ESTIMATION AND CONTROL WITH LIMITED INFORMATION AND UNRELIABLE FEEDBACK

BY

YOAV SHARON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Associate Professor Daniel M. Liberzon, Chair
Associate Professor Yi Ma, Co-Director of Research
Professor Naira Hovakimyan
Assistant Professor Todd P. Coleman

Abstract

Advancement in sensing technology is introducing new sensors that can provide information that was not available before. This creates many opportunities for the development of new control systems. However, the measurements provided by these sensors may not follow the classical assumptions from the control literature. As a result, standard control tools fail to maximize the performance in control systems utilizing these new sensors. In this work we formulate new assumptions on the measurements applicable to new sensing capabilities, and develop and analyze control tools that perform better than the standard tools under these assumptions. Specifically, we make the assumption that the measurements are quantized. This assumption is applicable, for example, to low resolution sensors, remote sensing using limited bandwidth communication links, and vision-based control. We also make the assumption that some of the measurements may be faulty. This assumption is applicable to advanced sensors such as GPS and video surveillance, as well as to remote sensing using unreliable communication links.

The first tool that we develop is a dynamic quantization scheme that makes a control system stable to any bounded disturbance using the minimum number of quantization regions. Both full state feedback and output feedback are considered, as well as nonlinear systems. We further show that our approach remains stable under modeling errors and delays. The main analysis tool we use for proving these results is the nonlinear input-to-state stability property. The second tool that we analyze is the Minimum Sum of Distances estimator that is robust to faulty measurements. We prove that this robustness is maintained when the measurements are also corrupted by noise, and that the estimate is stable with respect to such noise. We also develop an algorithm to compute the maximum number of faulty measurements that this estimator is robust to. The last tool we consider is motivated by vision-based control systems. We use a nonlinear optimization that is taking place over both the model parameters and the state of the plant in order to estimate these quantities. Using the example of an automatic landing controller, we demonstrate the improvement in performance attainable with such a tool.

To my Parents

Acknowledgments

I wish to express my deep gratitude to Daniel Liberzon and Yi Ma, my joint PhD advisers. I was very fortunate to work with these two distinguished researchers and teachers. They provided me with paths for research, but gave me the freedom to decide which one to follow. They were always available to offer me their wisdom, expertise and encouragement to overcome the challenges I faced. Daniel, in particular, spent many hours and days meticulously improving our publications, in order to make them as professional and accessible as possible.

I am also indebted to Gilad Adiv and my M.Sc. adviser Michael Margaliot who, each in his own way, prepared me during my prior research and academic activities for this PhD program.

Studying at University of Illinois at Urbana-Champaign has been an exciting experience I will forever cherish. Many of the excellent faculty at this university contributed to my training and this research, whether it was through their outstanding teaching or through more direct interaction. I would like to give special thanks to Todd Coleman, Robert Fossum, Naira Hovakimyan, Prashant Mehta, and to visiting professor Roberto Tempo. I would also like to thank the support staff in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory and, in particular, the Decision and Control Group: John Bambenek, Jana Lenz and Becky Lonberger.

The interaction, discussions and arguments I had with my colleagues helped make this research much better than it would otherwise have been. Among these especially are Nir Friedman, Arvind Ganesh, Peter Hokayem, Hossein Mobahi, Aneel Tanwani, Stephan Trenn, Andrew Wagner and John Wright. These and other colleagues also provided me with their invaluable friendship and made my stay at the university much more enjoyable.

I was blessed to be surrounded by such good friends in addition to my colleagues. Alina waited patiently for me during the many long hours of my research, but was always available to encourage me when I needed it. My old friend Amit Batikoff, despite the distance, always kept in touch to make sure I was doing well. Yaniv Eytani and Dan Goldwasser supported me through the hard times and were there to celebrate with me in the good times. There have been many others I cannot list here.

Finally, none of this could have happened without the endless love and support I received from my family: my parents, to whom this dissertation is dedicated, my sisters, my grandmother, and also Bertha, Benoni, and Ana.

My work here at Illinois was mainly supported by the NSF ECCS-0701676 grant. Additional grants which supported this research include NSF ECS-0134115 CAR, NSF CRS-EHS-0509151, NSF CCF-TF-0514955, ONR YIP N00014-05-1-0633 and NSF IIS 07-03756.

Table of Contents

Chapter 1

Main Introduction

Most control systems require feedback information in order to compensate for the unknowns in the system. In general these unknowns include the initial conditions, external disturbance, unmodeled dynamics, and delays. The most common type of feedback studied in the literature is output feedback corrupted by additive Gaussian noise. Classic sensors that measure the quantity of interest directly, such as ammeter, voltmeter, gyroscopes and accelerometers, and are located next to the controller, produce this type of feedback. Notable popular estimators for this type of feedback are the Luenberger Observer and the Kalman Filter, the latter being optimal under certain conditions. However, with the introduction of new types of advanced sensors and sensing capabilities, the classic assumptions on the feedback no longer hold and new types of feedback with different characteristics need to be dealt with. In this work we consider these new types of feedback and show that, by using new estimators that we develop, better performance can be achieved.

Specifically, we consider feedback that is:

Quantized: While the state is assumed to take values in a continuum (usually $\mathbb{R}^n$), the feedback available to the controller can take at most a finite number of different values. This implies that the state space is divided into a finite number of subsets, each corresponding to one feedback value. Two different valuations of the state in the same subset are indistinguishable given a single feedback measurement. Furthermore, while the state may evolve either continuously or discretely in time, the feedback is always discrete — there is only a finite number of feedback measurements in any given time interval.

Faulty: The feedback contains both faulty and noisy measurements. The noise affects all the measurement while only a small subset of the measurements is faulty. Faulty measurements, however, can be arbitrarily corrupted and there is no explicit indication that these measurements are faulty.

Motivation for these types of feedback is as follows. Many if not most control systems suffer some degree of quantization. Essentially every use of a digital controller incurs quantization due to the limited binary representation of real valued numbers and finite clock rate. The effects of this quantization, however, are usually, but not always, small relative to the other noises in the system and can therefore be neglected. In

addition to the digitization as a source of quantization, we identify two other main sources of quantization. The first source is limited data rate in the communication link connecting the sensors to the controller. The second source is the sensors themselves. A typical example, one that we will study in more detail here, is a video camera with a finite number of pixels and a finite capture rate. In general, any sensor with a specified resolution, range of operation, and sampling frequency is essentially a quantizer. These two sources of quantization can be much more significant than the quantization due to the digitization, in which case they cannot be neglected.

Corrupted measurements are usually associated with occasionally failing sensors that do not send any indication when they fail. Corruption may also occur when the assumptions by which the sensors operate fail to hold. For example, with a GPS sensor it is assumed that the signal arrives directly from the satellite and not after being reflected from surrounding objects. This assumption, however, may not always hold. Another example is when the sensor is a camera, and the quantities of interest need to be extracted from the image using a computer vision algorithm. Such algorithms, whose reliability is generally much lower than that of classic sensors, may misinterpret the image and produce erroneous results. Finally, communication links that need to transmit the feedback information from the sensors to the controller may become less reliable as the bandwidth is increased and fewer correction bits are used.

Our goal here is to still be able to compensate for the unknowns in the system, while relying on these new types of feedback. In general we seek to achieve stability with respect to the unknowns — the deviation of the system response from the desired response should be comparable to the deviation of the unknown signals or parameters from their nominal values. Naturally we also want to minimize the deviation of the system response from the desired one given the unknowns in the system. In the case of faulty sensors, we also want to achieve robustness — no matter how corrupted the faulty measurements are, the response of the system should be independently bounded.

We start by developing a controller that mitigates the effects of quantization using dynamic quantization. We continue with analyzing an estimator that is robust to faulty measurements. And we finish by designing a controller for a vision-based control system.

1.1 Organization of this Thesis

The following are more detailed summaries of each chapter:

In **Chapter 2** we consider the problem of achieving input-to-state stability (ISS) with respect to external disturbances for control systems with quantized measurements. Quantizers considered in this chapter

have an adjustable "center" and "zoom" parameters. Both the full state feedback and the output feedback cases are considered. Similarly to previous techniques from the literature, our proposed controller switches repeatedly between "zooming out" and "zooming in." However, here we use two modes to implement the "zooming in" phases, which gives us the important benefit of using the minimal number of quantization regions. Our analysis is trajectory-based and utilizes a cascade structure of the closed-loop hybrid system. We further show that the control system remains stable under modeling errors and delays in the plant dynamics using a specially adapted small-gain theorem. The main results are developed for linear systems, but we also discuss their extension to nonlinear systems under appropriate assumptions. These results were also published in [57, 58, 60, 59].

In **Chapter 3** we consider the problem of estimating a state $\boldsymbol{x}$ from noisy and corrupted linear measurements $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{z} + \boldsymbol{e}$, where $\boldsymbol{z}$ is a dense vector of small-magnitude noise and $\boldsymbol{e}$ is a relatively sparse vector whose entries can be arbitrarily large. We study the behavior of the ℓ^1 estimator $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - A\boldsymbol{x}\|_1$, and analyze its breakdown point with respect to the number of corrupted measurements $\|\boldsymbol{e}\|_0$. We show that the breakdown point is independent of the noise. We introduce a novel algorithm for computing the breakdown point for any given A, and provide a simple bound on the estimation error when the number of corrupted measurements is less than the breakdown point. As a motivating example, we apply our algorithm to design a robust state estimator for an autonomous vehicle, and show how it can significantly improve performance over the Kalman filter. These results were also published in [62].

In **Chapter 4** we consider a class of control systems where the plant model is unknown and the feedback contains only partial quantized measurements of the state. We use a nonlinear optimization that is taking place over both the model parameters and the state of the plant in order to estimate these quantities. We propose a computationally efficient algorithm for solving the optimization problem, and prove its convergence using tools from convex and non-smooth analysis. We demonstrate the importance of this class of control systems, and our method of solution, using the following application: A fixed wing airplane that follows a desired glide slope on approach to landing. The only feedback is from a camera mounted at the front of the airplane and focused on a runway of unknown dimensions. The quantization is due to the finite resolution of the camera. Using this application, we also compare our method to the basic method prevalent in the literature, where the optimization is only taking place over the plant model parameters. These results were also published in [61].

In **Chapter 5** we provide additional results which were obtained as part of the investigations reported in

the other chapters, but which have not reached sufficient maturity to be reported as separate chapters or to be included in any of the other chapters.

Chapter 2

Input to State Stabilizing Controller for Systems with Coarse Quantization

2.1 Introduction

We refer the reader to the main introduction of this thesis for the motivation behind the study of quantized feedback. The study of the influence of quantization on the behavior of feedback control systems can be traced back at least to [27]. In the literature on quantization, the quantized control system is typically regarded as a perturbation of the ideal (unquantized) one. Two principal phenomena account for changes in the system's behavior caused by quantization. The first one is saturation: if the quantized signal is outside the range of the quantizer, then the quantization error is large, and the system may significantly deviate from the nominal behavior (e.g., become unstable). The second one is deterioration of performance near the target point (e.g., the equilibrium to be stabilized): as this point is approached, higher precision is required, and so the presence of quantization errors again distorts the properties of the system. These effects can be precisely characterized using the tools of system theory, specifically, Lyapunov functions and perturbation analysis; see, e.g., [43, 11, 4] for results in this direction. We refer to this line of work as the "perturbation approach." The more recent work [31], also falling into this category, is particularly relevant because it reveals the importance of input-to-state stability—a concept we define below—for characterizing the robustness of the controller to quantization errors for general nonlinear systems.

An alternative point of view which this chapter takes, pioneered by Delchamps [11], is to regard the quantizer as an information-processing device, i.e., to view the quantized signal as providing a limited amount of information about the real quantity of interest (system state, control input, etc.) which is encoded using a finite alphabet. This "information approach" seems especially suitable in modern applications such as networked and embedded control systems. The main question then becomes: how much information is really needed to achieve a given control objective? In the context of stabilization of linear systems, one can explicitly calculate the minimal information transmission rate that will dominate the expansiveness of the underlying system dynamics. Results in this direction are reported in [75, 4, 44, 50, 1, 32] and in the papers cited in the next paragraph; see also [34, 46, 9, 28] for extensions to nonlinear systems.

All the aforementioned works only addressed stability in the absence of external disturbances. The papers that did address the issue of external disturbances are cited below. They differ mainly in the stability property they aim to achieve, and in their assumptions on the external disturbance. Papers [22], [70] and [38] designed a controller which guarantees stability only for a disturbance whose magnitude is lower than some known value. In the paper [45] mean square stability in the stochastic setting is obtained by utilizing statistical information about the disturbance (a bound on its appropriate moment). The paper [39] designed a controller with which it is possible to bound the plant's state in probability. With the expense of one additional feedback bit, no further information about the disturbance is required. Note that these two latter papers use (and prove) stochastic stability notions. All of these papers followed the information approach. Deterministic stability for a completely unknown bounded disturbance was initially shown in [35]. By generalizing the perturbation approach of [4, 31], the deterministic stability property achieved in [35] is input-to-state stability (ISS) which, apart from ensuring a bounded state response to every bounded disturbance, also ensures asymptotic stability (convergence to the origin) when the disturbance converges to zero. The approach of [35] was also shown to produce ℓ_2 stability in [28] (also, [29]).

In this chapter we also address the problem of achieving ISS for deterministic systems and completely unknown disturbance. In contrast to [35], which followed the perturbation approach, our first and main contribution here is that we do this following the information approach. The main advantage of using the information approach is that it requires fewer, possibly many fewer, quantization regions, which also translates to lower data rate. As a result, a better understanding is achieved of how much information is required for ISS disturbance attenuation. In fact, when all state variables are observed (quantized state feedback) we are able to achieve a data rate which can be arbitrarily close to the minimal data rate required for stabilization with no disturbance. We stress that following the information approach and not the perturbation approach necessitates significantly different design and analysis tools than what is described in [35].

Our second contribution is that we also consider the case where the state space is only partially measured, the situation commonly referred to as output feedback. This is a significant generalization of the approach described in [32], where only a specific observer was given and no disturbances were considered. The papers [45], [39] and [9] do formulate a system with output feedback, but it is assumed there that a state estimate is generated before the quantization is applied ([9] does not deal with disturbances). Here we generate the state estimate from the quantized measurements. We argue that this setting is much more reasonable when the quantization is due to physical or practical constraints on the sensor (as opposed to just a data rate constraint); refer to Remark 2.2.2 for more details. We emphasize that our results are novel even for the state feedback case.

Our third contribution, which was not discussed in any of the previous papers, is stability to modeling errors where the system model is known only approximately, and may also vary over time. We show that under small enough modeling errors the system remains ISS in a local practical sense. We prove this stability result using a specially adapted small-gain theorem.

Our fourth contribution is an extension to the case where (possibly time-varying) delays are present in addition to state quantization. Although we assume there are no external disturbances when dealing with delays, we do rely on the ISS property which we establish here after we show that error signals which arise due to delays can be regarded as external disturbances. The ISS small-gain analysis employed here to deal with delays is similar in spirit to that used in [33], but it becomes more challenging due to the dynamics of the quantizer which are necessary to achieve minimal data rate (in [33] only a static quantizer was considered). We also restrict ourselves to linear plant dynamics in the context of delays, but our method is nonlinear in nature and we expect it to naturally extend to suitable nonlinear systems along the lines of §2.6. Among other noteworthy references dealing with quantization and delays, using approaches different from ours, we mention [19], [10], and [54]. The first two of these papers employ Lyapunov-Krasovskii functionals for linear and nonlinear systems, respectively, while the last one handles nonlinear systems by sending time information along with the encoded state.

