delta_0)$, while for $k = 1, \dots, \hat{m} - 1$ it follows by recursive application of Lemma 30.*iii*. ■

Let us provide the details of the computations allowing us to prove equations (6.31) and (6.32) in Example 28 of Section 6.2.2.

Example 28: Analysis of ψ for logarithmic quantization of $\mathbb{R}$.

Proof of *i*: by definition,

$$\begin{aligned} \mathscr{E}(r_2^2) = & \sup_{x \in \mathcal{E}_{P,r_2^2}} \|\psi(x)\|_\infty = \\ = & \sup_{x \in \mathcal{E}_{P,r_2^2}} \big|q_e(Kx)\big| = \\ = & \sup_{y \in \mathbb{R} \,:\, |y| \le \mu_1} \big|q_e(y)\big| , \end{aligned}$$

where $\mu_1 := \max\limits_{x \in \mathcal{E}_{P,r_2^2}} |Kx|$.

From this relation, it immediately follows the continuity of $\mathscr{E}(r_2^2)$, in fact: μ_1 is continuous with respect to r_2^2 and $|q_e|$ is continuous with respect to y (the latter holds because q_e is the quantization error associated to a nearest neighbor quantizer, see also Fig. 5.3).

To complete the proof, it is sufficient to show that

$$\max_{x \in \mathcal{E}_{P,r_2^2}} |Kx| = r_2 \sqrt{K P^{-1} K'} \tag{A.26}$$

and that $\sup\limits_{y \in \mathbb{R} \,:\, |y| \le \mu_1} \big|q_e(y)\big|$ is equal to the right–hand side of equation (6.31).

The one in equation (A.26) is a standard constrained maximization problem that can be solved through the application of the Lagrange multipliers method. As $\mathcal{E}_{P,r_2^2}$ is symmetric with respect to the origin, then, with $L(x, \lambda) := Kx + \lambda(x'Px - r_2^2)$, a value of $x \in \mathcal{E}_{P,r_2^2}$ maximizing $|Kx|$ is so that the following system is satisfied for some $\lambda \in \mathbb{R}$:

$$\begin{cases} \dfrac{\partial L}{\partial x} = K + 2\lambda x' P = 0 & \text{(A.27a)} \\[2ex] \dfrac{\partial L}{\partial \lambda} = x' P x - r_2^2 = 0 \,. & \text{(A.27b)} \end{cases}$$

Equation (A.27a) implies that $2\lambda x' P x = -Kx$, then, using equation (A.27b), $2\lambda r_2^2 = -Kx$. By equation (A.27a), we can also come to $2\lambda x = -P^{-1}K'$ which combined with the former expression yields $(Kx)x = r_2^2 P^{-1} K'$. Multiplying both sides of the latter expression by K, we find $(Kx)^2 = r_2^2 K P^{-1} K'$ so that equation (A.26) follows.

Finally, the equivalence between $\sup\limits_{y \in \mathbb{R} \,:\, |y| \le \mu_1} \big|q_e(y)\big|$ and the right–hand side of equation (6.31) is just a generalization of equation (5.35) including the case $\varrho_0 \le \frac{u_0(\theta+1)}{2\theta}$. The proof can be easily obtained following that of equation (5.35) given in Appendix A.4.1.

Proof of *ii*: it follows by the same arguments used to prove case *i*. We have only to notice that, by definition of $\|K\|_\infty$, $\mu_2 := \max\limits_{x \in Q_n(\Delta)} |Kx| = \|K\|_\infty \frac{\Delta}{2}$.

A.6 Appendix to Chapter 7

A.6.1 Basic properties of the Laplace Transformation

We briefly recall the properties of the Laplace transformation that turns out to be useful for our presentation. In the sequel, the real functions to which the Laplace transformation is applied are tacitly assumed to be bounded on every finite interval. By s we denote a complex–valued variable whereas β denote a real–valued variable. The real part of s is denoted by $\mathcal{R}e(s)$.

Let $g(x)$ be a real function, the *Laplace transform* of g is defined by

$$G(s) := \int_0^{+\infty} g(x)e^{-sx}dx\,.$$

The Laplace transform of g is also conveniently denoted by

$$\mathcal{L}[g(x)](s) := G(s)\,.$$

Theorem 18 *Let $g(x)$ be a real function with Laplace transform $G(s)$ converging on the right half–plane $\{s \in \mathbb{C} \,|\, \mathcal{R}e(s) > \sigma^*\}$. Then $G(s)$ is an analytic function in the interior of the half–plane of convergence and its n–th derivative is $G^{(n)}(s) = (-1)^n\mathcal{L}[x^n g(x)](s)$.*

Proof. See [35] (Theorem 6.1, page 26). ■

Theorem 19 (Initial value theorem) *Let $g(x)$ be such that $\exists \lim_{x\to 0^+} g(x)$ and $\exists\, G(\beta)$ for some $\beta > 0$. Then*

$$\exists \lim_{\beta\to+\infty} \beta G(\beta) = \lim_{x\to 0^+} g(x)\,.$$

Proof. See [35] (Theorem 33.4, page 226). ■

Theorem 20 (Final value theorem) *Let $g(x)$ be such that $\exists \lim_{x\to+\infty} g(x) = g_\infty \in \mathbb{R} \cup \{\pm\infty\}$ and G is convergent on the right half–plane $\{s \in \mathbb{C} \,|\, \mathcal{R}e(s) > 0\}$. Then*

$$\exists \lim_{\beta\to 0^+} \beta G(\beta) = g_\infty\,.$$

Proof. When $g_\infty \in \mathbb{R}$, see [35] (Theorem 34.3, page 233). Let us prove the case $g_\infty = +\infty$: $\forall\, M > 0$, $\exists\, \hat{x}(M)$ such that $\forall\, x \geq \hat{x}(M)$, $g(x) \geq M$. Hence, $\beta G(\beta) = \beta \int_0^{+\infty} e^{-\beta x} g(x)dx = \beta \int_0^{\hat{x}(M)} e^{-\beta x} g(x)dx + \beta \int_{\hat{x}(M)}^{+\infty} e^{-\beta x} g(x)dx \;\geq\; \beta K(\beta) + M\beta \int_{\hat{x}(M)}^{+\infty} e^{-\beta x}dx$, where $K(\beta) := \int_0^{\hat{x}(M)} e^{-\beta x} g(x)dx$ is a bounded function of β. Therefore, $\beta G(\beta) \geq \beta K(\beta) + Me^{-\beta\hat{x}(M)}$ and $\liminf_{\beta\to 0+} \beta G(\beta) \geq M$. Since M can be chosen arbitrarily large, the thesis follows. ■