The chapter is organized as follows. In §2.2.1 we define the system and the specific quantizer we will use. In §2.2.2 we define the desired stability property, an extension of the ISS property. In §2.3 we present the proposed controller. In §2.4 we state and prove our main results pertaining to the first three contributions. In §2.5 we show that we can arbitrarily approach the minimum data-rate for the unperturbed system. In §2.6 we show how our results can be extended to nonlinear systems. And, finally, in §2.7 we state and prove the results pertaining to delays. In §2.8 we show that the small-gain theorem applies to our modified ISS notion. We defer to §2.9 the proofs of our technical lemmas.

2.2 Problem Statement

2.2.1 System Definition

The continuous-time dynamical system we are to stabilize is as follows ($t \in \mathbb{R}_{\geq 0}$):

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + D\boldsymbol{w}(t) \\ \boldsymbol{y}(t) &= C\boldsymbol{x}(t) \end{aligned} \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{n_x}$ is the state, $\boldsymbol{u} \in \mathbb{R}^{n_u}$ is the control input, $\boldsymbol{w} \in \mathbb{R}^{n_w}$ is an unknown disturbance, and $\boldsymbol{y} \in \mathbb{R}^{n_y}$ is the measured output ($n_y \leq n_x$).

While $\boldsymbol{y}$ is what the sensors measure, we assume that the information available to the controller is $\boldsymbol{z} : \{kT_s \,|k \in \mathbb{Z}_{\geq 0}\} \to \mathbb{R}^{n_y}$, which is a sampled and quantized version of $\boldsymbol{y}$:

$$\boldsymbol{z}(kT_s) = Q(\boldsymbol{y}(kT_s); \boldsymbol{c}(kT_s), \mu(kT_s)) \tag{2.2}$$

where Q is a quantization function and $T_s > 0$ is the time-sampling interval. The quantization parameters, $\boldsymbol{c} : \{kT_s \,|k \in \mathbb{Z}_{\geq 0}\} \to \mathbb{R}^{n_y}$ and $\mu : \{kT_s \,|k \in \mathbb{Z}_{\geq 0}\} \to \mathbb{R}_{>0}$, are generated by the controller. For convenience we will use the notation $\boldsymbol{z}_k \doteq \boldsymbol{z}(kT_s)$, and similarly for other variables, so (2.2) becomes $\boldsymbol{z}_k = Q(\boldsymbol{y}_k; \boldsymbol{c}_k, \mu_k)$. We refer to the special case where $C = I$, the identity matrix, as the quantized state feedback problem. We refer to the general case where C is arbitrary as the quantized output feedback problem.

We will present our results using the following (square) quantizer. We assume N is an odd number, $N \geq 3$, which counts the number of quantization regions per observed dimension. The quantizer is denoted by $\left(Q_1, \ldots, Q_{n_y}\right)^T = Q(\boldsymbol{y}; \boldsymbol{c}, \mu)$ where each scalar component is defined as follows (see Figure 2.1 for an illustration):

$$Q_i(\boldsymbol{y}; \boldsymbol{c}, \mu) \doteq c_i + \begin{cases} (-N+1)\mu & y_i - c_i \leq (-N+2)\mu \\ (-N+3)\mu & (-N+2)\mu < y_i - c_i \leq (-N+4)\mu \\ \vdots & \vdots \\ 0 & -\mu < y_i - c_i \leq \mu \\ \vdots & \vdots \\ (N-3)\mu & (N-4)\mu < y_i - c_i \leq (N-2)\mu \\ (N-1)\mu & (N-2)\mu < y_i - c_i. \end{cases} \tag{2.3}$$

We will refer to $\boldsymbol{c}$ as the *center* of the quantizer, and to μ as the *zoom factor*. Note that what will actually be transferred from the quantizer to the controller will be an index to one of the quantization regions. The controller, which either generates the values $\boldsymbol{c}$ and μ or knows the rule by which they are generated,[1] will use this information to convert the received index to the value of Q as given in (2.3).

Remark 2.2.1. All of our results, except for those in §2.5, actually apply to a more general family of

[1]The quantization parameters $\boldsymbol{c}$ and μ can be available to the sensors (or the sensor side of the communication link) depending on the source of quantization. When the source of quantization is the communication, and there is sufficient computation capability on the sensor side of the communication link, the quantization parameters $\boldsymbol{c}$ and μ may be generated simultaneously on both sides of the communication link. When the source of quantization is the sensors, these quantities can be generated by the controller only and then sent to the sensors.

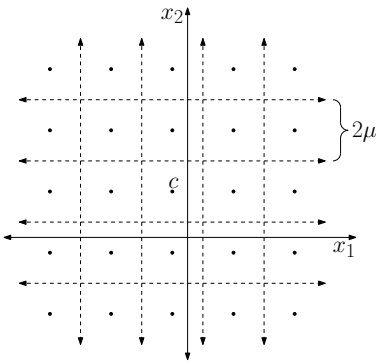

Figure 2.1: Illustration of the quantizer for the two-dimensional output subspace, $N = 5$. The dashed lines define the boundaries of the quantization regions. The black dots define the quantization values.

quantizers. For an arbitrary quantizer, we denote by $\mathcal{Q}(\boldsymbol{c}, \mu)$ the (finite) set of possible values of $Q(\cdot; \boldsymbol{c}, \mu)$. A quantizer belongs to the family of quantizers to which our results apply if there exist real numbers $M > 1$ and $0 \leq H \leq N - 1$ such that for all $\boldsymbol{y}$, $\boldsymbol{c}$ and μ there exists a set $\mathcal{Q}_{INT}(\boldsymbol{c}, \mu) \subsetneq \mathcal{Q}(\boldsymbol{c}, \mu)$ for which the following implications hold:

$$|\boldsymbol{y} - \boldsymbol{c}| < M\mu \quad \Rightarrow \quad |Q(\boldsymbol{y}; \boldsymbol{c}, \mu) - \boldsymbol{y}| < \mu \tag{2.4}$$

$$|\boldsymbol{y} - \boldsymbol{c}| < (M - H)\mu \quad \Rightarrow \quad Q(\boldsymbol{y}; \boldsymbol{c}, \mu) \in \mathcal{Q}_{INT}(\boldsymbol{c}, \mu) \tag{2.5}$$

$$Q(\boldsymbol{y}; \boldsymbol{c}, \mu) \in \mathcal{Q}_{INT}(\boldsymbol{c}, \mu) \quad \Rightarrow \quad |Q(\boldsymbol{y}; \boldsymbol{c}, \mu) - \boldsymbol{y}| < \mu. \tag{2.6}$$

It is easy to see that the square quantizer above belongs to this family with $M = N$, $H = 2$ and with $\mathcal{Q}_{INT}(\boldsymbol{y}; \boldsymbol{c}, \mu) = \left\{ \left(c_1 + q_1\mu, \ldots, c_{n_y} + q_{n_y}\mu\right) | q_i \in [-N + 3, -N + 5, \ldots N - 3], \forall i \right\}$ when the ∞-norm is considered.

Remark 2.2.2. In the literature on quantization there appear to be two different methods of positioning the partial measurement constraint (output feedback) in the feedback loop. One approach, followed by [45], [39] and [9], assumes that while not all the state variables are observed, those that are observed are measured continuously. These continuous measurements are fed into an observer which generates a state estimate. This state estimate is sent through a communication link to the controller (and thus has to be quantized). The second approach, followed by [32] and this chapter, assumes that the measurements of the observed state variables are quantized, and from these quantized measurements a state estimate needs to be generated. The reason for having two approaches is the different possible sources of quantization: Both approaches can handle the case when the communication is the source of quantization; however, only the second approach can handle the case when the sensors are the source of quantization.

In this chapter we will use the ∞-norm unless otherwise specified. For vectors, $|\boldsymbol{x}| \doteq |\boldsymbol{x}|_\infty \doteq \max_i |x_i|$.

For continuous-time signals, $\|\boldsymbol{w}\|_{[t_1,t_2]} \doteq \max_{t\in[t_1,t_2]} |\boldsymbol{w}(t)|_\infty$, $\|\boldsymbol{w}\| \doteq \|\boldsymbol{w}\|_{[0,\infty)}$. For discrete-time signals, $\|\boldsymbol{z}\|_{\{k_1,\ldots,k_2\}} \doteq \max_{k\in\{k_1,\ldots,k_2\}} |\boldsymbol{z}_k|_\infty$, $\|\boldsymbol{z}\| \doteq \|\boldsymbol{z}\|_{\{0,\ldots,\infty\}}$. For matrices we use the induced norm corresponding to the specified norm (∞-norm if none specified). For piecewise continuous signals we will use the superscripts $+$ and $-$ to denote the right and left continuous limits, respectively: $\boldsymbol{x}_k^+ \doteq \boldsymbol{x}^+(kT_s) \doteq \lim_{t\searrow 0} \boldsymbol{x}(kT_s+t)$, $\boldsymbol{x}_k^- \doteq \boldsymbol{x}^-(kT_s) \doteq \lim_{t\nearrow 0} \boldsymbol{x}(kT_s+t)$.

2.2.2 Desired Stability Property

Ideally we would want our closed-loop system to be asymptotically stable. In the presence of a non-vanishing disturbance, even linear state feedback systems without quantization cannot be driven asymptotically to the origin. Therefore, we aim for a weaker property: that the system be bounded and converge to a ball around the origin whose size depends on the magnitude of the disturbance. Furthermore, when the disturbance vanishes, we expect to recover asymptotic stability. This desired behavior is encapsulated by the (global) ISS property, originally defined in [63] as follows:

$$|\boldsymbol{x}(t)| \leq \beta\left(|\boldsymbol{x}(t_0)|, t-t_0\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right), \qquad \forall t \geq t_0 \geq 0 \tag{2.7}$$

where γ is a function of class $\mathcal{K}_\infty$ and β is a function of class $\mathcal{KL}$[2].

In addition to the original state variables, $\boldsymbol{x}$, the closed-loop system will contain other variables. Of these additional variables, the zoom factor in particular will not exhibit an ISS relation with respect to the disturbance. We refer the reader to the discussion in [35, §III.B] which explains why it is hard and probably impossible to have both the original state and the zoom factor exhibit an ISS relation with respect to the disturbance. Nevertheless, the value of the zoom factor at an arbitrary initial time will affect the ISS relation between the disturbance and the state. Therefore, the property that we will achieve, which we refer to as *parameterized input-to-state stability*, is defined as:

$$\begin{aligned} |\boldsymbol{x}(t)| &\leq \beta\left(|\boldsymbol{x}(t_0)|, t-t_0; \mu(t_0)\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu(t_0)\right), \qquad \forall t \geq t_0 \geq 0 \\ \mu(t) &\leq \delta\left(\|\boldsymbol{x}\|_{[t_0,t]}; \mu(t_0)\right) \end{aligned} \tag{2.8}$$

where the functions $\beta(\cdot,\cdot;\mu)$ and $\gamma(\cdot;\mu)$ are of class $\overline{\mathcal{KL}}$ and class $\overline{\mathcal{K}}_\infty$, respectively. We say that a function $\beta(\nu,t;\mu)$ is of class $\overline{\mathcal{KL}}$ when, as a function of its first two arguments with the third argument fixed, it is of class $\mathcal{KL}$, and it is a continuous function of its third argument when the first two arguments are fixed. We

[2] A function $\alpha : [0,\infty) \to [0,\infty)$ is said to be of class $\mathcal{K}$ if it is continuous, strictly increasing, and $\alpha(0) = 0$. A function $\alpha : [0,\infty) \to [0,\infty)$ is said to be of class $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and also unbounded. A function $\beta : [0,\infty) \times [0,\infty) \to [0,\infty)$ is said to be of class $\mathcal{KL}$ if $\beta(\cdot,t)$ is of class $\mathcal{K}$ for each fixed $t \geq 0$ and $\beta(s,t)$ decreases to 0 as $t \to \infty$ for each fixed $s \geq 0$.

say that a function $\gamma(\nu;\mu)$ if of class $\overline{\mathcal{K}}_\infty$ when as a function of its first argument with the second argument fixed, it is of class $\mathcal{K}_\infty$, and it is a continuous function of its second argument when the first argument is fixed.

In the case of modeling errors, even this cannot in general be achieved. Namely, we cannot achieve a global result, only a local one; furthermore, even with no external disturbance, the system will only be practically stable, not asymptotically stable. The weaker result we do achieve in the case of modeling error is *local practical input-to-state stability*: There exist $x_{\max}$, $w_{\max}$ and $\delta_{A,\max}$ such that if $\delta_A \leq \delta_{A,\max}$ where $\delta_A \in \mathbb{R}_{\geq 0}$ is a measure of the modeling errors, then

$$\begin{aligned} &|\boldsymbol{x}(t)| \leq \beta\left(|\boldsymbol{x}(t_0)|, t-t_0\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) + \lambda(\delta_A) \quad \forall t \geq t_0 \geq 0, \\ &\forall\, |\boldsymbol{x}(0)| < x_{\max} \quad \forall\, \|\boldsymbol{w}\|_{[0,t]} < w_{\max}. \end{aligned} \tag{2.9}$$

In (2.9) β is a function of class $\mathcal{KL}$, and γ and λ are functions of class $\mathcal{K}_\infty$. This property is along the lines of the input-to-state practical stability (ISpS) [25].

2.3 Controller Design

2.3.1 Overview of the Controller Design

Our controller switches between three different modes of operation. The motivation for each of these modes is given in this subsection.

Our quantizer consists of quantization regions of finite size, for which the quantization error, $\boldsymbol{e}_k = \boldsymbol{z}_k - \boldsymbol{y}_k$, can be bounded, and regions of infinite size, where the quantization error is unbounded. We will refer to these regions as bounded and unbounded quantization regions, respectively. Due to the fact that there are only a finite number of quantization regions to cover the infinite-size output subspace $\mathbb{R}^{n_y}$, only a subset of finite size of this subspace can be covered by the bounded quantization regions. The size of this subset, however, can be adjusted dynamically by changing the parameters of the quantizer. We refer to this subset which is covered by only bounded quantization regions as the unsaturated region. Our controller follows the general framework which was introduced in [4, 31] to stabilize the system from an unknown initial condition using dynamic quantization. In [35] this approach was developed further to achieve disturbance attenuation. This framework consists of two main modes of operation, generally referred to as the *zoom-in* and the *zoom-out* mode. During the *zoom-out* mode, the unsaturated region is enlarged until the measured output is captured in this region and a state estimate with a bounded estimation error can be established. This is followed by

a switch to the *zoom-in* mode. During the *zoom-in* mode, the size of the quantization regions is reduced in order to have the state estimate converge to the true state. The reduction of the size of the quantization regions inevitably reduces the size of the unsaturated region. As the size of this region is reduced, eventually the unknown disturbance may drive the measured output outside the unsaturated region. To regain a new state estimate with a bounded estimation error, the controller will switch back to the *zoom-out* mode. By switching repeatedly between these two modes, an ISS relation can be established. In this chapter we use the name *capture* mode for the *zoom-out* mode.

In our quantizer there are $2n_y$ unbounded quantization regions. To achieve the minimum data-rate, however, we are required to use the unbounded regions not only to detect saturation, but also to reduce the estimation error. This dual use is accomplished by dividing the *zoom-in* mode into two modes: a *measurement-update* mode and an *escape-detection* mode. After receiving r successive measurements in bounded quantization regions, where r is the observability index of the pair (A, C), we are able to define a region in the state space which must contain the state if there were no disturbance. We enlarge this region proportionally to its current size to accommodate some disturbance. In the *measurement-update* mode we cover this containment region using both the bounded and the unbounded regions of the quantizer. This way we are able to use the smallest quantization regions, which leads to the fastest reduction in the estimation error. The drawback with this mode is that if a strong disturbance comes in, we will not be able to detect it. Therefore, in the *escape-detection* mode we use larger quantization regions to cover the containment region using only the bounded regions. Thus, if a strong disturbance does come in, we will be able to detect it as the quantized output measurement will correspond to one of the unbounded regions. Note that having these two *zoom-in* modes is especially critical when there are only three quantization regions per dimension. If we had used only the *escape-detection* mode, which is necessary to detect escape, then the unsaturated region would contain only one quantization region. Having only one quantization in the unsaturated region does not provide any additional information, besides the distinction of whether an escape occurred, that can be used to reduce the estimation error. Following are the precise details on how to design the controller.

2.3.2 Preliminaries

In this section we assume that $A \equiv A_0$ is fixed and known. Extension to varying, unknown A will be discussed in §2.4.3.

We define the sampled-time versions of A, $\boldsymbol{u}$ and $\boldsymbol{w}$ as ($k \in \mathbb{Z}_{\geq 0}$):

$$A_d \doteq \exp(T_s A_0), \quad \boldsymbol{x}_k \doteq \boldsymbol{x}(kT_s), \quad \boldsymbol{u}_k^d \doteq \int_0^{T_s} \exp(A_0(T_s - t)) B\boldsymbol{u}(kT_s + t)\, dt,$$
$$\boldsymbol{w}_k^d \doteq \int_0^{T_s} \exp(A_0(T_s - t)) D\boldsymbol{w}(kT_s + t)\, dt.$$

With these definitions we can write

$$\boldsymbol{x}_{k+1} = A_d \boldsymbol{x}_k + \boldsymbol{u}_k^d + \boldsymbol{w}_k^d. \tag{2.10}$$

We assume that (A, B) is a controllable pair, so there exists a matrix K such that $A + BK$ is Hurwitz. By construction A_d is full rank, and in general (unless T_s belongs to some set of measure zero) the observability of the pair (A, C) implies that (A_d, C) is an observable pair (see [64, Proposition 6.2.11]). Thus there exists a positive integer r, the observability index, such that

$$\widetilde{C} \doteq \begin{pmatrix} CA_d^{-r+1} \\ \vdots \\ CA_d^{-1} \\ C \end{pmatrix} = \begin{pmatrix} C \\ CA_d \\ \vdots \\ CA_d^{r-1} \end{pmatrix} A_d^{-r+1} \tag{2.11}$$

has full column rank. For state feedback systems $r = 1$ and $\tilde{C}$ is the identity matrix.

2.3.3 Implementation

Our controller consists of three elements: an observer which generates a state estimate $\hat{x}(t)$ (with the notation $\hat{x}_k \doteq \hat{x}(kT_s)$); a switching logic which updates the parameters for the quantizer and sends update commands to the observer; and a stabilizing control law which computes the control input based on the state estimate.

For simplicity of presentation, we assume the stabilizing control law consists of a static nominal state feedback:

$$\boldsymbol{u}(t) = K\hat{\boldsymbol{x}}(t). \tag{2.12}$$

However, any control law that will render the closed-loop system ISS with respect to the disturbance and the state estimation error will work with our controller.

Given an update command from the switching logic, the observer generates an estimate of the state based on current and previous quantized measurements. We require the state estimate to be exact in the absence of measurement error and disturbance, and to be a linear function of the measurements. For concreteness,

we use the following state estimate from [32] which is based on the pseudo-inverse:

$$\hat{\boldsymbol{x}}_k = G\left(\boldsymbol{z}; \boldsymbol{u}^d; k\right) \doteq \widetilde{C}^{\dagger} \begin{bmatrix} \boldsymbol{z}_{k-r+1} + C\sum_{i=1}^{r-1} A_d^{-i} \boldsymbol{u}_{k-r+i}^d \\ \vdots \\ \boldsymbol{z}_{k-1} + CA_d^{-1} \boldsymbol{u}_{k-1}^d \\ \boldsymbol{z}_k \end{bmatrix}, \qquad \widetilde{C}^{\dagger} \doteq \left(\widetilde{C}^T \widetilde{C}\right)^{-1} \widetilde{C}^T. \tag{2.13}$$

In [58] we presented additional approaches to generate a state estimate that satisfy the above requirements, and compared their properties. Note that we must have at least r successive measurements to generate a state estimate. Therefore, (2.13) is defined only for $k \geq r-1$. In the special case of state feedback, on which we will comment further as we present our results, the state estimate will then be generated simply as $\hat{\boldsymbol{x}}_k = \boldsymbol{z}_k$.

The observer continuously updates the state estimate based on the nominal system dynamics:

$$\dot{\hat{\boldsymbol{x}}}(kT_s + t) = A_0 \hat{\boldsymbol{x}}(kT_s + t) + B\boldsymbol{u}\left(kT_s + t\right), \qquad t \in [0, T_s). \tag{2.14}$$

The control input is integrated continuously to generate $\boldsymbol{u}_k^d$ (initialized to zero at every $t = kT_s$):

$$\dot{\boldsymbol{u}}_k^d = \exp\left(A\left(\left(k+1\right)T_s - t\right)\right) B\boldsymbol{u}\left(t\right) \qquad \forall t \in [kT_s, \left(k+1\right)T_s].$$

2.3.4 Switching Logic

The switching logic will keep and update a discrete time step variable, $k \in \mathbb{N}$, whose value will correspond to the current sampling time of the continuous system – at each sampling time, the switching logic will update $\hat{x}_k \doteq \hat{x}\left(kT_s\right)$ where k is the discrete time step. At each discrete time step, the switching logic will operate in one of three modes: *capture*, *measurement update* or *escape detection*. The current mode will be stored in the variable $mode(k) \in \{capture, update, detect\}$. The switching logic will also use $p_k \in \mathbb{Z}$ and $saturated(k) \in \{\textbf{true}, \textbf{false}\}$ as auxiliary variables.

We assume the control system is activated at $k = 0$ ($t = 0$). We initialize $\hat{\boldsymbol{x}}_0 = \boldsymbol{0}$, $mode(0) = capture$, $p_0 = 0$, and $\mu_{-1} = s$, where s is any positive constant which will be regarded as a design parameter. We also have the following design parameters: $\alpha \in \mathbb{R}_{>0}$, $\Omega_{\text{out}} \in \mathbb{R}$ such that $\Omega_{\text{out}} > \|A\|$, and $P \in \mathbb{Z}$ such that $P \geq r+1$. In §2.3.5 we provide a qualitative discussion on how each design parameter affects the system performance. We also define

$$F\left(\mu; k\right) \doteq \left\| CA_d \widetilde{C}^{\dagger} \right\| \left\| \mu \right\|_{\{k-r,\ldots,k-1\}} \tag{2.15}$$

which in the case of state feedback reduces to $F(\mu;k) \doteq \|A_d\|\,\mu_{k-1}$.

At each discrete time step, k, the switching logic is implemented by sequentially executing the following algorithms:

Algorithm 1 Preliminaries

if $mode(k) = capture$ **then**
  set $\mu_k = \Omega_{out}\mu_{k-1}$
else if $mode(k) = update$ **then**
  set

$$\mu_k = \frac{F(\mu;k) + \alpha\|\mu\|_{\{k-r-p_{k-1}\ldots k-1-p_{k-1}\}}}{N} \tag{2.16}$$

else if $mode(k) = detect$ **then**
  set

$$\mu_k = \frac{F(\mu;k) + \alpha\|\mu\|_{\{k-r-p_{k-1}\ldots k-1-p_{k-1}\}}}{N-2} \tag{2.17}$$

end if
have the observer record $\boldsymbol{z}_k = Q\left(\boldsymbol{y}(kT_s); C\hat{\boldsymbol{x}}_k, \mu_k\right)$
if $\exists i$ such that $(\boldsymbol{z}_k)_i = (C\hat{\boldsymbol{x}}_k)_i \pm (N-1)\mu_k$ **then**
  set $saturated(k) =$ **true**
else
  set $saturated(k) =$ **false**
end if
initialize the next mode to be the same as the current mode: $mode(k+1) = mode(k)$

Algorithm 2 *capture* mode

if $mode(k) = capture$ **then**
  if $saturated(k)$ **then**
    set $p_k = 0$
  else
    set $p_k = p_{k-1} + 1$
    if $p_k = r$ **then**
      set $p_k = 0$
      have the observer update the state estimate: $\hat{\boldsymbol{x}}_k = G\left(\boldsymbol{z}; \boldsymbol{u}_d; k\right)$
      switch to the *measurement update* mode: set $mode(k+1) = update$
    end if
  end if
end if

Algorithm 3 *measurement update* mode

if $mode(k) = update$ **then**
  set $p_k = p_{k-1} + 1$
  have the observer update the state estimate: $\hat{\boldsymbol{x}}_k = G\left(\boldsymbol{z}; \boldsymbol{u}_d; k\right)$
  if $p_k = P - r$ **then**
    switch to the *escape detection* mode: set $mode(k+1) = detect$
  end if
end if

Algorithm 4 *escape detection* mode

if $mode(k) = detect$ **then**
  if not $saturated(k)$ **then**
    set $p_k = p_{k-1} + 1$
    have the observer update the state estimate: $\hat{\boldsymbol{x}}_k = G(\boldsymbol{z}; \boldsymbol{u}_d; k)$
    if $p_k = P$ **then**
      set $p_k = 0$
      switch to the *measurement update* mode: set $mode(k+1) = update$
    end if
  else
    set $p_k = 0$ and $\mu_k = s$
    switch to *capture* mode: set $mode(k+1) = capture$
  end if
end if

2.3.5 Sensitivity to Design Parameters

Any choice for the design parameters will render the closed-loop system ISS as long as the convergence property, do be defined in §2.4, holds. However, different choices will result in a different ISS gain and overshoot, and may also affect performance measures which are not expressed by the ISS definition, such as energy gain and data rate. By ISS gain we refer to the γ function in the ISS definition, and by overshoot we refer to the β function in the ISS definition. A gain or overshoot will be smaller (or bigger) if it is smaller (or bigger) for any chosen bound on the disturbance, for any initial condition and for all $t \geq 0$.

The parameter α expresses the sensitivity of the system to the disturbance and it is bounded from above by the requirement to satisfy the convergence property. The ISS gain and the overshoot will decrease as α is increased. However, increasing α will slow down the convergence of the system in the *zoom-in* sequence. By taking more quantization regions we may use a larger α, but that will also require higher data rate. The parameter P will have similar effects: higher P will result in faster convergence and smaller data rate, but the ISS gain and overshoot will be increased.

The parameters μ_0, s, and Ω_{out} have more complicated effects on the ISS gain and the overshoot, and their optimal values, when the goal is to minimize the ISS gain or the overshoot, depend on the characteristics of the disturbance and the initial condition. The choice of μ_0 will depend on the expected magnitude of the initial condition, and it will affect the overshoot only. The choice of s will depend on the expected magnitude of the disturbance and it will affect the ISS gain only. Last, the choice of Ω_{out} will depend on the expected deviation of the initial condition and the disturbance from their expected values. None of these three design parameters will affect the data rate.

2.4 Main Results

2.4.1 The Convergence Property

We define the following convergence property which implies that in an infinite sequence in which the switching logic is never in the *capture* mode (a result of having no disturbance), $\lim_{k\to\infty}\mu_k = 0$.

Set μ' as

$$\begin{aligned}
\mu'_k &= 1, && k \in \{0 \dots r-1\} \\
\mu'_k &= \frac{F(\mu';k)+\alpha}{N}, && k \in \{r \dots P-1\} \\
\mu'_k &= \frac{F(\mu';k)+\alpha}{N-2}, && k \in \{P \dots P+r-1\}.
\end{aligned} \tag{2.18}$$

If there exists $\sigma < 1$ for which the following holds

$$\|\mu'\|_{\{P,\dots,P+r-1\}} \leq \sigma \tag{2.19}$$

then we say that the observer has the *convergence property.*

Whether the observer has the convergence property depends on the choice of the design parameters α and P. The following Lemma (proved in the Appendix) gives a sufficient and easy to verify condition for the existence of design parameters with which the observer will have the convergence property.

Lemma 2.4.1. *If the following condition holds:*

$$\sigma_{pi} \doteq \frac{1}{N}\left\|CA_d\widetilde{C}^{\dagger}\right\| < 1 \tag{2.20}$$

then it is always possible to choose P and α such that the observer will possess the convergence property.

In the state feedback case we do not need an observer as the updates of the state estimate become simply $\hat{\boldsymbol{x}}_k = G(\boldsymbol{z}, \boldsymbol{u}_d, k) = \boldsymbol{z}_k$. In this case we just say that the control system has or does not have the convergence property. Note also that in this case (2.20) becomes $\|A_d\|/N < 1$.

2.4.2 Results for When the System Model Is Known

The state estimation error is defined as

$$\tilde{\boldsymbol{x}}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t). \tag{2.21}$$

In the simpler case where $A \equiv A_0$, the evolution of the state estimation error is independent of the state. This property is critical in proving the following proposition, which is the main technical step for deriving the desired stability results.

Proposition 2.4.1. *If we implement the controller with the above algorithm and the observer has the convergence property, then the state estimation error of the closed-loop system will satisfy the parameterized-ISS property with respect to the disturbance:*

$$\begin{aligned} |\tilde{\boldsymbol{x}}(t)| &\leq \beta_e\left(|\tilde{\boldsymbol{x}}(t_0)|, t-t_0; \mu(t_0)\right) + \gamma_e\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu(t_0)\right), \qquad \forall t \geq t_0 \geq 0 \\ \mu(t) &\leq \delta_e\left(\|\tilde{\boldsymbol{x}}\|_{[t_0,t]}, \mu(t_0)\right) \end{aligned} \tag{2.22}$$

where β_e, γ_e and δ_e have the same properties as β, γ and δ in (2.8)*, respectively.*

Our first stability result is the following:

Theorem 2.4.2. *With the assumptions of Proposition 2.4.1, the closed-loop system will be parameterized-ISS with respect to the disturbance:*

$$\begin{aligned} \left|\begin{pmatrix} \boldsymbol{x}(t) \\ \tilde{\boldsymbol{x}}(t) \end{pmatrix}\right| \leq& \beta\left(\left|\begin{pmatrix} \boldsymbol{x}(t_0) \\ \tilde{\boldsymbol{x}}(t_0) \end{pmatrix}\right|, t-t_0; \mu(t_0)\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu(t_0)\right), \qquad \forall t \geq t_0 \geq 0 \\ \mu(t) \leq& \delta\left(\|\tilde{\boldsymbol{x}}\|_{[t_0,t]}; \mu(t_0)\right) \end{aligned}$$

where β, γ and δ have the same properties as in (2.8)*. When $t_0 = 0$, this can be reduced to*

$$|\boldsymbol{x}(t)| \leq \beta\left(|\boldsymbol{x}(0)|, t\right) + \gamma\left(\|\boldsymbol{w}\|_{[0,t]}\right), \qquad \forall t \geq 0$$

where β and γ have the same properties as in (2.7)*.*

An illustrative simulation of the proposed controller is given in Figure 2.2. The proofs of Proposition 2.4.1 and Theorem 2.4.2 will follow the statements of the technical lemmas below. The proofs of the technical lemmas are deferred to §2.9.