Theorem 21 *Let f and g be real functions with Laplace transforms F and G converging on the right half–plane $\{s \in \mathbb{C} \,|\, \mathcal{R}e(s) > 0\}$. Assume that $\forall\, x \geq 0$, $g(x) \geq 0$, $\lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} g(x) = +\infty$ and that*

$$\lim_{x\to+\infty} \frac{f(x)}{g(x)} = 1\,,$$

then

i) $\lim_{\beta\to 0^+} \dfrac{F(\beta)}{G(\beta)} = 1\,;$

ii) $\lim_{\beta\to 0^+} \dfrac{F'(\beta)}{G'(\beta)} = 1\,.$

Proof. *i)* Let M_0 be such that $\forall\, x \geq M_0\,,\ g(x) > 0\,.$ For $x \geq M_0$ we can write $f(x) = g(x) + h(x)g(x)$ with $\lim_{x\to+\infty} h(x) = 0\,.$ For any given $\epsilon > 0\,,$ let $M > M_0$ be such that $\forall\, x \geq M\,,\ |h(x)| \leq \epsilon\,.$
$F(s) = G(s) + \int_0^M \big(f(x) - g(x)\big)e^{-sx}dx + \int_M^{+\infty} h(x)g(x)e^{-sx}dx\,.$
Because for $\mathcal{R}e(s) > 0\,,\ |e^{-sx}| < 1\,,$ then $\big|\int_0^M \big(f(x)-g(x)\big)e^{-sx}dx\big| \leq \int_0^M |f(x)-g(x)|dx = K(\epsilon)$ where $K(\epsilon)$ is a finite constant (thanks to the boundedness assumption on f and $g\,$) not depending on $s\,.$ Also, $\big|\int_M^{+\infty} h(x)g(x)e^{-sx}dx\big| \leq \epsilon\int_M^{+\infty} g(x)e^{-\mathcal{R}e(s)x}dx \leq \epsilon\, G\big(\mathcal{R}e(s)\big)\,,$ where in the last inequality we use the fact that $g(x) \geq 0$ on $[0\,,\,M]\,.$
We hence have that $\left|\frac{F(s)}{G(s)} - 1\right| \leq \epsilon\frac{G\left(\mathcal{R}e(s)\right)}{|G(s)|} + \frac{K(\epsilon)}{|G(s)|}\,.$ Since for $\beta > 0\,,\ G(\beta) > 0\,,$ then $\left|\frac{F(\beta)}{G(\beta)} - 1\right| \leq \epsilon + \frac{K(\epsilon)}{G(\beta)}\,.$ The thesis follows because, as a consequence of the "Final value theorem", it holds that $\lim_{\beta\to 0^+} G(\beta) = +\infty\,.$

ii) $\lim_{x\to+\infty} \dfrac{xf(x)}{xg(x)} = 1\,,$ the thesis then follows by part *i* and Theorem 18. ■

Corollary 11 *If* $\exists\,\lambda > 0$ *such that* $\lim_{x\to+\infty} \frac{g(x)}{Ax^\lambda} = 1\,,$ *then* $G(s)$ *exists for* $\mathcal{R}e(s) > 0$ *and*

$$\lim_{\beta\to 0^+} \frac{G(\beta)}{A\,\Gamma(\lambda+1)\beta^{-(\lambda+1)}} = 1\,.$$

Proof. It is a particular case of Theorem 21.*i* (see also [35], Theorem 34.1 – page 231). ■

Lemma 31 *Let* $\theta > 1$ *and*

$$f(x) := \begin{cases} 0 & \text{if} \quad x \in [0\,,\,1] \\ \log_\theta x & \text{if} \quad x > 1\,, \end{cases} \tag{A.28}$$

then

i) $\lim_{\beta\to 0^+} \dfrac{F(\beta)}{-\frac{\log\beta}{\beta}} = \dfrac{1}{\log\theta}\,;$

ii) $F'(\beta) = -\Big(\frac{F(\beta)}{\beta} + \frac{e^{-\beta}}{\beta^2\log\theta}\Big)\,,$ *in particular* $\lim_{\beta\to 0^+} \dfrac{F'(\beta)}{\frac{\log\beta}{\beta^2}} = \dfrac{1}{\log\theta}\,.$

Proof. *i)* For $\beta > 0\,,$

$$\begin{aligned} F(\beta) = &\ \tfrac{1}{\log\theta}\int_1^{+\infty} e^{-\beta x}\log x\,dx = \tfrac{1}{\beta\log\theta}\int_1^{+\infty} \tfrac{e^{-\beta x}}{x}\,dx = \\ = &\ \tfrac{1}{\beta\log\theta}\int_\beta^{+\infty} \tfrac{e^{-y}}{y}\,dy = \tfrac{1}{\beta\log\theta}\big[\int_\beta^1 \tfrac{1}{y}\,dy + \int_\beta^1 \tfrac{e^{-y}-1}{y}\,dy + \int_1^{+\infty} \tfrac{e^{-y}}{y}\,dy\big] = \\ = &\ \tfrac{1}{\beta\log\theta}[-\log\beta + \psi(\beta) + c_2]\,, \end{aligned}$$

where $c_2 > 0$ and $\lim_{\beta\to 0^+} \psi(\beta) = c_1 < 0$: the thesis follows.

ıı) By Theorem 18, for $\beta > 0$, $F'(\beta) = -\frac{1}{\log\theta}\int_1^{+\infty} e^{-\beta x} x \log x\, dx = -\frac{1}{\beta\log\theta}\int_1^{+\infty} e^{-\beta x}(\log x + 1)\, dx = -\Big(\frac{F(\beta)}{\beta} + \frac{1}{\beta\log\theta}\int_1^{+\infty} e^{-\beta x}\, dx\Big) = -\Big(\frac{F(\beta)}{\beta} + \frac{e^{-\beta}}{\beta^2\log\theta}\Big)$. ■

A.6.2 Appendix to Section 7.3.1

The following Lemma provides a relation between the growth of $g(x)$ and that of d_k with k.

Lemma 32 *Given a quantization $\mathcal{I}$, there exists $\lambda > 0$ such that*

$$\lim_{k\to+\infty} \frac{d_k}{k^\lambda} = +\infty$$

if and only if there exists $\gamma > 0$ such that the corresponding g–function is such that

$$\lim_{x\to+\infty} \frac{g(x)}{x^\gamma} = 0\,.$$

Proof. If $\lambda > 0$ is such that $\lim_{k\to+\infty} \frac{d_k}{k^\lambda} = +\infty$, then $\forall\, M > 0$, $\exists\, k_M$ such that $\forall\, k \geq k_M$,

$$d_k \geq M k^\lambda\,. \tag{A.29}$$

Consider a quantization $\mathcal{I}_M$ such that $\mathcal{D}_{\mathcal{I}_M} = \big\{ \pm M k^\lambda \,|\, k \in \mathbb{N}\big\}$ and let g_M be the corresponding g–function. By inequality (A.29) and the definition of g–function, it holds that $\forall\, x \geq d^2_{k_M}$, $g(x) \leq g_M(x)$. Therefore, $\forall\, M > 0$, $\limsup_{x\to+\infty} \frac{g(x)}{x^{1/2\lambda}} \leq \lim_{x\to+\infty} \frac{g_M(x)}{x^{1/2\lambda}} = \frac{1}{M^2}$ (see equation (7.19)). Namely, the thesis holds with $\gamma = \frac{1}{2\lambda}$.