Lemma 2.4.3. *Assume that for some time step k' we have $mode(k'+1) = update$ and $p(k') = 0$ (i.e. an update measurement sequence starts at $k'+1$). If $\forall k \in \{k'+1 \ldots k'+P+1\}$, $mode(k) \neq capture$ (i.e. by time step $k'+P$ the controller has not switched to the capture mode) then $\|\mu\|_{\{k'-r+1+P \ldots k'+P\}} \leq \sigma \|\mu\|_{\{k'-r+1 \ldots k'\}}$.*

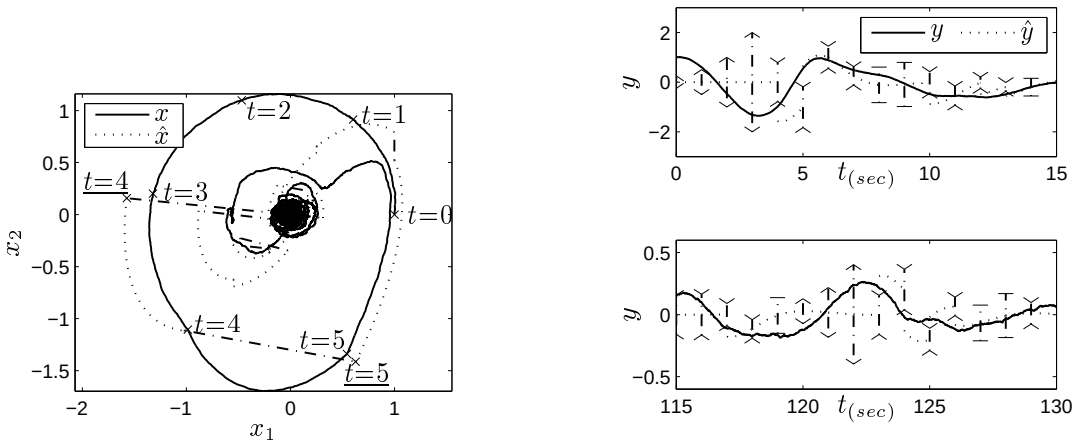

Figure 2.2: Simulation of the proposed controller. Simulated here is a two-dimensional dynamical system: $\dot{\boldsymbol{x}}(t) = [0.1, -1; 1, 0.1]\,\boldsymbol{x}(t) + [0; 1]\,\boldsymbol{u}(t) + [1, 0; 0, 1]\,\boldsymbol{w}(t)$, where only the first dimension is observed, $\boldsymbol{y}(t) = [1, 0]\,\boldsymbol{x}\,(t)$, through a quantizer with $N = 3$. The solid line in the left plot is the trajectory of the system (starting at $\boldsymbol{x}\,(0) = [1; 0]$). The dotted line in that plot is the state estimate. The dash-dotted lines represent the jumps in the state estimate after a new measurement is received. The locations of the trajectory and the state estimate at the first few sampling times are marked by $\times$. The underlined time indications correspond to the state estimate. The two plots on the right show time segments of the measured output ($T_s = 1s$). The solid line is the unquantized output ($\boldsymbol{y}$) of the system and the dotted line is its estimate. The vertical dash-dotted lines depict the single bounded quantization region. The controller is in the *capture* mode where these vertical lines are bounded by arrows facing outward, in the *update* mode where the arrows are facing inward, and in the *detect* mode where the vertical lines are bounded by small horizontal lines. Looking at both the left plot and the top right plot, one can observe the initial transient of the system: At $t = 3$ a sufficient number (two) of unsaturated measurements were collected and the controller switches to the *update* mode; this causes the state estimate to jump at $t = 4$ from the origin to $\approx [-1.6; 0.2]$; and at $t = 5$ the state estimate jumps even closer to the true state. Looking at the bottom right plot, one can observe the steady-state behavior of the simulation, where an escape of the trajectory due to disturbance is detected at $t = 119s$, and then the trajectory is recaptured at $t = 122s$. The design parameters were: $P = 6$, $\mu\,(0) = 0.25$, $\Omega_{out} = 2$, $\alpha = 0.02$, $s = 0.05$, $K = [0.6, -1.5]$. The disturbance followed the zero-mean normal distribution with standard deviation of 0.2.

Lemma 2.4.4. *There exist constants $\zeta_D > 0$ and $\zeta_\mu > 0$ with the following properties: If for some time step k' we have $mode(k'+1) = update$ and $p(k') = 0$, and the input is such that*

$$\|\mu\|_{\{k'-r+1\ldots k'\}} > \frac{1}{\alpha}\zeta_D \left\|\boldsymbol{w}^d\right\|_{\{k'-r+1,k'+P-2\}}, \tag{2.23}$$

then $mode(m) = update$ $\forall m \in \{k'+2\ldots k'+P-r\}$, $mode(m) = detect$ $\forall m \in \{k'+P-r+1\ldots k'+P\}$, $mode(k'+P+1) = update$, and

$$\|\tilde{\boldsymbol{x}}\|_{\{k',\ldots,k'+P-1\}} \leq \zeta_\mu \|\mu\|_{\{k'-r+1\ldots k'\}}. \tag{2.24}$$

Lemma 2.4.5. *Assume that for some time step k' we have $mode(k'+1) = update$ and $p(k') = 0$. Let $k_2 = \min\{k'+P, \min\{k \,|mode(k+1) = capture, k > k'\}\}$. There exists a constant $\zeta_w > 0$ such that if the disturbance does not satisfy* (2.23)*, then*

$$\|\tilde{\boldsymbol{x}}\|_{\{k'\ldots k_2-1\}} \leq \zeta_w \left\|\boldsymbol{w}^d\right\|_{\{k'-r+1\ldots k'+P-2\}}.$$

Lemma 2.4.6. *There exist functions $\tilde{\delta}_1(\nu;\rho) : \mathbb{R}^2_{>0} \to \mathbb{R}_{>0}$ and $T_1^*(\nu;\rho) : \mathbb{R}^2_{>0} \to \mathbb{R}_{>0}$, each nondecreasing in ν when ρ is fixed, and constants $\zeta_C > 0$ and $\zeta_b > 0$, with the following properties: For any time step k_2 such that $mode(k_2+1) = capture$ there exists $k_3 > k_2$ such that $k_3 < k_2 + T_1^*\left(|\tilde{x}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_2}\right)$, $mode(k_3+1) = update$, $p(k_3) = 0$, $\|\tilde{\boldsymbol{x}}\|_{\{k_2\ldots k_3\}} \leq \tilde{\delta}_1\left(|\tilde{x}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_2}\right)$ and $\|\mu\|_{k_3-r+1\ldots k_3} \leq \mu_{k_2}\Omega_{out}^{T_1^*(\nu;\rho)}$; the functions $\tilde{\delta}_1$ and T_1^* satisfy $\tilde{\delta}_1(\nu;\rho) \leq \rho\zeta_b\Omega_{out}^{T_1^*(\nu;\rho)}$ $\forall \nu, \rho$.*

Lemma 2.4.7. *Let k_2 be an arbitrary time step. There exist a constant $\zeta_s > 0$, a class $\mathcal{K}$ function ε, and functions $\tilde{\delta}_2(\nu;\rho) : \mathbb{R}^2_{>0} \to \mathbb{R}_{>0}$ and $T_2^*(\nu;\rho) : \mathbb{R}^2_{>0} \to \mathbb{R}_{>0}$ with the following properties: If $|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\| \leq \varepsilon(\mu_{k_2})$ then $k_3 \doteq k_2 + T_2^*\left(|\tilde{x}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_2}\right)$ satisfies $mode(k_3+1) = update$, $p(k_3) = 0$ and $\|\tilde{\boldsymbol{x}}\|_{\{k_2,\ldots,k_3\}} \leq \tilde{\delta}_2\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_2}\right)$, $\|\mu\|_{k_3-r+1\ldots k_3} \leq \mu_{k_2}\zeta_s\sigma^{T_2^*/P}$; when ρ is fixed the function $\tilde{\delta}_2(\cdot;\rho)$ is of class $\mathcal{K}_\infty$; the functions $\tilde{\delta}_2$ and T_2^* satisfy $\tilde{\delta}_2(\nu;\rho) \leq \rho\zeta_s\sigma^{T_2^*(\nu;\rho)/P}/\|C\|$ $\forall \nu, \rho$.*

Proof of Proposition 2.4.1: Assume first that t_0 is at a sampling time and let k_0 be such that $k_0 T_s = t_0$. We say that time step k has the *SS properties* if $mode(k+1) = update$, $p(k) = 0$ and (2.23) does not hold with $k' = k$. The proof will proceed in four steps: in the first step we will derive a bound on the trajectory from k_0 to k_1 for some k_1 that has the SS properties; in the second step we will derive a bound on the trajectory from k_1 to infinity; in the third step we will combine these two bounds and derive the ISS bound on the estimation error; in the fourth step we will derive the bound on the zoom factor.

1. Assume first that $mode(k_0) = capture$. Define k_1 to be the first time step after k_0 with the SS prop-

erties. If such a time step does not exist, define $k_1 \doteq \infty$. By Lemma 2.4.6 we know there exists $k^* \leq k_0 + T_1^* \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right)$ such that $\|\tilde{\boldsymbol{x}}\|_{k_0 \ldots k^*} \leq \tilde{\delta}_1 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right)$. Together with Lemmas 2.4.3 and 2.4.4 we also have that if $k_1 > k^*$ then $|\tilde{\boldsymbol{x}}_k| \leq \zeta_\mu \mu_{k_0} \Omega_{out}^{T_1^*} \sigma^{\left\lfloor \frac{k-k^*}{P} \right\rfloor} \leq \zeta_\mu \mu_{k_0} \Omega_{out}^{T_1^*} \sigma^{\left\lfloor \frac{k-T_1^*}{P} \right\rfloor}$ $\forall k \in \{k^* \ldots k_1\}$. As Lemma 2.4.6 also states $\tilde{\delta}_1 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right) \leq \mu_{k_0} \zeta_b \Omega_{out}^{T_1^*}$, we can finally derive $|\tilde{\boldsymbol{x}}_k| \leq \tilde{\beta}_c \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|, k - k_0; \mu_{k_0}\right)$ $\forall k \in \{k_0 \ldots k_1\}$ where

$$\tilde{\beta}_c(\nu, k; \rho) \doteq \min \left\{ \tilde{\delta}_1(\nu; \rho), \rho \left(\frac{\Omega_{out}}{\sigma^{1/P}} \right)^{T_1^*(\nu;\rho)} \sigma^{\frac{k}{P}-1} \max\{\zeta_\mu, \zeta_b\} \right\}. \tag{2.25}$$

If $mode(k_0) \neq capture$ then there is a time step k_1', $k_0 - P < k_1' \leq k_0$, such that $mode(k_1'+1) = update$ and $p(k_1') = 0$. If in addition (2.23) does not hold with $k' = k_1'$, then we define $k_1 = k_1'$, and thus we have, vacuously, $|\tilde{\boldsymbol{x}}_k| \leq 0$ $\forall k \in \{k_0 \ldots k_1\}$. If (2.23) does hold with $k' = k_1'$, then with k_1 defined as the first time step after k_0 with the SS properties, we can write: $|\tilde{\boldsymbol{x}}_k| \leq \zeta_\mu \mu_{k_0} \sigma^{\left\lceil \frac{k-k_0}{P} \right\rceil}$ $\forall k \in \{k_0 \ldots k_1\}$. Taking into consideration that $mode(k_0) = capture$ only if $\mu_{k_0} \geq s$, we can now write $|\tilde{\boldsymbol{x}}_k| \leq \tilde{\beta}_1 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|, k - k_0; \mu_{k_0}\right)$ $\forall k \in \{k_0 \ldots k_1\}$ where

$$\tilde{\beta}_1(\nu, k; \rho) \doteq \begin{cases} \max\left\{ \tilde{\beta}_c(\nu, k; \rho), \zeta_\mu \rho \sigma^{\left\lceil \frac{k-k_0}{P} \right\rceil} \right\} & \rho \geq s \\ \zeta_\mu \rho \sigma^{\left\lceil \frac{k-k_0}{P} \right\rceil} & \rho < s. \end{cases} \tag{2.26}$$

Assume now that $|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\| \leq \varepsilon(\mu_{k_0})$ where $\varepsilon(\cdot)$ comes from Lemma 2.4.7. By that lemma there exists $T_2^* = T_2^* \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right)$ such that $\|\tilde{\boldsymbol{x}}\|_{k_0 \ldots T_2^*} \leq \tilde{\delta}_2 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right)$. Also from Lemma 2.4.7, together with Lemmas 2.4.3 and 2.4.4, if $k_1 > k^*$ then $|\tilde{\boldsymbol{x}}_k| \leq \zeta_\mu \mu_{k_0} \zeta_s \sigma^{T_2^*/P} \sigma^{\left\lfloor \frac{k-T_2^*}{P} \right\rfloor}$ $\forall k \in \{T_2^* \ldots k_1\}$. As we are also given from Lemma 2.4.7 that $\tilde{\delta}_2 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_0}\right) \leq \mu_{k_0} \zeta_s \sigma^{T_2^*/P} / \|C\|$, we can finally derive $\forall k \in \{k_0 \ldots k_1\}$: $|\tilde{\boldsymbol{x}}_k| \leq \tilde{\beta}_2 \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|, k - k_0; \mu_{k_0}\right)$ where

$$\tilde{\beta}_2(\nu, k; \rho) \doteq \min \left\{ \tilde{\delta}_2(\nu; \rho), \rho \zeta_s \sigma^{\lfloor k/P \rfloor} \max\{\zeta_\mu, 1/\|C\|\} \right\}. \tag{2.27}$$

For fixed ν and ρ, both $\lim_{k\to\infty} \tilde{\beta}_1(\nu, k; \rho) = 0$ and $\lim_{k\to\infty} \tilde{\beta}_2(\nu, k; \rho) = 0$. Also, for fixed k and ρ, both $\tilde{\beta}_1(\nu, k; \rho)$ and $\tilde{\beta}_2(\nu, k; \rho)$ are continuous and nondecreasing with respect to ν. However, only $\tilde{\beta}_2$ satisfies $\tilde{\beta}_2(0, k; \rho) = 0$ $\forall k \forall \rho$, and $\tilde{\beta}_2$ is a valid bound on the trajectory only when $|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\| \leq \varepsilon(\mu_{k_0})$. Nevertheless, it is possible to construct a class $\overline{\mathcal{KL}}$ function, $\hat{\beta}(\nu, k; \rho)$, such that $\hat{\beta}(\nu, k; \rho) \geq \tilde{\beta}_2(\nu, k; \rho)$ when $\nu \leq \varepsilon(\rho)$ and $\hat{\beta}(\nu, k; \rho) \geq \tilde{\beta}_1(\nu, k; \rho)$ otherwise. With this $\hat{\beta}(\nu, k; \rho)$ we can write $|\tilde{\boldsymbol{x}}_k| \leq \hat{\beta} \left(|\tilde{\boldsymbol{x}}_{k_0}| + \zeta_C \left\|\boldsymbol{w}^d\right\|, k - k_0; \mu_{k_0}\right)$ $\forall k \in \{k_0 \ldots k_1\}$.

Note that all the functions mentioned above are continuous in ν and ρ, $\forall \nu \in \mathbb{R}_{\geq 0}$ and $\forall \rho \in \mathbb{R}_{>0}$. They are not, however, all continuous at $\rho = 0$ (or even defined) since $\lim_{\rho \searrow 0} T_1^* (\nu; \rho) = \infty$ for every $\nu > 0$. Nevertheless, $\hat{\beta} (\nu, k; \rho)$ is continuous at $\rho = 0$. This is due to ε being of class $\mathcal{K}$, which implies that for sufficiently small ρ, $\hat{\beta} (\nu, k; \rho) = \tilde{\beta}_1 (\nu, k; \rho) = \zeta_\mu \rho \sigma \lceil (k - k_0) / P \rceil$.

2. Let k_2 be the first time step after k_1 such that $mode(k_2) = detect$ and $mode(k_2 + 1) = capture$. If such a time step does not exist, we set $k_2 = \infty$. From Lemma 2.4.5 we have that $\|\tilde{\boldsymbol{x}}\|_{\{k_1 \dots k_2\}} \leq \zeta_w \left\|\boldsymbol{w}^d\right\|$. Let k_4 be the first time step after k_2 such that $mode(k_4 + 1) = update$, $p(k_4) = 0$ and (2.23) does not hold with $k' = k_4$. Replacing k_0 with k_2 in the previous step, we can write

$$\|\tilde{\boldsymbol{x}}\|_{\{k_2 \dots k_4\}} \leq \hat{\beta} \left(|x_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|, k - k_2; \mu_{k_2} \right) \leq \hat{\beta} \left((\zeta_w + \zeta_C) \left\|\boldsymbol{w}^d\right\|, 0; s \right) \doteq \tilde{\gamma} \left(\left\|\boldsymbol{w}^d\right\| \right).$$

Since k_4 also satisfies the SS properties as does k_1, we can repeat these arguments for future time steps and arrive at $\|\tilde{\boldsymbol{x}}\|_{\{k_1 \dots \infty\}} \leq \hat{\gamma} \left(\left\|\boldsymbol{w}^d\right\| \right)$, where $\hat{\gamma} (\nu) \doteq \max \{\zeta_w \nu, \tilde{\gamma} (\nu)\}$. Note that $\hat{\gamma} (\cdot)$ is of class $\mathcal{K}_\infty$.

3. Combining the last two steps, we can derive the first condition for the parametrized ISS property at the discrete times: for all $k \in \{0 \dots \infty\}$,

$$|\tilde{\boldsymbol{x}}_k| \leq \beta_e (|\tilde{\boldsymbol{x}}_{k_0}|, k; \mu_{k_0}) + \gamma_e \left(\left\|\boldsymbol{w}^d\right\|; \mu_{k_0} \right)$$

where $\beta_e (\nu, k; \mu) \doteq \hat{\beta} (2\nu, k; \mu)$ and $\gamma_e (\nu; \mu) \doteq \hat{\beta} (2\zeta_C \nu, 0; \mu) + \hat{\gamma} (\nu)$. Note that indeed β_e and γ_e of class $\overline{\mathcal{KL}}$ and $\overline{\mathcal{K}}_\infty$, respectively.

The extension from the discrete analysis to continuous time, with the estimation error defined as $\tilde{\boldsymbol{x}}(t) \doteq \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ for every $t \geq t_0$, can be proved along the lines of [48, Theorem 6]. This proves the first line of (2.22).

4. To construct the bound on μ we consider the three phases of the trajectory: initial *capture* sequence, *zoom-in* sequences and subsequent *capture* sequences. If $mode(k_0) = capture$ we start with μ_{k_0} and we grow the zoom factor until for r successive time steps we have $(N - 2) \mu_k > |\tilde{\boldsymbol{y}}_k|$. Thus at the initial *capture* sequence we have

$$\|\mu\| \leq \Omega_{out}^r \max \{\mu_{k_0}, \|C\| \|\tilde{\boldsymbol{x}}\| / (N - 2)\}. \tag{2.28}$$

At a *zoom-in* sequence we may initially enlarge μ by a factor of $\|\mu'\|$ with μ' defined according to (2.18). However, after this possible initial enlargement, μ is decreased by a factor of σ every P steps.

At subsequent *capture* sequences we start with $\mu_k = s$ and enlarge it again until for r successive time steps we have $(N-2)\,\mu_k > |\tilde{\boldsymbol{y}}_k|$. Combining all these observations, we can set γ_μ from (2.22) as

$$\gamma_\mu(\nu, \mu_0) \doteq \|\mu'\|\, \Omega_{out}^r \max\{\mu_0, s, \|C\|\, \|\tilde{\boldsymbol{x}}\| / (N-2)\}.$$

■

Proof of Theorem 2.4.2: With $A + BK$ being Hurwitz, the stabilizing control law, $u = K\hat{x}$, renders the closed-loop system

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + B\boldsymbol{u} + D\boldsymbol{w} = (A + BK)\,\boldsymbol{x} + BK\tilde{\boldsymbol{x}} + D\boldsymbol{w} \tag{2.29}$$

ISS with respect to the disturbance and the estimation error. Combining this ISS property with the ISS property proved in Proposition 2.4.1, and applying a cascade argument similar to what was used to prove [63, Proposition 7.2], we can conclude that the closed-loop system is ISS with respect to the disturbance. ■

2.4.3 Modeling Errors

We represent modeling errors as $A(t) = A_0 + \Delta A(t)$ with only A_0 known and $\Delta A(t) \not\equiv 0$. It is assumed, though, that $\|\Delta A(t)\| \leq \delta_A$ for some $\delta_A \in \mathbb{R}_{\geq 0}$ and $\forall t \geq 0$. To deal with such modeling errors the only change needed in the design is in the stabilizing control law, where K will be chosen such there exist two positive definite symmetric matrices, P and Q, for which the following holds:

$$P(A_0 + \Delta A + BK) + (A_0 + \Delta A + BK)^T P + Q < 0, \qquad \|\Delta A(t)\| < \delta_A. \tag{2.30}$$

It is well-known and easy to show using a Lyapunov argument that if (2.30) holds then the system (2.29) has the ISS stability property with respect to the estimation error and disturbance:

$$|\boldsymbol{x}(t)| \leq \beta_x(\boldsymbol{x}(t_0), t - t_0) + \gamma_{x,e}\left(\|\tilde{\boldsymbol{x}}\|_{[t_0,t]}\right) + \gamma_{x,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right), \quad \forall t > t_0 > 0 \tag{2.31}$$

where β_x is of class $\mathcal{KL}$ and $\gamma_{x,w}$ and $\gamma_{w,x}$ are of class $\mathcal{K}_\infty$. Such a stabilizing gain matrix K can be found by using linear matrix inequality (LMI) techniques.

With this stabilizing control law, we derive our second stability result:

Theorem 2.4.8. *Assume the observer has the convergence property and the stabilizing control law is chosen so that* (2.30) *holds for some* $\delta_A > 0$. *Then the closed-loop system has the local practical ISS property* (2.9) *for some* $\delta_{A,\max} > 0$, $x_{max} > 0$ *and* $w_{max} > 0$.

Proof: The estimation error now follows the following dynamics between sampling times:

$$\dot{\tilde{\boldsymbol{x}}} = A\tilde{\boldsymbol{x}} - \Delta A\boldsymbol{x} - D\boldsymbol{w}. \tag{2.32}$$

Therefore its evolution is no longer independent of the state of the system. The proposed controller in this case will render the estimation error parameterized-ISS with respect to both the disturbance and the system's state:

$$\begin{aligned} &|\tilde{\boldsymbol{x}}(t)| \leq \beta_e\left(\tilde{\boldsymbol{x}}(t_0), t-t_0; \mu(t_0)\right) + \gamma_{e,x}\left(\delta_A \|\boldsymbol{x}\|_{[t_0,t]}; \mu(t_0)\right) + \gamma_{e,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu(t_0)\right), \\ &\mu(t) \leq \gamma_\mu\left(\|\tilde{\boldsymbol{x}}\|_{[t_0,t]}, \mu(t_0)\right), \qquad \forall t \geq t_0 \geq 0. \end{aligned}$$

Due to the interleaved dependency of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ on each other we can no longer apply the cascade theorem. However, since $\boldsymbol{x}_1$ which follows (2.1) is continuous, we can now apply a variation of the small-gain theorem, Theorem 2.8.1, and arrive at the result stated in the theorem. ■

Note that because for every fixed μ, $\gamma_{e,x}(r,\mu)$ grows faster than any linear function of r both at $r=0$ and at $r=\infty$, we cannot choose $r_0 = 0$ or $r_1 = \infty$ in the proof of Theorem 2.8.1. These super-linear gains are not an artifact of our design. Recently Martins [37] showed, using techniques from information theory, that it is impossible to achieve ISS with linear gain for any linear system with finite data rate feedback.

2.5 Approaching the Minimal Data Rate

Several papers ([50],[22],[70],[45],[39]) present the same lower bound on the data rate necessary to stabilize a given system. This bound, in terms of the bit-rate (R) to be transmitted, is

$$R > R_{min} \doteq \frac{\sum_{|\eta_j|\geq 1} \log_2 |\eta_j|}{T_s} \tag{2.33}$$

where the η_j's are the eigenvalues of the discrete open-loop matrix $\Phi \doteq \exp(AT_s)$. Note that the bound (2.33) is independent of the disturbance characteristics and is applicable to systems with no disturbances.

Lemma 2.5.1. *To achieve ISS in the state feedback case, it is necessary and sufficient for the data rate to satisfy* (2.33).

Proof (sketch): Since (2.33) is necessary for asymptotic stability in the disturbance-free case, it is also necessary to achieve disturbance rejection in the ISS sense (which reduces to asymptotic stability when the disturbance is zero). Now for the sufficiency. Consider the scalar unstable case, $n_x = 1$, $A \equiv a > 0$. In this

case the minimum data rate is $R_{min} = \log_2 \exp(T_s a)/T_s$. The data rate of our scheme is $\log_2(N)/T_s$ where we require N to be an odd integer which, together with P and α, satisfies the convergence property. Note that for this simple case, in the limit as $\alpha \searrow 0$ the convergence property becomes

$$\exp(T_s a)^P / \left(N^{P-r}(N-2)^r\right) < 1. \tag{2.34}$$

As the above is equivalent to $(\exp(T_s a)/N)^P < (N-2)^r/N^r$, we can see that for any $N > \exp(T_s a)$ we can find P large enough to satisfy (2.34). Because of the continuous dependence of the convergence property on α, if (2.34) is satisfied then there exists $\alpha > 0$ which satisfies the convergence property. Thus to be able to find design parameters that satisfy the convergence property, we only need $N > \exp(T_s a)$. To deal with the constraint that N is an integer, we can use a different number of quantization regions at each sampling time. This makes our data rate $\log_2\left(\tilde{N}\right)/T_s$ with $\tilde{N}$ being the average number of quantization regions per sampling time. As $\tilde{N}$ does not have to be integer, we can have $\tilde{N}$ approach $\exp(T_s a)$ and make our data rate arbitrarily close to the minimum data rate.

Extension to the multidimensional case with distinct real eigenvalues is trivial if we allocate a different number of quantization regions for each unstable mode of the system. Extension to systems with imaginary eigenvalues and to non-diagonalizable systems can be made along the lines of [70]. ■

2.6 Extension to Nonlinear Systems

The crucial properties of linear systems which are used in the proof of Theorem 2.4.2 are (a) that the continuous, unquantized, closed-loop system is ISS with respect to the estimation error and the disturbance, and (b) that the update law for the estimated state between the sampling times (2.14) is such that the estimation error grows between these sampling times according to

$$\lim_{t \nearrow T_s} \|\tilde{\boldsymbol{x}}(kT_s + t)\| \leq \lambda_e \|\tilde{\boldsymbol{x}}(kT_s)\| + \lambda_w \|\boldsymbol{w}\|_{[kT_s,(k+1)T_s]} + \lambda_x \|\boldsymbol{x}\|_{[kT_s,(k+1)T_s]} \tag{2.35}$$

where λ_e, λ_w and λ_x are known constants. For linear systems these constants are $\lambda_e = \|A_d\|$, $\lambda_w = \left\|\int_0^{T_s} \exp(A_0(T_s - t))\, D dt\right\|$ and $\lambda_x = \left\|\int_0^{T_s} \exp(A_0(T_s - t))\, dt\right\| \delta_A$, which follows easily from (2.10). If (2.35) holds globally, $\lambda_x = 0$ (as in the case where the exact system model is known), and the number of quantization regions allows the observer to satisfy the convergence property, then the quantized closed-loop system will be parameterized ISS with respect to the disturbance. If $\lambda_x \neq 0$, as in the case where modeling errors exist, then an additional small-gain condition must be satisfied in order to achieve the local practical

parameterized ISS property.

Neither property is unique to linear systems and both can also be formulated for nonlinear systems. This leads to a better conceptualization of our results. Consider a nonlinear system

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{w}(t)) \tag{2.36}$$

with $\boldsymbol{y}(t) = \boldsymbol{x}(t)$ (state feedback). State feedback control laws that render unquantized systems ISS with respect to either external disturbances or measurement errors have been proposed for certain nonlinear systems; see for example the discussions in [31, 28] and the references therein. Designing state feedback control laws that render unquantized systems ISS with respect to *both* external disturbances and measurement errors is still considered an open problem. The two closest results, for systems in strict feedback from, appear in [18, §6.2.2] and [17].

Assume that (2.36) satisfies the Lipschitz property: There exist $l_x > 0$, $l_w > 0$, $L_x > 0$ and $L_w > 0$ such that

$$|f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}) - f(\hat{\boldsymbol{x}}, \boldsymbol{u}, 0)| \leq L_x|\boldsymbol{x} - \hat{\boldsymbol{x}}| + L_w|\boldsymbol{w}|, \qquad \forall|\boldsymbol{x}| < l_x, \forall|\hat{\boldsymbol{x}}| < l_x, \forall|\boldsymbol{w}| < l_w \tag{2.37}$$

holds. When the Lipschitz property holds globally, which is the case with linear systems, $l_x = l_w = \infty$. Assuming the exact system model is known, if we update our state estimate between sampling times according to $\dot{\hat{\boldsymbol{x}}} = f(\hat{\boldsymbol{x}}, \boldsymbol{u}, 0)$, then (2.35) holds with

$$\lambda_e \doteq e^{T_s L_x}, \qquad \lambda_w \doteq \int_0^{T_s} e^{(T_s - \tau)L_x} d\tau L_w. \tag{2.38}$$

To make the convergence property applicable to state feedback nonlinear systems, the only change needed is to redefine

$$F(\mu; k) \doteq \lambda_e \|\mu\|_{\{k-r,\ldots,k-1\}}. \tag{2.39}$$

A sufficient condition for the control system to have the convergence property remains $\lambda_e/N < 1$.

The above discussion leads to our third stability result:

Theorem 2.6.1. *Consider a state feedback nonlinear system:*

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{w}(t)), \qquad \boldsymbol{z}_k = Q(\boldsymbol{x}_k; c_k, \mu_k) \tag{2.40}$$

where f has the Lipschitz property (2.37)*, and for which there exists a static feedback $\boldsymbol{u} = k(\boldsymbol{x})$ which renders the dynamics $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}, k(\boldsymbol{x} + \boldsymbol{e}), \boldsymbol{w})$ ISS with respect to $\boldsymbol{e}$ and $\boldsymbol{w}$. If $e^{T_s L_x}/N < 1$ then there exists a*

choice of α and P with which the control system has the convergence property with $F(\mu; k)$ defined in (2.39). With this choice of α and P and a choice of $\Omega_{out} > e^{T_s L_x}$ and $s > 0$, the system will have the parameterized ISS property for some β and γ if it can be guaranteed that $\|\boldsymbol{x}\| < l_x$ and $\|\boldsymbol{w}\| < l_w$. For $|\boldsymbol{x}(0)| < x_{max}$ and $\|\boldsymbol{w}\| < w_{max}$ such that

$$\beta(x_{max}, 0; s) + \gamma(w_{max}; s) \leq l_x \qquad \text{and} \qquad w_{max} \leq l_w \tag{2.41}$$

this will be guaranteed and therefore the system will have the local practical parametrized ISS property. If the Lipschitz property holds globally, then the closed-loop system will have the parameterized ISS property.

A natural question would be what is the necessary number of quantizations regions needed to achieve ISS for a given bound on $|\boldsymbol{x}(0)|$ and $\|\boldsymbol{w}\|$. Unfortunately, the theorem does not give a direct answer to this question. Nevertheless, we can say the following: Given x_{max}, $l_w = w_{max}$, l_x, $\lambda_e = e^{T_s L_x}$, such that both (2.37) and

$$\beta_x(x_{max}, 0) + \gamma_{x,e}\left(\max\left\{\lambda_e\left(x_{max} + \frac{w_{max}}{\lambda_e - 1}\right), \frac{\lambda_e^3}{\lambda_e - 1} w_{max}\right\}\right) + \gamma_{x,w}(w_{max}) < l_x$$

hold, where β_x, $\gamma_{x,e}$ and $\gamma_{x,w}$ are the ISS gains of the state feedback control law, there exist appropriate design parameters P, Ω_{out}, α, N and s with which the closed-loop system will have the local practical parametrized ISS property.

The proof of Theorem 2.6.1 follows the same lines as the proof of Theorem 2.4.2 and it is therefore omitted. See also [34] for a similar result but without disturbances.