Vice versa, if $\gamma > 0$ is such that $\lim_{x\to+\infty} \frac{g(x)}{x^\gamma} = 0$, then $\forall\, \epsilon > 0$, $\exists\, x_\epsilon$ such that $\forall\, x \geq x_\epsilon$, $g(x) \leq \epsilon x^\gamma$. In particular, since $k = g(d_k^2)$ and g is a non decreasing function, then $\forall\, k > g(x_\epsilon)$, $d_k^2 > x_\epsilon$ and hence $k \leq \epsilon d_k^{2\gamma}$. Namely, $\forall\, \epsilon > 0$ and $\forall\, k > g(x_\epsilon)$, $\frac{d_k}{k^{1/2\gamma}} \geq \frac{1}{\epsilon^{1/2\gamma}}$, therefore the thesis holds with $\lambda = \frac{1}{2\gamma}$. ■

A.6.3 Appendix to Section 7.3.2

Proof of Proposition 24. We follow the arguments used to prove Proposition 23. Let $f(x)$ be the function defined in equation (A.28) in Appendix A.6.1: by assumption, the g–function associated to $\mathcal{I}$ is such that $\lim_{x\to+\infty} \frac{g(x)}{(M/2)f(x)} = 1$. Hence, by Lemma 31.*ı* and Theorem 21.*ı* in Appendix A.6.1,

$$\lim_{\beta\to 0^+} \frac{G(\beta)}{-\frac{M\log\beta}{2\beta\log\theta}} = 1\,.$$

Using the expression for $\beta\mathcal{E}(\beta)$ deriving from equation (7.10b) we get

$$\begin{aligned}\lim_{\beta\to 0^+}\frac{\beta\mathcal{E}(\beta)}{\frac{1}{-\log\beta}} &= \lim_{\beta\to 0^+}\frac{1+\frac{\beta G'(\beta)}{G(\beta)}}{\frac{1}{\log\beta}} = \\ &\overset{(a)}{=} \lim_{\beta\to 0^+}\frac{1-\beta\frac{\frac{F(\beta)}{\beta}+\frac{e^{-\beta}}{2\beta^2\log\theta}}{F(\beta)}}{\frac{1}{\log\beta}} = \\ &= \lim_{\beta\to 0^+}\frac{-e^{-\beta}\log\beta}{2\beta F(\beta)\log\theta} = \\ &\overset{(b)}{=} 1\,,\end{aligned}$$

where equality (a) holds by Theorem 21 and Lemma 31.*ii*, equality (b) follows by Lemma 31.*i*. To sum up,

$$\begin{aligned}\lim_{\beta\to 0^+}\frac{e^{\mathbb{H}\left(\mathcal{E}(\beta)\right)}}{\log\mathcal{E}(\beta)} &= \lim_{\beta\to 0^+}\frac{\left(1+2\beta G(\beta)\right)e^{\beta\mathcal{E}(\beta)}}{\log\mathcal{E}(\beta)} = \\ &= \lim_{\beta\to 0^+}\frac{\frac{-M}{\log\theta}\log\beta}{-\log(-\log\beta)-\log\beta} = \frac{M}{\log\theta}\,. \qquad \blacksquare\end{aligned}$$

A.6.4 Appendix to Section 7.4.2

Proof of Lemma 22. *Proof of part i*: Suppose by contradiction that $\exists\,\tilde{x}\in J_e$ such that $|\tilde{x}|\geq\sqrt{\eta}$ and $|\varphi(\tilde{x})|>|\tilde{x}|$. Consider $\mu_0=\frac{\eta}{\tilde{x}^2}\delta_{\tilde{x}}+\left(1-\frac{\eta}{\tilde{x}^2}\right)\delta_0$, then $\mathcal{E}_e(\mu_0)=\eta$ whereas $\mu_1=\frac{\eta}{\tilde{x}^2}\delta_{\varphi(\tilde{x})}+\left(1-\frac{\eta}{\tilde{x}^2}\right)\delta_{\varphi(0)}$ and $\mathcal{E}_e(\mu_1)=\frac{\eta}{\tilde{x}^2}\varphi(\tilde{x})^2>\eta$.
Proof of part ii: Suppose by contradiction that $\exists\,\tilde{x}\in J_e$ such that $|\varphi(\tilde{x})|>|\tilde{x}|$. Consider $\mu_0=\delta_{\tilde{x}}$, then $\mathcal{E}_e(\mu_0)=\tilde{x}^2$ whereas $\mu_1=\delta_{\varphi(\tilde{x})}$ and $\mathcal{E}_e(\mu_1)=\varphi(\tilde{x})^2>\tilde{x}^2=\mathcal{E}_e(\mu_0)$.
Proof of part iii: For any given $\mu_0\in\mathcal{P}r(\mathbb{R})$, $\mathcal{E}_e(\mu_1)=\int_{\varphi^{-1}(J_e)}\varphi(x)^2d\mu_0\leq\int_{J_e}\varphi(x)^2d\mu_0\leq\sigma^2\int_{J_e}x^2d\mu_0=\sigma^2\mathcal{E}_e(\mu_0)$, where the first inequality holds because, since $\varphi(J_0)\subseteq J_0$, then $\varphi^{-1}(J_e)\subseteq J_e$. $\blacksquare$

Proof of Lemma 24. First notice that the class $\mathcal{P}_{\text{all}}$ is admissible: indeed, it is sufficient to show that $\mathcal{P}_{\text{all}}$ is closed under the dynamics φ and this is an easy consequence of the fact that φ is standard logarithmic.
Let $x_0>0$ and $\theta=\frac{|a|+\sigma}{|a|-\sigma}$ be as in Definition 43.*ii*. For φ as in the assumptions of the lemma, we claim that it is possible to construct a sequence $\{y_n\}_{n\in\mathbb{N}}$ such that

$$\begin{cases}\forall\,n\in\mathbb{N}\,, \quad y_n\in J_e & \text{(A.30a)}\\ \forall\,n\in\mathbb{N}\,, \quad \gamma_{nm(\sigma)}<|y_n|<x_0\theta^{nm(\sigma)+1} & \text{(A.30b)}\\ \text{for } n\geq 1\,, \quad \varphi(y_n)=y_{n-1}\,, & \text{(A.30c)}\end{cases}$$

where, for $h\in\mathbb{N}$, $\gamma_h=\frac{|a|-\sigma}{|a|}x_0\theta^{h+1}+\frac{r_0}{|a|}$.
The thesis follows by the claim, in fact: consider $\mu_0:=(1-q)\sum_{i=0}^{+\infty}q^i\delta_{y_i}$, for some $q\in]0\,,\,1[$. The energy of such a distribution is $\mathcal{E}(\mu_0)=(1-q)\sum_{i=0}^{+\infty}q^iy_i^2<(1-q)(x_0\theta)^2\sum_{i=0}^{+\infty}(q\theta^{2m(\sigma)})^i$