2.7 Extension to Time Delays

Incorporating time delays, we assume for every $k \in \mathbb{Z}_{\geq 0}$ the controller receives the information $\boldsymbol{z}_k = \boldsymbol{z}(kT_s)$ only at time $kT_s + \delta_k$ where $\delta_k \in [0, T_s)$ is the delay. The delay is unknown to the controller and it does not need to be fixed. We set $\delta_{\max} \doteq \sup_{k \geq 0} \delta_k$. To simplify the derivation, we assume that there is no external disturbance ($\boldsymbol{w} \equiv 0$), and that there are no modeling errors ($A \equiv A_0$). With these settings we established the following result:

Theorem 2.7.1. *Given an implementation of the controller above with any valid choice for the design parameters such that the control system has the convergence property, the closed loop system will have the following semiglobal stability property: For every $x_{\max} \geq 0$, there exists a sufficiently small but strictly*

positive $\bar{\bar{\delta}}_{\max}$ *such that if* $\delta_{\max} \leq \bar{\bar{\delta}}_{\max}$ *then the following bound,* $\forall t \geq 0$*:*

$$|\boldsymbol{x}(t)| \leq \beta(|\boldsymbol{x}(0)|, t) + \gamma(\delta_{\max}) \tag{2.42}$$

holds whenever $|\boldsymbol{x}(0)| \leq x_{\max}$*, where the function* β *is of class* $\mathcal{KL}$ *and* γ *is of class* $\mathcal{K}$*.*

Remark 2.7.1. Known results on delays, [33] for example, provide what can be interpreted as a more general result than (2.42), in which the time 0 is replaced with t_0 and the bound holds for arbitrary t_0. In fact, an intermediate step in proving Theorem 2.7.1 (see (2.57) below) does provide a similar result which holds for arbitrary t_0. However, results for systems with delays which hold for arbitrary t_0 require knowledge of a history of the state over some nonzero time interval. By constraining ourselves to $t_0 = 0$ we are able to get a bound which only depends on the state at this time instance.

We start by giving a brief overview of the proof of Theorem 2.7.1. In addition to the state signal, $\boldsymbol{x}(t)$, we define a state estimation error signal, $\tilde{\boldsymbol{x}}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t-\delta)$ (the explicit dependence of δ on t will be provided in the proof itself). We also define two additional signals, $\boldsymbol{\theta}_x(t) = \boldsymbol{x}(t-\delta) - \boldsymbol{x}(t)$ and $\boldsymbol{\theta}_e(t) = \tilde{\boldsymbol{x}}(t-\delta) - \tilde{\boldsymbol{x}}(t)$. We use a small-gain argument between $\boldsymbol{x}$ and $\boldsymbol{\theta}_x$ in Lemma 2.7.3 to show that for a sufficiently small delay, there exists an ISS relation between the state estimation error signal (as the only input) and the state signal. We establish that the two signals $\boldsymbol{\theta}_x(t)$ and $\boldsymbol{\theta}_e(t)$ enter the system as external disturbances, and recall in Corollary 2.7.4 our previous result that the state estimation error signal possesses the ISS property with respect to external disturbances. We then use a small-gain argument between $\tilde{\boldsymbol{x}}$ and $\boldsymbol{\theta}_e$ in Lemma 2.7.6 to show that for a sufficiently small delay, there exists a local ISS relation between the state signal (as the only input) and the state estimation error signal. Finally, in the proof of Theorem 2.7.1 we use another small-gain argument between these two established ISS relations to derive the desired result.

We adopt the following notation from [71]: $\boldsymbol{x}_d(t) \doteq \|\boldsymbol{x}\|_{[t-\Delta,t]}$ and $\tilde{\boldsymbol{x}}_d(t) \doteq \|\tilde{\boldsymbol{x}}\|_{[t-\Delta,t]}$ where

$$\Delta \doteq 2T_s + \delta_{\max}.$$

The proof of Theorem 2.7.1 will come after the following intermediate results. The proofs of the intermediate results are deferred to §2.9

Lemma 2.7.2. *Let a system with state* $\boldsymbol{x}$ *satisfy the following relation,* $\forall t \geq t_0 \geq \Delta$*:*

$$|\boldsymbol{x}(t)| \leq \beta_x(|\boldsymbol{x}(t_0)|, t - t_0) + \gamma_x\Big(\|\boldsymbol{x}_d\|_{[t_0,t]}\Big) + \gamma_w\Big(\|\boldsymbol{w}\|_{[t_0,t]}\Big) \tag{2.43}$$

where $\beta_x \in \mathcal{KL}$, and $\gamma_x, \gamma_w \in \mathcal{K}_\infty$. If $\gamma_x(r) < \lambda r$ for some $\lambda < 1$, then for every function $\gamma \in \mathcal{K}_\infty$ such that

$$\gamma(\nu) \geq \left(1 + \sqrt{\frac{\lambda}{1-\lambda}}\right)\left(1 + \lambda\left(1 + \sqrt{\frac{\lambda}{1-\lambda}}\right)\right)\gamma_w(\nu) \tag{2.44}$$

there exists a function $\beta \in \mathcal{KL}$ such that $\forall t \geq t_0 \geq \Delta$:

$$|\boldsymbol{x}_d(t)| \leq \beta\left(|\boldsymbol{x}_d(t_0)|, t - t_0\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right). \tag{2.45}$$

Define $k(t) \doteq \max\{k \in \mathbb{Z}_{\geq 0} \,|kT_s + \delta_k \leq t\}$, the index of the last sampling which arrived at the controller before time t. With this definition we can write

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + BK\left(\boldsymbol{x}\left(t - \delta_{k(t)}\right) + \tilde{\boldsymbol{x}}(t)\right) = (A + BK)\boldsymbol{x}(t) + BK\left(\boldsymbol{\theta}_x(t) + \tilde{\boldsymbol{x}}(t)\right) \tag{2.46}$$

where $\boldsymbol{\theta}_x(t) \doteq \boldsymbol{x}\left(t - \delta_{k(t)}\right) - \boldsymbol{x}(t)$ and $\tilde{\boldsymbol{x}}(t) \doteq \hat{\boldsymbol{x}}(t) - \boldsymbol{x}\left(t - \delta_{k(t)}\right)$.

Lemma 2.7.3. *There exists a sufficiently small, but strictly positive, $\bar{\delta}_{\max}$, such that if $\delta_{\max} \leq \bar{\delta}_{\max}$ then the following ISS relation, $\forall t \geq t_0 \geq \Delta$:*

$$|\boldsymbol{x}_d(t)| \leq \beta_x\left(|\boldsymbol{x}_d(t_0)|, t - t_0\right) + \gamma_x\left(\|\tilde{\boldsymbol{x}}_d\|_{[t_0,t]}\right) \tag{2.47}$$

holds where $\beta_x \in \mathcal{KL}$ and $\gamma_x \in \mathcal{K}_\infty$ is a linear function.

Define $\bar{k}(t) = \lfloor t/T_s \rfloor$. Another way to expand (2.46) is as follows, $\forall t \geq \delta_0$:

$$\begin{aligned}
\dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{u}(t) = A\boldsymbol{x}(t) + BK\hat{\boldsymbol{x}}(t) \\
&= A\boldsymbol{x}(t) + BK\left(\hat{\boldsymbol{x}}\left(t + \delta_{\bar{k}(t)}\right) + \hat{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}\left(t + \delta_{\bar{k}(t)}\right)\right) \\
&= A\boldsymbol{x}(t) + B\boldsymbol{u}\Big(t + \delta_{\bar{k}(t)}\Big) + BK\Big(\boldsymbol{\theta}_e\Big(t + \delta_{\bar{k}(t)}\Big) + \boldsymbol{\theta}_x(t)\Big)
\end{aligned} \tag{2.48}$$

where $\boldsymbol{\theta}_e(t) \doteq \tilde{\boldsymbol{x}}\left(t - \delta_{k(t)}\right) - \tilde{\boldsymbol{x}}(t)$. For $t \geq T_s + \delta_1$, $t \neq kT_s + \delta_k \ \forall k$, the state estimate evolves according to (2.14) and thus the estimation error, $\tilde{\boldsymbol{x}}$, evolves according to

$$\dot{\tilde{\boldsymbol{x}}}(t) = \dot{\hat{\boldsymbol{x}}}(t) - \dot{\boldsymbol{x}}\left(t - \delta_{k(t)}\right) = A\tilde{\boldsymbol{x}}(t) - BK\left(\boldsymbol{\theta}_e(t) + \boldsymbol{\theta}_x\left(t - \delta_{k(t)}\right)\right). \tag{2.49}$$

Denoting

$$\begin{aligned}\boldsymbol{w}(t) &\doteq -BK\left(\boldsymbol{\theta}_e(t)+\boldsymbol{\theta}_x\left(t-\delta_{k(t)}\right)\right),\\ \boldsymbol{w}_k^d &\doteq \int_{kT_s+\delta_k}^{(k+1)T_s+\delta_k} e^{A(k+1)T_s+\delta_k-t}\boldsymbol{w}(t)\,dt,\end{aligned}$$

we have that $\forall k \geq 1$:

$$\boldsymbol{c}_{k+1}-\boldsymbol{x}_{k+1}=\boldsymbol{c}\left(\left(k+1\right)T_s\right)-\boldsymbol{x}\left(\left(k+1\right)T_s\right)=e^{T_sA}\tilde{\boldsymbol{x}}\left(kT_s+\delta_k\right)+\boldsymbol{w}_k^d \doteq e^{T_sA}\tilde{\boldsymbol{x}}_k+\boldsymbol{w}_k^d \tag{2.50}$$

where $\boldsymbol{c}$ is the quantization parameter defining the center of the quantizer. In Proposition 2.4.1 we proved that if the system satisfies (2.50), then the following holds:

Corollary 2.7.4. *There exist functions $\beta_{e,d} \in \overline{\mathcal{KL}}$ and $\gamma_{e,d} \in \overline{\mathcal{K}}_\infty$ such that $\forall k \geq k_0 \geq 1$:*

$$\begin{aligned}|\tilde{\boldsymbol{x}}_k| \leq& \beta_{e,d}\left(\left|\tilde{\boldsymbol{x}}_{k_0}\right|, k-k_0; \mu_{k_0}\right)+\gamma_{e,d}\left(\left\|\boldsymbol{w}^d\right\|_{\{k_0,\ldots,k-1\}}; \mu_{k_0}\right)\\ \mu_k \leq& \psi\left(\left\|\tilde{\boldsymbol{x}}\right\|_{\{k_0,\ldots,k-1\}}; \mu_{k_0}\right).\end{aligned} \tag{2.51}$$

The function $\psi(\cdot,\cdot)$ as a function of its first argument when its second argument is fixed, is continuous, non-decreasing and non-negative. As a function of its second argument when its first argument is fixed, it is continuous.

Lemma 2.7.5. *The delayed estimation error, $\tilde{\boldsymbol{x}}_d$, satisfies the following relation, $\forall t \geq t_0 \geq \Delta$:*

$$\left|\tilde{\boldsymbol{x}}_d(t)\right| \leq \tilde{\beta}_e\left(\left|\tilde{\boldsymbol{x}}_d(t_0)\right|, t-t_0; \mu_{k(t_0)}\right)+\tilde{\gamma}_e\left(\delta_{\max}\left\|\tilde{\boldsymbol{x}}_d\right\|_{[t_0,t]}; \mu_{k(t_0)}\right)+\tilde{\gamma}_w\left(\delta_{\max}\left\|\boldsymbol{x}_d\right\|_{[t_0,t]}; \mu_{k(t_0)}\right) \tag{2.52}$$

where $\tilde{\beta}_e \in \overline{\mathcal{KL}}$ and $\tilde{\gamma}_e, \tilde{\gamma}_w \in \overline{\mathcal{K}}_\infty$.

Lemma 2.7.6. *For any $d'>0$, $x'_{\max}$, $\bar{x}_{\max}$ and $\mu_{\max}$ there exists a sufficiently small, but strictly positive, $\bar{\delta}_{\max}$, such that if $\delta_{\max} \leq \bar{\delta}_{\max}$ then the following ISS relation, $\forall t \geq t_0 \geq \Delta$:*

$$\left|\tilde{\boldsymbol{x}}_d(t)\right| \leq \beta_e\left(\left|\tilde{\boldsymbol{x}}_d(t_0)\right|, t-t_0\right)+\gamma_e\left(\delta_{\max}\left\|\boldsymbol{x}_d\right\|_{[t_0,t]}\right)+d' \tag{2.53}$$

where $\beta_e \in \mathcal{KL}$ and $\gamma_e \in \mathcal{K}$ holds for all $\forall\left|\tilde{\boldsymbol{x}}_d(\Delta)\right| \leq x'_{\max}$, $\forall\left\|\boldsymbol{x}_d\right\|_{[\Delta,\infty]} \leq \bar{x}_{\max}$, and $\forall \mu_{k(\Delta)} \leq \mu_{\max}$. Furthermore, $\forall \delta_{\max} \leq \bar{\delta}_{\max}$, we can write

$$\left\|\tilde{\boldsymbol{x}}_d\right\|_{[\Delta,t]} \leq \bar{\gamma}_1\left(\delta_{\max}\right)+\bar{\gamma}_e\left(\left\|\boldsymbol{x}_d\right\|_{[\Delta,t)}; \delta_{\max}\right) \qquad \forall t \geq \Delta \tag{2.54}$$

where

$$\lim_{\delta_{\max}\searrow 0} \bar{\gamma}_1(\delta_{\max}) = \max_{\mu\in[0,\mu_{\max}]} \tilde{\beta}_e(x'_{\max}, 0; \mu)$$
$$\lim_{\delta_{\max}\searrow 0} \bar{\gamma}_e(\cdot; \delta_{\max}) = 0. \tag{2.55}$$

Proof of Theorem 2.7.1: Let $x'_{\max}$, $\mu_{\max}$ be given and assume it can be guaranteed that $|\boldsymbol{x}_d(\Delta)| \leq x'_{\max}$, $|\tilde{\boldsymbol{x}}_d(\Delta)| \leq x'_{\max}$ and $\forall \mu_{k(\Delta)} \leq \mu_{\max}$. We start with $\bar{\delta}_{\max}$ such that (2.47) holds for all $\delta_{\max} < \bar{\delta}_{\max}$. We choose $\bar{x}_{\max}$ such that

$$\bar{x}_{\max} > \beta_x(x'_{\max}, 0) + \gamma_x\left(\sup_{\mu\in[0,\mu_{\max}]} \tilde{\beta}_e(x'_{\max}, 0; \mu)\right). \tag{2.56}$$

By decreasing $\bar{\delta}_{\max}$ if necessary, we can get that $\forall \delta_{\max} < \bar{\delta}_{\max}$, (2.54) holds $\forall \|\boldsymbol{x}_d\|_{[\Delta,\infty]} \leq \bar{x}_{\max}$. Combining (2.47) and (2.54) we can write, using the linearity of γ_x,

$$|\boldsymbol{x}(t)| \leq d + \gamma_x\left(\bar{\gamma}_e\left(\|\boldsymbol{x}\|_{[\Delta,t]}; \delta_{\max}\right)\right)$$

where

$$d \doteq \beta_x(x'_{\max}, 0) + \gamma_x(\bar{\gamma}_1(\delta_{\max})).$$

From (2.55) we see that by decreasing $\bar{\delta}_{\max}$ further if necessary we can get $\lambda < 1$ such that $\forall \delta_{\max} < \bar{\delta}_{\max}$,

$$\gamma_x(\bar{\gamma}_e(r; \delta_{\max})) < \lambda r, \qquad \forall r \in [r'_0, r'_1]$$

for any $0 < r'_0 < \bar{x}_{\max} < r'_1$ and

$$\frac{1}{1-\lambda} d < \bar{x}_{\max}.$$

Using Lemma 2.8.2, we can conclude that $\|\boldsymbol{x}_d\|_{[\Delta,\infty)} \leq \bar{x}_{\max}$. Finally we decrease $\bar{\delta}_{\max}$ even further if necessary so that $\forall \delta_{\max} < \bar{\delta}_{\max}$

$$\gamma_x(\gamma_e(\delta_{\max} r)) < r \qquad \forall r \in [r'_0, r'_1].$$

That, together with (2.47), (2.53), and Corollary 2.8.3 gives us

$$|\boldsymbol{x}_d(t)| \leq \beta'\left(\left|\begin{pmatrix} \boldsymbol{x}_d(t_0) \\ \tilde{\boldsymbol{x}}_d(t_0) \end{pmatrix}\right|, t - t_0\right) + d \qquad \forall t \geq t_0 \geq \Delta \tag{2.57}$$

where $\beta' \in \mathcal{KL}$. The last term above, d, is nonzero due to $d' > 0$ in (2.53) and $r'_0 > 0$ when $\delta_{\max} > 0$. However, we can make both d' and r'_0 arbitrarily small, and therefore also d, by taking a sufficiently small $\delta_{\max} > 0$. Thus we can replace d with $\gamma(\delta_{\max}) \in \mathcal{K}$.