by inequality (A.30b). Hence, if $q \in]0, \frac{1}{\theta^{2m(\sigma)}}[$, $\mathcal{E}(\mu_0) < +\infty$. By equation (A.30c),

$$\forall\, i \in \mathbb{N} \quad \text{and} \quad \forall\, t \in \mathbb{N}, \quad \mu_t(y_i) \geq \mu_0(y_{i+t}) = (1-q)q^{i+t}, \tag{A.31}$$

therefore,

$$\begin{aligned}
\mathcal{E}_e(\mu_t) = & \int_{J_e} x^2 d\mu_t \\
\geq & \sum_{i=0}^{+\infty} y_i^2 \mu_t(y_i) \\
\overset{(a)}{\geq} & (1-q)\sum_{i=0}^{+\infty} q^{i+t} y_i^2 \\
\overset{(b)}{>} & (1-q)q^t \sum_{i=0}^{+\infty} q^i \gamma_{im(\sigma)}^2 \\
\overset{(c)}{>} & (1-q)\Big(\tfrac{|a|+\sigma}{|a|} x_0\Big)^2 q^t \sum_{i=0}^{+\infty} (q\theta^{2m(\sigma)})^i \\
= & \tfrac{1-q}{1-q\theta^{2m(\sigma)}} \Big(\tfrac{|a|+\sigma}{|a|} x_0\Big)^2 q^t,
\end{aligned}$$

where inequality (a) follows by inequality (A.31), inequality (b) by inequality (A.30b) and inequality (c) holds because $\gamma_h > \frac{|a|-\sigma}{|a|} x_0 \theta^{h+1}$. Hence,

$$\limsup_{t \to +\infty} \frac{\log \mathcal{E}_e(\mu_t)}{t} \geq \log q$$

and

$$\mathcal{T}_e \leq - \sup_{q \in]0, \frac{1}{\theta^{2m(\sigma)}}[} \limsup_{t \to +\infty} \frac{\log \mathcal{E}_e(\mu_t)}{t} \leq \log \theta^{2m(\sigma)}.$$

That is,

$$\mathcal{T}_e \leq 2m(\sigma) \log \frac{|a|+\sigma}{|a|-\sigma}.$$

Let us prove the claim, namely, let us show the existence of a sequence $\{y_n\}_{n \in \mathbb{N}}$ satisfying the conditions (A.30). With $x_h := x_0 \theta^h$, since $|\varphi(x_h)| = \sigma x_h$ (see equation (7.23)), $\exists\, \hat{h} \in \mathbb{N}$ such that $\forall\, h \geq \hat{h}$, $\varphi(x_h) \in J_e$. Hence, provided that x_0 is redefined as $x_0 := x_{\hat{h}}$, we can assume that $\forall\, h \in \mathbb{N}$, $|\varphi(x_h)| > r_0$. For $h \in \mathbb{N}$, let

$$I_h^+ :=]\gamma_h, x_{h+1}[,$$

where $\gamma_h \in]x_h, x_{h+1}[$ is such that $\varphi(\gamma_h) = \operatorname{sign}(a) r_0$. The existence of such a γ_h is guaranteed by equation (7.23) and by the fact that $|\varphi(x_h)| > r_0$ (see Fig. A.5). Moreover, it is easy to figure out that $\gamma_h = \frac{|a|-\sigma}{|a|} x_0 \theta^{h+1} + \frac{r_0}{|a|}$. Let

$$I_h^- := -I_h^+.$$

Accordingly,

$$\begin{cases}
\text{if } a > 0, & \varphi(I_h^+) =]r_0, \sigma x_{h+1}[& \text{(A.32a)} \\
\text{if } a < 0, & \begin{cases} \varphi(I_h^+) =]-\sigma x_{h+1}, -r_0[\\ \varphi(I_h^-) =]r_0, \sigma x_{h+1}[. \end{cases} & \text{(A.32b)}
\end{cases}$$

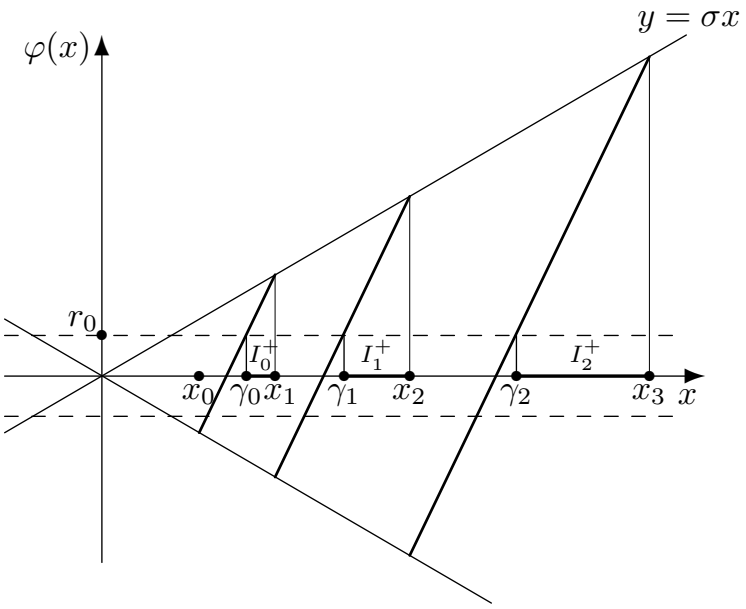

Figure A.5: Construction of the intervals I_h^+'s in the proof of Lemma 24 .

It holds that

$$\text{if} \quad h \geq m(\sigma)\,, \quad \text{then} \quad \sigma x_{h+1} \geq x_{h-m(\sigma)+1}\,. \tag{A.33}$$

In fact: $\sigma x_{h+1} \geq x_{h-m(\sigma)+1} \Leftrightarrow \sigma x_0 \theta^{h+1} \geq x_0 \theta^{h-m(\sigma)+1} \Leftrightarrow \theta^{m(\sigma)} \geq \frac{1}{\sigma} \Leftrightarrow m(\sigma) \geq -\log_\theta \sigma$ which is the case since $m(\sigma) = \lceil -\log_\theta \sigma \rceil$.
Conditions (A.32) and (A.33) imply that, for $h \geq m(\sigma)$,

$$\begin{cases} \text{if } a > 0\,, & \varphi(I_h^+) \supset I_{h-m(\sigma)}^+ & \text{(A.34a)} \\ \text{if } a < 0\,, & \begin{cases} \varphi(I_h^+) \supset I_{h-m(\sigma)}^- \\ \varphi(I_h^-) \supset I_{h-m(\sigma)}^+\,. \end{cases} & \text{(A.34b)} \end{cases}$$

We are now in the position to construct the desired sequence. Let us do it recursively: fix any $y_0 \in I_0^+$. In the case $a > 0$, assume that $\forall\, i \leq n$, y_i has been found which satisfies the properties in equation (A.30) and $y_i \in I_{im(\sigma)}^+$, then y_{n+1} can be determined as follows: $y_n \in I_{nm(\sigma)}^+ \subset \varphi\big(I_{(n+1)m(\sigma)}^+\big)$ by equation (A.34a) , therefore $\exists\, y_{n+1} \in I_{(n+1)m(\sigma)}^+$ such that $\varphi(y_{n+1}) = y_n$. The case $a < 0$ is similar, in this case $y_i \in (-1)^i I_{im(\sigma)}^+$, namely the sequence is alternating. ■

Proof of Theorem 17. For any $\sigma \in]0\,,\, 1[$, φ_σ satisfies the hypotheses of both Lemma 23 and Lemma 24 . Hence, part $\imath$ is a direct consequence of Lemma 23 $\big($and inequality (7.33)$\big)$ and of Lemma 24 $\big($and inequality (7.32)$\big)$.
The fact that, for $\mathcal{E} \to +\infty$, $\mathcal{N}_\sigma(\mathcal{E}) \sim C(\sigma) \log \mathcal{E}$, with $C(\sigma)$ as in equation (7.34) , was shown in equation (7.24) .

Inequality (7.35) is a consequence of the inequality proved in part $\imath$ and of the fact that, on an interval $]0\,,\,\sigma_0[$ (for sufficiently small $\sigma_0 > 0$), C is a decreasing function of σ and both f_1^{-1} and f_2^{-1} are decreasing functions of $\mathcal{T}_e$.
Let us show that $\lim_{\mathcal{T}_e\to+\infty} \frac{C_1(\mathcal{T}_e)}{C_2(\mathcal{T}_e)} = 1$:

$$\begin{aligned}
\lim_{\mathcal{T}_e\to+\infty} \frac{C_1(\mathcal{T}_e)}{C_2(\mathcal{T}_e)} = &\quad \lim_{\mathcal{T}_e\to+\infty} \frac{C\big(f_1^{-1}(\mathcal{T}_e)\big)}{C\big(f_2^{-1}(\mathcal{T}_e)\big)} = \\
= &\quad \left\{ \begin{array}{l} \lim_{y\to 0^+} \dfrac{C(y)}{C\big((f_2^{-1}\circ f_1)(y)\big)} = \\ y := f_1^{-1}(\mathcal{T}_e) \end{array} \right. \\
\overset{\text{(a)}}{=} &\quad \lim_{y\to 0^+} \frac{C(y)}{C\big(\frac{y(|a|-y)}{|a|+y}\big)} = \\
\overset{\text{(b)}}{=} &\quad \lim_{y\to 0^+} \frac{\log \frac{a^2+2y|a|-y^2}{a^2+y^2}}{\log \frac{|a|+y}{|a|-y}} = 1\,,
\end{aligned}$$

where in equality (a) we used the expression of f_1 given in equation (7.32) and the fact that, by equation (7.33), $f_2^{-1}(\mathcal{T}_e) = e^{-\mathcal{T}_e/2}$; in equality (b) we used the expression of C given in equation (7.34).
Finally, for $\mathcal{T}_e \to +\infty$, $C_2(\mathcal{T}_e) \sim \frac{|a|}{2} e^{\mathcal{T}_e/2}$, because

$$C_2(\mathcal{T}_e) = C\big(f_2^{-1}(\mathcal{T}_e)\big) = \frac{1}{\log \frac{|a|+e^{-\mathcal{T}_e/2}}{|a|-e^{-\mathcal{T}_e/2}}}\,.$$

■

Bibliography

[1] P. Antsaklis and J. Baillieul, Eds. (2004) Special issue on "Networked Control Systems", *IEEE Trans. Autom. Control*, 49(9).

[2] Y. Anzai (1974) A note on reachability of discrete–time quantized control systems, *IEEE Trans. Autom. Control*, 19(5); pages: 575–577.

[3] J.P. Aubin (1991) Viability Theory, *Boston, MA Birkhäuser.*

[4] S.I. Azuma and T. Sugie (2007), An Optimal Dynamic Quantization Scheme for Control With Discrete–Valued Input, Proc. of the *American Control Conference*, pages: 3576–3581.

[5] J. Baillieul (2001) Feedback Designs in Information–Based Control, Proc. of the *Workshop on Stochastic Theory and Control, Kansas*; pages: 35–57. Springer-Verlag.

[6] V. Balakrishnan and S. Boyd (1992) On computing the Worst–Case Peak Gain of Linear Systems, *Systems and Control Letters*, 19; pages: 265–269.

[7] B. Bamieh and M.A. Dahleh (1998) Open problems in ℓ_1 optimal control, In *Open problems in Mathematical Systems and Control Theory – Springer*; pages: 31–33.

[8] J.E. Bertram (1958) The effect of quantization in sampled feedback systems, *AIEE Trans. Appl. Ind.*, Pt. II vol. 77; pages: 177–181.

[9] D.P. Bertsekas and I.B. Rhodes (1971) Recursive state estimation for a set–membership description of uncertainty, *IEEE Trans. Autom. Control*, 16; pages: 117–128.

[10] A. Bicchi, A. Marigo and B. Piccoli (2002) On the Reachability of Quantized Control Systems, *IEEE Trans. Autom. Control*, 47(4); pages: 546–563.

[11] F. Blanchini (1999) Set Invariance in Control, *Automatica*, 35(11); pages: 1747–1767.

[12] F. Blanchini and S. Miani (2008) Set–Theoretic Methods in Control, *Boston, MA Birkhäuser.*

[13] P. Bolzern, P. Colaneri, G. De Nicolao and U. Shaked (2002) Guaranteed H_∞ robustness bounds for Wiener filtering and prediction, *Int. J. of Robust and Nonlinear Control*, 12; pages: 41–56.

[14] V.S. Borkar and S. Mitter (1997) LQG Control With Communication Constraints, In *Communications, Computation, Control and Signal Processing – Dordrecht, Boston*, pages: 365–373.

[15] S. Boyd and J. Doyle (1987) Comparison of peak and RMS gains for discrete–time systems, *Systems and Control Letters*, 9; pages: 1–6.

[16] R. Brockett (1997) Minimum Attention Control, Proc. of the *36th IEEE Conference on Decision and Control*, pages: 2628–2632.

[17] R. Brockett and D. Liberzon (2000) Quantized feedback stabilization of linear systems, *IEEE Trans. Autom. Control*, 45(7); pages: 1279–1289.

[18] F. Bullo and D. Liberzon (2006) Quantized control via locational optimization, *IEEE Trans. Autom. Control*, 51(1); pages: 2–13.

[19] P. Caravani and E. De Santis (2006) Quantized control via robust controlled invariance, Proc. of the *45th IEEE Conference on Decision and Control*, pages: 5501–5506.

[20] R. Carli, F. Fagnani, A. Speranzon and S. Zampieri (2008) Communication Constraints in the Average Consensus Problem, *Automatica*, 44(3); pages: 671–684.