We now bound the evolvement of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ from $t = 0$ to $t = \Delta$. Initially $\hat{\boldsymbol{x}} = 0$ and at the first sampling by our quantizer $|\hat{\boldsymbol{x}}(\delta_0)| < 2|\boldsymbol{x}(0)|$, leading to $\|\hat{\boldsymbol{x}}\|_{[0,T_s+\delta_1]} \leq e^{(T_s+\delta_{\max})\|A+BK\|} 2|\boldsymbol{x}(0)| \doteq \rho_1 |\boldsymbol{x}(0)|$. Thus

$$\|\boldsymbol{x}\|_{[0,T_s+\delta_1]} \leq e^{(T_s+\delta_{\max})\|A\|} |\boldsymbol{x}(0)| + (T_s + \delta_{\max}) e^{(T_s+\delta_{\max})\|A\|} \|BK\| \rho_1 |\boldsymbol{x}(0)| \doteq \rho_2 |\boldsymbol{x}(0)|.$$

Then $\left|\tilde{\boldsymbol{x}}^-(T_s + \delta_1)\right| \leq (\rho_1 + \rho_2)|\boldsymbol{x}(0)|$. Our quantizer has the property that $|\tilde{\boldsymbol{x}}(T_s + \delta_1)| \leq \left|\tilde{\boldsymbol{x}}^-(T_s + \delta_1)\right|$, so that $|\hat{\boldsymbol{x}}(T_s + \delta_1)| \leq (\rho_1 + 2\rho_2)|\boldsymbol{x}(0)|$. Repeating these arguments, we can derive the bound $|\boldsymbol{x}_d(\Delta)| \leq \rho|\boldsymbol{x}(0)|$ and $|\tilde{\boldsymbol{x}}_d(\Delta)| \leq \rho|\boldsymbol{x}(0)|$ for some $\rho > 0$.

Noting that $k(\Delta) = 2$, we can also bound $\mu_{k(\Delta)} \leq s\Omega_{out}^2$. To complete the proof, find $\bar{\delta}_{\max}$ such that (2.57) holds for $x'_{\max} = \rho x_{\max}$ and $\mu_{\max} = s\Omega_{out}^2$, and set

$$\beta(\nu; t) \doteq \beta'(\rho\nu, \max\{0, t - \Delta\}).$$

■

2.8 Small-Gain Theorem for Local Practical Parameterized ISS

The following is a modification of the small-gain theorem ([25, Theorem 2.1]). It states that the interconnection of an ISS system and a parameterized ISS system, under a small-gain condition, results in a local practical ISS system.

Theorem 2.8.1. *Consider two systems whose state variables, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, satisfy the ISS and the parameterized ISS properties, respectively:*

$$\begin{aligned}
|\boldsymbol{x}_1(t)| &\leq \beta_1(|\boldsymbol{x}_1(t_0)|, t - t_0) + \gamma_1\left(\|\boldsymbol{x}_2\|_{[t_0,t]}\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) \\
|\boldsymbol{x}_2(t)| &\leq \beta_2(|\boldsymbol{x}_2(t_0)|, t - t_0; \mu(t_0)) + \gamma_2\left(\delta\|\boldsymbol{x}_1\|_{[t_0,t]}; \mu(t_0)\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu(t_0)\right) \\
\mu(t) &\leq \gamma_\mu\left(\|\boldsymbol{x}_2\|_{[t_0,t]}, \mu(t_0)\right), \qquad \forall t \geq t_0 \geq 0.
\end{aligned} \tag{2.58}$$

Assume the first trajectory, $\boldsymbol{x}_1$, is continuous. Then the interconnected system will satisfy the local practical input-to-state stability property: $\exists \delta_{\max}, x_{\max}, w_{\max} \in \mathbb{R}_{>0}$ such that $\forall \delta \leq \delta_{\max}$, $\forall |\boldsymbol{x}_1(0)| < x_{\max}$, $\forall |\boldsymbol{x}_2(0)| <$

$x_{\max}$, $\forall \|\boldsymbol{w}\|_{[0,t]} < w_{\max}$,

$$|\boldsymbol{x}_1(t)| \leq \beta_{ic}\left(\left|\begin{pmatrix} \boldsymbol{x}_1(t_0) \\ \boldsymbol{x}_2(t_0) \end{pmatrix}\right|, t - t_0\right) + \gamma_{ic}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) + \lambda(\delta) \tag{2.59}$$

for all $t \geq t_0 \geq 0$ where β_{ic} is of class $\mathcal{KL}$ and γ_{ic} and λ are of class $\mathcal{K}_\infty$. The function s_∞ is continuous in all its variables and satisfies $s_\infty(0,0,\rho,0) = 0\ \forall \rho \geq 0$.

Remark 2.8.1. The proof below follows the first part of the proof of [25, Theorem 2.1] with necessary modifications. Two things make the setting of Theorem 2.8.1 different from the setting of [25, Theorem 2.1]. These are the additional state, μ, and the fact that the small-gain condition only holds locally. We need to show that our system can be written in the formulation of [25, Theorem 2.1]. We achieve this by showing that as long as the small-gain condition holds, the signal $\boldsymbol{x}_1$, and subsequently all the other signals, are bounded in the interior of the region in which the small-gain condition holds. Since the signal $\boldsymbol{x}_1$ is continuous, it cannot jump to where the small-gain condition does not hold, and we can conclude that the small-gain condition must hold indefinitely. Note that the second signal, $\boldsymbol{x}_2$, corresponding to our state estimation error, may not be continuous at the sampling times. In the state feedback case it can be shown that the norm of the estimation error only decreases at these instances of discontinuity, making it possible to use apply Lemma 2.8.2 on $\boldsymbol{x}_2$ in the proof of Theorem 2.8.1. In the output feedback case, however, the norm of the estimation error may in fact increase at these instances of discontinuity, making this argument not applicable (it is still applicable on $\boldsymbol{x}_1$ even in this case).

The proof of Theorem 2.8.1 will come after the following intermediate results.

Lemma 2.8.2. *Assume for some t_0 the signal $\boldsymbol{x}$ satisfies*

$$|\boldsymbol{x}(t)| \leq d + \gamma\left(\|\boldsymbol{x}\|_{[t_0,t]}\right), \qquad \forall t \geq t_0, \tag{2.60}$$

and

$$\lim_{\tau \nearrow t} |\boldsymbol{x}(\tau)| \geq |\boldsymbol{x}(t)|, \qquad \forall t \geq t_0 \tag{2.61}$$

(any discontinuity in the signal results in a decrease of the norm of the signal). Assume further that for some $r_2 > r_1 > 0$, $\lambda < 1$

$$\gamma(r) < \lambda r \qquad \forall r \in [r_1, r_2] \tag{2.62}$$

and

$$x_{\max} \doteq \max\left\{r_1, \frac{1}{1-\lambda}d\right\} < r_2. \tag{2.63}$$

Then given that $|\boldsymbol{x}(t_0)| \leq x_{\max}$*, we have*

$$\|\boldsymbol{x}\|_{[t_0,\infty)} \leq x_{\max}.$$

Proof: Assume on the contrary that there exists $t' \geq t_0$ such that $|\boldsymbol{x}(t')| > x_{\max}$. Choose $\varepsilon = x_{\max} + \min\left\{\|\boldsymbol{x}\|_{[t_0,\infty)} - x_{\max}, r_2 - x_{\max}\right\}/2$ so that $t = \inf\left\{\tau \geq t_0 \middle| |x(\tau)| \geq \varepsilon\right\}$ is well-defined. By definition of t, $\|\boldsymbol{x}\|_{[t_0,t)} \leq x_{\max} + \varepsilon$ and from $|\boldsymbol{x}(t_0)| \leq x_{\max}$ and (2.61), $t > t_0$ and $|\boldsymbol{x}(t)| = x_{\max} + \varepsilon$. Thus $\|\boldsymbol{x}\|_{[t_0,t]} = x_{\max} + \varepsilon < r_2$. From (2.60) and (2.62) we can now write $\|\boldsymbol{x}\|_{[t_0,t]} \leq d + \lambda\|\boldsymbol{x}\|_{[t_0,t]}$, and conclude using (2.63) that $\|\boldsymbol{x}\|_{[t_0,t]} \leq x_{\max}$. This contradicts $\|\boldsymbol{x}\|_{[t_0,t]} = x_{\max} + \varepsilon$. ■

A corollary of the small-gain theorem [25, Theorem 2.1] gives us the following local result:

Corollary 2.8.3. *Given* $\beta_1, \beta_2 \in \mathcal{KL}$*,* $\gamma_{1,x}, \gamma_{2,x} \in \mathcal{K}_\infty$*, and* $\rho < 1$*, there exists* $\beta \in \mathcal{KL}$ *and* $\gamma, \lambda_1, \lambda_2, \lambda_0 \in \mathcal{K}_\infty$ *with the following property. For every three signals* $\boldsymbol{x}_1$*,* $\boldsymbol{x}_2$*,* $\boldsymbol{w}$ *satisfying* $\forall t \geq t_0 \geq 0$

$$\begin{aligned}
|\boldsymbol{x}_1(t)| \leq& \beta_1\left(|\boldsymbol{x}_1(t_0)|, t - t_0\right) + \gamma_{1,x}\left(\|\boldsymbol{x}_2\|_{[t_0,t]}\right) + \gamma_{1,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) + d_1 \\
|\boldsymbol{x}_2(t)| \leq& \beta_2\left(|\boldsymbol{x}_2(t_0)|, t - t_0\right) + \gamma_{2,x}\left(\|\boldsymbol{x}_1\|_{[t_0,t]}\right) + \gamma_{2,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) + d_2
\end{aligned}$$

where $\gamma_{1,w}, \gamma_{2,w} \in \mathcal{K}$ *and* $d_1, d_2 \in \mathbb{R}_{\geq 0}$*, and every* $r_1 > r_0 > 0$ *satisfying the small-gain condition*

$$\gamma_{1,x}\left(\gamma_{2,x}(r)\right) \leq \rho r, \qquad \forall r \in [r_0, r_1],$$

if it can be guaranteed that

$$\|\boldsymbol{x}_1\|_{[0,\infty]} \leq r_1, \tag{2.64}$$

then $\forall t \geq t_0 \geq 0$*:*

$$\begin{aligned}
|\boldsymbol{x}_1(t)| \leq& \beta\left(\begin{vmatrix} \boldsymbol{x}_1(t_0) \\ \boldsymbol{x}_2(t_0) \end{vmatrix}, t - t_0\right) + \gamma\left(\gamma_{1,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right)\right) + \gamma\left(\gamma_{2,w}\left(\|\boldsymbol{w}\|_{[t_0,t]}\right)\right) + \\
& \lambda_1(d_1) + \lambda_2(d_2) + \lambda_0(r_0).
\end{aligned}$$

Proof of Theorem 2.8.1: Choose arbitrary $r_1 > r_0 > 0$, $\bar{\mu}$ such that $\bar{\mu} > \gamma_\mu\left(\gamma_2\left(r_1; \mu(0)\right); \mu(0)\right)$, $\rho < 1$

and consider the following small-gain condition:

$$\gamma_1\left(\gamma_2\left(\delta r,\mu\right)\right)\leq\rho r,\qquad\forall r\in\left[r_0,r_1\right]\subset\mathbb{R}_{\geq 0},\quad\forall\mu\in\left[0,\bar{\mu}\right]. \tag{2.65}$$

For every fixed μ, $\gamma_1\left(\cdot\right)$ and $\gamma_2\left(\cdot;\mu\right)$ are of class $\mathcal{K}_\infty$. Thus for every $r\in[r_0,r_1]$ and every $\mu\in[0,\bar{\mu}]$ there exists a small enough but strictly positive δ_A for which the small-gain condition holds. Since $[r_0,r_1]\times[0,\bar{\mu}]$ is a compact set and all the functions in (2.65) are continuous, the minimum of δ satisfying the small-gain condition over this whole set must also be strictly positive. Set $\delta_{\max}$ to be this minimum.

Since ρ in (2.65) is strictly smaller than 1, there exist $\alpha>0$ and $\rho'<1$ such that

$$\gamma_1\left(\left(1+\alpha\right)\gamma_2\left(r;\mu\right)\right)\leq\rho' r,\quad\forall r\in\left[r_0,r_1\right],\quad\forall\mu\in\left[0,\bar{\mu}\right]. \tag{2.66}$$

For all nondecreasing functions γ and all $\alpha>0$, $a>0$ and $b>0$, we have $\gamma\left(a+b\right)\leq\gamma\left(\left(1+\alpha\right)a\right)+\gamma\left(\left(1+1/\alpha\right)b\right)$. Using this and (2.66) we can derive $\forall t\geq 0$,

$$\begin{aligned}
\left|\boldsymbol{x}_1(t)\right| \quad &\leq \quad \beta_1\left(\left|\boldsymbol{x}_1\left(0\right)\right|,0\right)+\gamma\left(\left\|\boldsymbol{w}\right\|\right)+\\
&\qquad \gamma_1\left(\beta_2\left(\left|\boldsymbol{x}_2\left(0\right)\right|,0;\mu\left(0\right)\right)+\gamma_2\left(\left\|\boldsymbol{x}_1\right\|_{[0,t]}\right)+\gamma\left(\left\|\boldsymbol{w}\right\|;\mu\left(0\right)\right)\right)\\
&\leq \quad \beta_1\left(\left|\boldsymbol{x}_1\left(0\right)\right|,0\right)+\gamma_1\left(\left(1+\alpha\right)\gamma_2\left(\left\|\boldsymbol{x}_1\right\|_{[0,t]};\mu\left(0\right)\right)\right)+\\
&\qquad \gamma_1\left(\left(1+\frac{1}{\alpha}\right)\left(\beta_2\left(\left|\boldsymbol{x}_2\left(0\right)\right|,0;\mu\left(0\right)\right)+\gamma\left(\left\|\boldsymbol{w}\right\|;\mu\left(0\right)\right)\right)\right)+\gamma\left(\left\|\boldsymbol{w}\right\|\right).
\end{aligned}$$

Define

$$\begin{aligned}
s_\infty\left(\left|\boldsymbol{x}_1\left(0\right)\right|,\left|\boldsymbol{x}_2\left(0\right)\right|,\mu\left(0\right),\left\|\boldsymbol{w}\right\|\right)\doteq&\frac{1}{1-\rho'}\left(\beta_1\left(\left|\boldsymbol{x}_1\left(0\right)\right|,0\right)+\gamma\left(\left\|\boldsymbol{w}\right\|\right)\right)+\\
&\frac{1}{1-\rho'}\gamma_1\left(\left(1+\frac{1}{\alpha}\right)\left(\beta_2\left(\left|\boldsymbol{x}_2\left(0\right)\right|,0;\mu\left(0\right)\right)+\gamma\left(\left\|\boldsymbol{w}\right\|;\mu\left(0\right)\right)\right)\right).
\end{aligned}$$