[21] H. Cartan (1995) Elementary Theory of Analytic Functions of One or Several Complex Variables, *Dover publications, New York*.

[22] A. Cervin, J. Eker, B. Bernhardsson and K.E. Årzén (2002) Feedback–Feedforward Scheduling of Control Tasks, *Real-Time Systems*, 1(23); pages: 25–53.

[23] X. Chen and J.T. Wen (1995) A linear matrix inequality approach to discrete–time mixed $l_1/\mathcal{H}_\infty$ control problems, Proc. of the *34th IEEE Conference on Decision and Control*, pages: 3670–3675.

[24] C. Choi and T.C. Tsao (2001) H_∞ Preview Control for Discrete–Time Systems, *ASME Journal of Dynamic Systems, Measurement, and Control*, 123; pages: 117–124.

[25] P. Colaneri, A. Locatelli and J.C. Geromel (1997) Control theory and design: a RH_2 and RH_∞ viewpoint, *Academic Press, San Diego*.

[26] T.M. Cover and J.A. Thomas (1991) Elements of Information Theory, *John Wiley & Sons, Inc. New York.*

[27] R.E. Curry (1970) Estimation and control with quantized measurements, *Cambridge, MA, M.I.T. Press*

[28] M.A. Dahleh and J.B. Pearson (1987) l^1–Optimal Feedback Controllers for MIMO Discrete–Time Systems, *IEEE Trans. Autom. Control*, 32(4); pages: 314–322.

[29] D.F. Delchamps (1990) Stabilizing a Linear System with Quantized State Feedback, *IEEE Trans. Autom. Control*, 35(8); pages: 916–924.

[30] J.C. Delvenne (2006) An Optimal Quantized Feedback Strategy for Scalar Linear Systems, *IEEE Trans. Autom. Control*, 51(2); pages: 298–303.

[31] C. De Persis (2006) Nonlinear Stabilizability via Encoded Feedback: The case of Integral ISS Systems, *Automatica*, 42(10); pages: 1813–1816.

[32] C.A. Desoer and M. Vidyasagar (1975) Feedback Systems: Input–Output Properties, *New York: Academic Press.*

[33] C.E. deSouza and L. Xie (1992) On the discrete–time bounded real lemma with application in the characterization of static state feedback H_∞ controller, *Systems and Control Letters*, 18; pages: 61–71.

[34] I.J. Diaz–Bobillo and M.A. Dahleh (1993) Minimization of the Maximum Peak–to–Peak Gain: The General Multiblock Problem, *IEEE Trans. Autom. Control*, 38(10); pages: 1459–1482.

[35] G. Doetsch (1974) Introduction to the Theory and Applications of the Laplace Transformation, *Springer-Verlag.*

[36] Z. Drezner (1995) Facility Location: A Survey of Applications and Methods, *New York: Springer–Verlag.*

[37] N. Elia and M.A. Dahleh (1997) Controller Design with Multiple Objectives, *IEEE Trans. Autom. Control*, 42(5); pages: 596–613.

[38] N. Elia and M.A. Dahleh (1998) A Quadratic Programming Approach for Solving the ℓ_1 Multiblock Problem, *IEEE Trans. Autom. Control*, 43(9); pages: 1242–1252.

[39] N. Elia and S. Mitter (2001) Stabilization of Linear Systems With Limited Information, *IEEE Trans. Autom. Control*, 46(9); pages: 1384–1400.

[40] N. Elia and E. Frazzoli (2002) Quantized Stabilization of Two–Input Linear Systems: A Lower Bound on the Minimal Quantization Density, In *Hybrid Systems: Computation and Control*, pages: 179–193.

[41] N. Elia (2004) When Bode Meets Shannon: Control–Oriented Feedback Communication Schemes, *IEEE Trans. Autom. Control*, 49(9) special issue on "Networked Control Systems"; pages: 1477–1488.

[42] F. Fagnani and S. Zampieri (2003) Stability analysis and synthesis for scalar linear systems with a quantized feedback, *IEEE Trans. Autom. Control*, 48(9); pages: 1569–1584.

[43] F. Fagnani and S. Zampieri (2003) Performance evaluations of quantized stabilizers, Proc. of the *42nd IEEE Conference on Decision and Control*, pages: 1897–1901.

[44] F. Fagnani and S. Zampieri (2004) Steady state and transient performance in memoryless quantized controllers, In Proc. of the *16th Int. Symposium on the Mathematical Theory of Networks and Systems*.

[45] F. Fagnani and S. Zampieri (2004) Quantized stabilization of linear systems: complexity versus performance, *IEEE Trans. Autom. Control*, 49(9) special issue on "Networked Control Systems"; pages: 1534–1548.

[46] F. Fagnani and S. Zampieri (2004) Tree structured vector quantization and quantized LQ optimal control, *unpublished manuscript*.

[47] F. Fagnani and S. Zampieri (2005) A symbolic approach to performance analysis of quantized feedback systems: the scalar case, *SIAM Journal on Control and Optimization*, 44; pages: 816–866.

[48] F. Fagnani and B. Picasso (2006) Stabilization of Linear Stochastic Systems Under Static Quantized Control: Complexity vs Performance, In Proc. of the *17th Int. Symposium on the Mathematical Theory of Networks and Systems*.

[49] L. Farina and S. Rinaldi (2000) Positive Linear Systems: Theory and Applications, *John Wiley & Sons, New York*.

[50] W. Feller (1957) An Introduction to Probability Theory and Its Applications, *John Wiley & Sons, Inc. New York*.

[51] M. Fu and L. Xie (2005) The Sector Bound Approach to Quantized Feedback Control, *IEEE Trans. Autom. Control*, 50(11); pages: 1698–1711.

[52] R.G. Gallager (2008) Principles of Digital Communication, *Cambridge University Press*.

[53] K. Glover (1984) All optimal Hankel–norm approximations of linear multivarable systems and their $\mathcal{L}_\infty$–error bounds, *Int. Journal of Control*, 39; pages: 1115–1193.

[54] G.C. Goodwin, H. Haimovich, D.E. Quevedo and J.S. Welsh (2004) A Moving Horizon Approach to Networked Control System Design, *IEEE Trans. Autom. Control*, 49(9) special issue on "Networked Control Systems"; pages: 1427–1445.

[55] J. Hespanha, A. Ortega and L. Vasudevan (2002) Towards the control of linear systems with minimum bit–rate, In Proc. of the *15th Int. Symposium on the Mathematical Theory of Networks and Systems*.

[56] L. Hou, A.N. Michel and H. Ye (1997) Some Qualitative Properties of Sampled–Data Control Systems, *IEEE Trans. Autom. Control*, 42(12); pages: 1721–1725.

[57] D. Hristu and K. Moransen (1999) Limited Communication Control, *Systems and Control Letters*, 37(4); pages: 193–205.

[58] Z. Hurak, M. Hromcik and M. Sebek (2002) On computing the ℓ_1 Norm of a Polynomial Matrix Fraction, In Proc. of the *IEEE Int. Symp. on Computer Aided Control System Design*, pages: 278–283.

[59] P.A. Iglesias and K. Glover (1991) State–space approach to discrete–time H_∞ control, *Int. J. of Control*, 54; pages: 1031–1073.

[60] H. Ishii and B.A. Francis (2002) Stabilizing a Linear Systems by Switching Control With Dwell Time, *IEEE Trans. Autom. Control*, 47(12); pages: 1962–1973.

[61] H. Ishii and B.A. Francis (2003) Quadratic stabilization of sampled–data systems with quantization, *Automatica*, 39(10); pages: 1793–1800.

[62] Z.P. Jiang and Y. Wang (2001) Input–to–state stability for discrete–time nonlinear systems, *Automatica*, 37(6); pages: 857–869.

[63] R.E. Kalman (1956) Nonlinear aspects of sampled–data control systems, In Proc. of the *Symposium on Nonlinear Circuit Theory*, vol. VII, Brooklyn, NY: Polytechnic Press.

[64] C.Y. Kao and S.R. Venkatesh (2002), Stabilization of linear systems with limited information–Multiple input case, Proc. of the *American Control Conference*, pages: 2406–2411.

[65] M. Khammash (1996) Solution of the ℓ_1 MIMO control problem without zero interpolation, In Proc. of the *IEEE Conference on Decision and Control*, pages: 4040–4045.

[66] J.L. Kelley (1975) General Topology, *Springer*.

[67] H.K. Khalil (2002) Nonlinear systems, *Prentice Hall*.

[68] V. Kucera (1991) Analysis and design of discrete linear control systems, *Prentice Hall*.

[69] H. Kwakernaak, R. Sivan and R.C.W. Strijbos (1991) Modern signals and systems, *Prentice Hall*.

[70] A.H. Land and A.G. Doig (1960) An Automatic Method for Solving Discrete Programming Problems, *Econometrica*, 28; pages: 497–520.

[71] A. Lasota and M.C. Mackey (1994) Chaos, fractals and noise, *Springer–Verlag*.

[72] K. Li and J. Baillieul (2004) Robust quantization for digital finite communication bandwidth (DFCB) control, *IEEE Trans. Autom. Control*, 49(9) special issue on "Networked Control Systems"; pages: 1573–1584.

[73] K. Li and J. Baillieul (2005) Problems in decentralized sensor–actuator networks, Proc. of the *44th IEEE Conference on Decision and Control*, pages: 3207–3212.

[74] D. Liberzon (2003) Switching in Systems and Control, *Boston, MA Birkhäuser.*

[75] D. Liberzon (2003) On Stabilization of Linear Systems With Limited Information, *IEEE Trans. Autom. Control*, 48(2); pages: 304–307.

[76] D. Liberzon (2003) Hybrid feedback stabilization of systems with quantized signals, *Automatica*, 39(9); pages: 1543–1554.

[77] D. Liberzon and J. Hespanha (2005) Stabilization of Nonlinear Systems With Limited Information Feedback, *IEEE Trans. Autom. Control*, 50(6); pages: 910–915.

[78] D. Liberzon (2006) Quantization, Time Delays, and Nonlinear Stabilization, *IEEE Trans. Autom. Control*, 51(7); pages: 1190–1195.

[79] D. Liberzon and D. Nĕsić (2007) Input–to–State Stabilization of Linear Systems With Quantized State Measurements, *IEEE Trans. Autom. Control*, 52(5); pages: 767–781.

[80] Q. Ling and M.D. Lemmon (2005) Stability of Quantized Control Systems Under Dynamic Bit Assignment, *IEEE Trans. Autom. Control*, 50(5); pages: 734–740.

[81] R. Mañé (1987) Ergodic Theory and Differentiable Dynamics, *Springer.*

[82] A.S. Matveev and A.V. Savkin (2004) The problem of LQG optimal control via a limited capacity communication channel, *Systems and Control Letters*, 53(1); pages: 51–64.

[83] A.S. Matveev and A.V. Savkin (2005) Multi–rate stabilization of linear multiple sensor systems via limited capacity communication channels, *SIAM Journal on Control and Optimization*, 44(2); pages: 584–617.

[84] R.K. Miller, M.S. Mousa and A.N. Michel (1988) Quantization and Overflow Effects in Digital Implementations of Linear Dynamic Controllers, *IEEE Trans. Autom. Control*, 33(7); pages: 698–704.

[85] R.K. Miller, A.N. Michel and J.A.Farrel (1989) Quantizer effects on steady–state error specifications of digital control systems, *IEEE Trans. Autom. Control*, 34(6); pages: 651–654.

[86] G.H. Moore (1982) Zermelo's Axiom of Choice: Its Origins, Developments, and Influence, *Springer.*

[87] P. Moroney (1983) Issues in the Implementation of Digital Compensators, *Cambridge, MA, M.I.T. Press*

[88] G.N. Nair and R.J. Evans (2003) Exponential stabilisability of finite–dimensional linear systems with limited data rates, *Automatica*, 39(4); pages: 585–593.

[89] G.N. Nair and R.J. Evans (2004) Stabilizability of stochastic linear systems with finite feedback data rates, *SIAM Journal on Control and Optimization*, 43(2); pages: 413–436.

[90] G.N. Nair, R.J. Evans and P.E. Caines (2004) Stabilizing decentralised linear systems under data rate constraints, Proc. of the *43rd IEEE Conference on Decision and Control*, pages: 3992–3997.

[91] G.N. Nair, F. Fagnani, S. Zampieri and R.J. Evans (2007) Feedback Control Under Data Rate Constraints: An Overview, *Proceedings of the IEEE*, 95(1); pages: 108–137.

[92] D. Nešić and D. Liberzon (2005) A small–gain approach to stability analysis of hybrid systems, Proc. of the *44th IEEE Conference on Decision and Control*, pages: 5409–5414.

[93] D. Nešić and D. Liberzon (2008) A unified framework for design and analysis of networked and quantized control systems, To appear in *IEEE Trans. Autom. Control.*

[94] A. Okabe, B. Boots, K. Sugihara and S.N. Chiu (2000) Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, *New York: Wiley.*

[95] L. Palopoli (2002) Design of Embedded Control Systems under real-time scheduling constraints, *PhD Thesis: ReTiS Lab – Scuola Superiore S. Anna – Pisa.*

[96] L. Palopoli, B. Picasso and A. Bicchi (2005) The interaction of different types of constraints on control design, In report *DR.–2.3 of EC – IST research project "Real–time Embedded Control of mobile SYstems with distributed Sensing (RECSYS)"*, pages: 5–25. Available on–line at `http://recsys.s3.kth.se/publications/reports/DR2.3.pdf`

[97] B. Picasso, F. Gouaisbaut and A. Bicchi (2002) Construction of invariant and attractive sets for quantized–input linear systems, Proc. of the *41st IEEE Conference on Decision and Control*, pages: 824–829.

[98] B. Picasso, S. Pancanti, A. Bemporad and A. Bicchi (2003) Receding Horizon Control of LTI systems with quantized inputs, In Gueguen, Engell and Zaytoon, editors, Proc. of the *1st IFAC Conference on Analysis and Design of Hybrid Systems*; pages: 259–264. Elsevier.

[99] B. Picasso and A. Bicchi (2004) Some Relations Between Ergodicity and Minimality Properties of Invariant Sets in Quantized Control Systems, Poster for *S.I.C.C. Workshop on Bifurcations in nonsmooth and hybrid dynamical systems: analysis, control and applications*; Milano.

[100] B. Picasso, L. Palopoli, A. Bicchi and K.H. Johansson (2004) Control of Distributed Embedded Systems in the Presence of Unknown–but–Bounded Noise, Proc. of the *43rd IEEE Conference on Decision and Control*, pages: 1448–1453.

[101] B. Picasso and A. Bicchi (2005) Control synthesis for practical stabilization of quantized linear systems, *Rendiconti del seminario matematico dell'Università e del Politecnico di Torino*, 63(4); pages: 397–410.

[102] B. Picasso and P. Colaneri (2006) Practical Stability Analysis for Quantized Control Systems via Small–Gain Theorem, In Proc. of the *5th IFAC Symposium on Robust Control Design*.

[103] B. Picasso and A. Bicchi (2007) On the Stabilization of Linear Systems Under Assigned I/O Quantization, *IEEE Trans. Autom. Control*, 52(10); pages: 1994–2000.

[104] B. Picasso and P. Colaneri (2008) A factorization approach for the ℓ_∞–gain of discrete–time linear systems, To appear in Proc. of the *17th IFAC World Congress*.

[105] B. Picasso and A. Bicchi (2008) Hypercubes are minimal controlled invariants for discrete time linear systems with quantized scalar input, *Journal of Nonlinear Analysis: Theory, Methods, and Applications*, 2(3) special issue on "Hybrid Systems and Applications"; pages: 706–720.

[106] B. Picasso, L. Palopoli and A. Bicchi (2008) Stabilization of quantized systems, To appear in Section 6.1 of *The Hycon Handbook of Hybrid Systems Control Theory–Tools–Applications*.

[107] B. Picasso and P. Colaneri (2008) Stabilization of discrete–time quantized linear systems: an H_∞/ℓ_1 approach, Submitted to the *47th IEEE Conference on Decision and Control*.

[108] D.E. Quevedo, G.C. Goodwin, J.A. De Doná (2004) Finite constraint set receding horizon control, *Int. J. of Robust and Nonlinear Control*, 14(4); pages: 355–377.

[109] J. Raisch (1994) Simple hybrid control systems – continuous FDLTI plants with symbolic measurements and quantized control inputs, In Proc. of the *11th Int. Conf. on Analysis and Optimization of Systems*, vol. 199 of Lectures Notes in Control and Information Sciences; pages: 369–376. Springer–Verlag.

[110] A. Sahai and S. Mitter (2006) The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link–Part I: Scalar Systems, *IEEE Trans. Information Theory*, 52(8); pages: 3369–3395.

[111] A. Sahai and S. Mitter (2008) Source coding and channel requirements for unstable processes, Submitted to *IEEE Trans. Information Theory*.

[112] F.C. Schweppe (1968) Recursive state estimation: Unknown but bounded errors and system inputs, *IEEE Trans. Autom. Control*, 13(1); pages: 22–28.

[113] J.S. Shamma (1996) Optimization of the ℓ^∞–Induced Norm Under Full State Feedback, *IEEE Trans. Autom. Control*, 41(4); pages: 533–544.

[114] E.D. Sontag (1998) Mathematical Control Theory (Deterministic Finite Dimensional Systems) , *Sprnger*.

[115] J. Stillwell (1980) Classical Topology and Combinatorial Group Theory , *Sprnger*.

[116] A.A. Stoorvogel (1992) The discrete time H_∞ control problem with measurement feedback, *SIAM Journal on Control and Optimization*, 30(1); pages: 182–202.

[117] A.A. Stoorvogel (1992) The H_∞ control problem: a state space approach, *Prentice-Hall, Englewood Cliffs*. Out of print, a pdf version is available at `http://homepage.mac.com/a.a.stoorvogel/`

[118] R. Su, S. Abdelwahed and S. Neema (2005) Computing finitely reachable containable region for switching systems, *IEE Proc. Control Theory and Applications*, 152(4); pages: 477–486.

[119] M. Sznaier and M. Damborg (1989) Control of constrained discrete–time linear systems using quantized controls, *Automatica*, 25(4); pages: 623–628.

[120] M. Sznaier and A. Sideris (1994) Feedback Control of Quantized Constrained Systems with Applications to Neuromorphic Controllers Design, *IEEE Trans. Autom. Control*, 39(7); pages: 1497–1502.

[121] M. Sznaier and J. Bu (1998) Mized $l_1/\mathcal{H}_\infty$ Control of MIMO Systems via Convex Optimization, *IEEE Trans. Autom. Control*, 43(9); pages: 1229–1241.

[122] D.C. Tarraf (2006) A Finite State Machine Framework for Robust Analysis and Control of Hybrid Systems, *PhD Thesis: Massachusetts Institute of Technology*.

[123] S.C. Tatikonda (2000) Control under communication constraints, *PhD Thesis: Massachusetts Institute of Technology*.

[124] S. Tatikonda (2003) Some scaling properties of large distributed control systems, Proc. of the *42nd IEEE Conference on Decision and Control*, pages: 3142–3147.

[125] S. Tatikonda, A. Sahai and S. Mitter (2004) Stochastic Linear Control Over a Communication Channel, *IEEE Trans. Autom. Control*, 49(9) special issue on "Networked Control Systems"; pages: 1549–1561.

[126] S.C. Tatikonda and S. Mitter (2004) Control Under Communication Constraints, *IEEE Trans. Autom. Control*, 49(7); pages: 1056–1068.

[127] T. Ushio and C.S. Hsu (1987) Chaotic rounding error in digital control systems *IEEE Trans. Circuits Syst.*, vol. CAS–34; pages: 133–139.

[128] M. Vidyasagar (1986) Optimal rejection of persistent bounded disturbances, *IEEE Trans. Autom. Control*, 31(6); pages: 527–534.

[129] H.K. Wimmer (2000) Extensions of the bounded real lemma of discrete–time systems, *Int. J. of Control*, 73(14); pages: 1322–1328.

[130] W. Wong and R. Brockett (1997) Systems with finite communication bandwidth constraints - part I: State estimation problems, *IEEE Trans. Autom. Control*, 42(9); pages: 1294–1299.

[131] W. Wong and R. Brockett (1999) Systems with finite communication bandwidth constraints - part II: Stabilization with limited information feedback, *IEEE Trans. Autom. Control*, 44(5); pages: 1049–1053.

[132] G. Zames (1981) Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses, *IEEE Trans. Autom. Control*, vol. 26, no. 2, pages: 301–320.