By the choice of $\bar{\mu}$, it is always possible to find $s_{\max}<r_1$, $x_{\max}>0$, $w_{\max}>0$ such that

$$\begin{aligned}
s_\infty\left(\left|\boldsymbol{x}_1\left(0\right)\right|,\left|\boldsymbol{x}_2\left(0\right)\right|,\mu\left(0\right),\left\|\boldsymbol{w}\right\|\right)\leq&s_{\max}\leq r_1,\\
\gamma_\mu\left(\beta_2\left(\left|\boldsymbol{x}_2\left(0\right)\right|,0;\mu\left(0\right)\right)+\gamma_2\left(s_{max};\mu\left(0\right)\right)+\gamma\left(\left\|\boldsymbol{w}\right\|;\mu\left(0\right)\right),\mu\left(0\right)\right)<&\bar{\mu}
\end{aligned}$$

$$\forall\left|\boldsymbol{x}_1\left(0\right)\right|<x_{\max},\qquad\forall\left|\boldsymbol{x}_2\left(0\right)\right|<x_{\max},\qquad\forall\left\|\boldsymbol{w}\right\|_{[0,t]}<w_{\max}. \tag{2.67}$$

Given that (2.67) holds, we can use Lemma 2.8.2 to get $\|\boldsymbol{x}_1\|_{[0,\infty)]} \leq s_{\max}$. Using (2.58) we can also derive the bound

$$\|\mu\| \leq \gamma_\mu \left(\beta_2 \left(\left|\boldsymbol{x}_2\left(0\right)\right|, 0; \mu\left(0\right)\right) + \gamma_2 \left(s_{max}; \mu\left(0\right)\right) + \gamma\left(\|\boldsymbol{w}\|; \mu\left(0\right)\right), \mu\left(0\right)\right) < \bar{\mu}.$$

And with this, we can write

$$\begin{aligned}
\left|\boldsymbol{x}_1\left(t\right)\right| &\leq \beta_1\left(\left|\boldsymbol{x}_1\left(t_0\right)\right|, t - t_0\right) + \gamma_1\left(\|\boldsymbol{x}_2\|_{[t_0,t]}\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) \\
\left|\boldsymbol{x}_2\left(t\right)\right| &\leq \max_{\mu\in[0,\bar{\mu}]} \beta_2\left(\left|\boldsymbol{x}_2\left(t_0\right)\right|, t - t_0; \mu\right) + \max_{\mu\in[0,\bar{\mu}]} \gamma_2\left(\|\boldsymbol{x}_1\|_{[t_0,t]}; \mu\right) + \max_{\mu\in[0,\bar{\mu}]} \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}; \mu\right)
\end{aligned}$$

for all $t \geq t_0 \geq 0$. Note that for every fixed $\mu \in \mathbb{R}_{\geq 0}$ the function $\beta_2\left(\cdot, \cdot; \mu\right)$ is a function of class $\mathcal{KL}$ and the functions $\gamma_2\left(\cdot; \mu\right)$ and $\gamma\left(\cdot; \mu\right)$ are of class $\mathcal{K}_\infty$. They are also all continuous in μ. Thus taking the maximum of these functions over μ is well defined and does not change their $\mathcal{KL}/\mathcal{K}_\infty$ characteristics. Note that we can actually satisfy the following small-gain condition $\forall \delta \leq \delta_{\max}$:

$$\gamma_1\left(\gamma_2\left(\delta r, \mu\right)\right) \leq \rho r, \qquad \forall r \in \left[\lambda\left(\delta\right), r_1\right] \subset \mathbb{R}_{\geq 0}, \quad \forall \mu \in \left[0, \bar{\mu}\right].$$

where $\lambda \in \mathcal{K}$. Corollary 2.8.3 now gives us (2.59). ∎

2.9 Proofs of the Technical Lemmas

Proof of Lemma 2.4.1: Assume α satisfies $\sigma_{pi} + \frac{\alpha}{N} \leq 1$ and for simplicity assume also that P is a multiple of r. Then for all $l \in \{1 \ldots P/r - 1\}$:

$$\|\mu'\|_{lr\ldots(l+1)r-1} \leq \sigma_{pi}^l + \sum_{m=0}^{l-1} \sigma_{pi}^m \frac{\alpha}{N} \doteq V\left(l\right).$$

Because $\sigma_{pi} < 1$ we have that $V\left(l\right)$ converges to $\frac{\alpha}{N(1-\sigma)}$ as $l \to \infty$. We also have

$$\begin{aligned}
\|\mu'\|_{P-r\ldots P-1} \leq \max\Bigg\{ &\frac{N}{N-2}\sigma_{pi} V\left(P/r - 1\right) + \frac{\alpha}{N-2}, \\
&\left(\frac{N}{N-2}\sigma_{pi}\right)^r V\left(P/r - 1\right) + \sum_{m=0}^{r-1} \left(\frac{N}{N-2}\sigma_{pi}\right)^m \frac{\alpha}{N-2}\Bigg\}.
\end{aligned}$$

Since we can make $V\left(P/r - 1\right)$ arbitrarily small by taking P to be large enough and α to be small enough, we can make $\|\mu'\|_{P-r\ldots P-1} < 1$, which satisfies the convergence property. ∎

We will use the following definition in the proofs below:

$$\widetilde{D} \doteq \begin{bmatrix} -C & -CA_d^{-1} & \cdots & -CA_d^{-r+2} \\ 0 & -C & \cdots & -CA_d^{-r+3} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -C \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Proof of Lemma 2.4.3: Set $\lambda = \|\mu\|_{\{k'-r+1\ldots k'\}}$. Between time steps $k'+1$ and $k'+P$, μ is updated according to (2.16) or (2.17). Note that $F(\mu; k)$ depends linearly, with positive coefficients, on $\mu_{k-r} \ldots \mu_{k-1}$. Therefore, it is easy to see by induction from $k = k'+1$ to $k = k'+P$ that $\mu_k \leq \lambda \mu'_{k-k'+r-1}$. As we have that condition (2.19) holds, the result of the lemma follows. ■

Proof of Lemma 2.4.4: The settings $mode(k'+1) = update$ and $p = 0$ imply that for $m \in \{k'-r+1 \ldots k'\}$ we had $saturated(m) = \textbf{false}$ and either $mode(m) = capture$ or $mode(m) = detect$. The structure of our quantizer is such that if $saturated(m) = \textbf{false}$ for some m, then $|\tilde{\boldsymbol{y}}_m| < \mu_m$ where $\tilde{\boldsymbol{y}}_m \doteq \boldsymbol{z}_m - \boldsymbol{y}_m$ denotes the quantization error. The observations can be written as

$$\boldsymbol{z}_{k-l} = CA_d^l \boldsymbol{x}_k - C\sum_{i=1}^{l} A_d^{-i} \boldsymbol{u}_{k-l+i-1}^d - C\sum_{i=1}^{l} A_d^{-i} \boldsymbol{w}_{k-l+i-1}^d + \tilde{\boldsymbol{y}}_{k-l}. \tag{2.68}$$

Since the state estimate (2.13) was chosen so that $\hat{\boldsymbol{x}}_k = \boldsymbol{x}_k$ in the absence of measurement errors and disturbances, we get together with (2.68) that

$$\tilde{\boldsymbol{x}}_k^+ = G \begin{bmatrix} \tilde{\boldsymbol{y}}_{k-r+1} \\ \vdots \\ \tilde{\boldsymbol{y}}_k \end{bmatrix} + G\widetilde{D} \begin{bmatrix} \boldsymbol{w}_{k-r+1}^d \\ \vdots \\ \boldsymbol{w}_{k-1}^d \end{bmatrix}. \tag{2.69}$$

When taking the next measurement at time step $k+1$, the distance between the real output, $\boldsymbol{y}_{k+1}$, and the center of the quantizer $C\hat{\boldsymbol{x}}_{k+1}^-$ is

$$\left|\boldsymbol{y}_{k+1} - C\hat{\boldsymbol{x}}_{k+1}^-\right| = \left|CA_d\tilde{\boldsymbol{x}}_k^+ + C\boldsymbol{w}_k^d\right| \leq F(\mu; k+1) + \left\|\left[CA_dG\widetilde{D} \,|\, C\right]\right\| \left\|\boldsymbol{w}^d\right\|_{[k'-r+1,k]}. \tag{2.70}$$

Given that (2.23) holds with

$$\zeta_D \doteq \left\|\left[CA_dG\widetilde{D} \,|\, C\right]\right\|,$$

we have from (2.16) that

$$\left|\boldsymbol{y}_{k+1} - C\hat{\boldsymbol{x}}_{k+1}^{-}\right| \leq N\mu_{k+1}. \tag{2.71}$$

The structure of our quantizer guarantees in this case that $\left|\tilde{\boldsymbol{y}}_{k+1}\right| \leq \mu_{k+1}$. We can now repeat these arguments and show that (2.69)–(2.71) holds for all $k \in \{k' \dots k' + P - r\}$.

At time steps $k' + P - r$ the controller will switch to $mode(k' + P - r + 1) = detect$, and we will have for $l = P - r + 1$ that $\left|\boldsymbol{y}_{k'+l} - C\hat{\boldsymbol{x}}_{k'+l}^{-}\right| \leq (N-2)\,\mu_{k'+l}$. This guarantees that both $\left|\tilde{\boldsymbol{y}}_{k'+l}\right| \leq \mu_{k'+l}$ and $saturated(k' + l) = \textbf{false}$, thus $mode(k' + l + 1) = detect$. Again, we can repeat these arguments for $l \in \{P - r + 2 \dots P\}$ with the exception that for $l = P$ the controller will set $mode(k' + \ l + \ 1) = update$.

Based on (2.69) we can bound the estimation error for $l \in \{0 \dots P - 1\}$ as

$$\left|\tilde{\boldsymbol{x}}_{k'+l}^{+}\right| \leq \|G\|\,\|\mu\|_{\{k'-r+1\dots k'+l\}} + \left\|G\widetilde{D}\right\|\left\|\boldsymbol{w}^d\right\|_{\{k'-r+1\dots k'+l-1\}} \leq \zeta_\mu\,\|\mu\|_{\{k'-r+1\dots k'\}}$$

where

$$\zeta_\mu \doteq \|G\|\,\|\mu'\|_{\{0\dots r+P-2\}} + \left\|G\widetilde{D}\right\|\frac{\alpha}{\zeta_D}.$$

Note that in the definition of ζ_μ we used the constants μ''s defined in (2.18). ∎

Proof of Lemma 2.4.5: If (2.23) does not hold, then it will not necessarily be true that $\|\tilde{\boldsymbol{y}}_k\| \leq \mu_k$, $\forall k \in \{k' + 1 \dots k' + P - r\}$. However, since now we have that

$$\|\tilde{\boldsymbol{y}}\|_{\{k'-r+1\dots k'\}} \leq \|\mu\|_{\{k'-r+1\dots k'\}} \leq \frac{1}{\alpha}\zeta_D\left\|\boldsymbol{w}^d\right\|_{\{k'-r+1,k'+P\}} \tag{2.72}$$

we can still bound the estimation error as follows. For $k \in \{k' \dots k_2\}$ we have

$$\begin{aligned}\left\|\tilde{\boldsymbol{x}}_k^{+}\right\| &\leq \|G\|\,\|\tilde{\boldsymbol{y}}\|_{\{k-r+1,\dots,k\}} + \left\|G\widetilde{D}\right\|\left\|\boldsymbol{w}^d\right\|_{\{k-r+1,\dots,k-1\}}\\ \left\|\tilde{\boldsymbol{y}}_{k+1}\right\| &\leq \|CA_d\|\left\|\tilde{\boldsymbol{x}}_k^{+}\right\| + \|C\|\left|\boldsymbol{w}_k^d\right|.\end{aligned}$$

Iterating these two inequalities and combining with (2.72) we get $\left|\tilde{\boldsymbol{x}}_k^{+}\right| \leq \zeta_w\left\|\boldsymbol{w}^d\right\|_{k'-r+1\dots k-1}$ where

$$\zeta_w \doteq \|G\|\,\|CA_dG\|^P\frac{1}{\alpha}\zeta_D + \sum_{m=1}^{P}\|G\|\,\|CA_dG\|^{m-1}\left(\left\|CA_dG\widetilde{D}\right\| + \|C\|\right) + \left\|G\widetilde{D}\right\|.$$

∎

Proof of Lemma 2.4.6: Let k_3 be the first time step after k_2 such that $mode(k_3 + 1) = update$ (let

$k_3 = \infty$ if no such time step exists). We now have that for all $l \in \{0 \dots k_3 - k_2\}$

$$\begin{aligned}\left|\tilde{\boldsymbol{x}}^-_{k_2+l}\right| &\leq \|A_d\|^l \left|\tilde{\boldsymbol{x}}_{k_2}\right| + \sum_{m=0}^{l-1} \|A_d\|^{l-m-1} \left|\boldsymbol{w}^d_{k_2+m}\right| \leq \|A_d\|^l \left|\tilde{\boldsymbol{x}}_{k_2}\right| + \frac{\|A_d\|^l - 1}{\|A_d\| - 1} \left\|\boldsymbol{w}^d\right\| \\ &\leq \|A_d\|^l \left(\left|\tilde{\boldsymbol{x}}_{k_2}\right| + \zeta_C \left\|\boldsymbol{w}^d\right\|\right) \end{aligned} \tag{2.73}$$

where $\zeta_C \doteq \frac{1}{\|A_d\|-1}$. Now, the zoom factor grows as $\mu_{k_2+l} = \mu_{k_2}\Omega^l_{out}$. Define

$$T^*_1(\nu;\rho) \doteq \max\left\{0, \log_{\Omega_{out}/\|A_d\|}\left(\frac{\|C\|\,\nu}{\rho\,(N-2)}\right) + 1\right\} + r - 1$$

and note that when ρ is fixed, $T^*_1(\cdot;\rho)$ is a nondecreasing function. Assuming $mode(k) = capture$ $\forall k \in \left\{k_2+1 \dots k_2 + \left\lfloor T^*_1\left(\left|\boldsymbol{x}_{k_2}\right| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right)\right\rfloor\right\}$, we will have

$$\left|\boldsymbol{y}_{k_2+\lfloor T^*_1\rfloor-r+1} - C\hat{\boldsymbol{x}}^-_{k_2+\lfloor T^*_1\rfloor-r+1}\right| \leq \|C\|\left|\tilde{\boldsymbol{x}}^-_{k_2+\lfloor T^*_1\rfloor-r+1}\right| \leq (N-2)\,\mu_{k_2+\lfloor T^*_1\rfloor-r+1}.$$

Thus $saturated(k_2 + \lfloor T^*_1\rfloor - r + 1) = \textbf{false}$ as well as $saturated(k_2 + \lfloor T^*_1\rfloor + l) = \textbf{false}$ for $l = -r+2 \dots 0$ which guarantees that $k_3 \leq k_2 + T^*_1 < \infty$ where k_3 is the first time step after k_2 such that $mode\,(k_3+1) = update$ and $p\,(k_3) = 0$. Using (2.73) we can bound the estimation error until the controller switches to the *measurement update* mode at k_3 by

$$\left\|\tilde{\boldsymbol{x}}\right\|_{\{k_2 \dots k_3\}} \leq \tilde{\delta}_1\left(\left|\tilde{\boldsymbol{x}}_{k_2}\right| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right), \qquad \tilde{\delta}_1(\nu;\rho) \doteq \|A_d\|^{T^*_1(\nu;\rho)}\nu.$$

Note also that $\tilde{\delta}_1\left(\left|\tilde{\boldsymbol{x}}_{k_2}\right| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right) \leq \mu_{k_2}\zeta_b\Omega^{T^*_1\left(\left|\tilde{\boldsymbol{x}}_{k_2}\right|+\zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right)}_{out}$ where $\zeta_b \doteq \frac{(N-2)}{\|C\|}$. ■

Proof of Lemma 2.4.7: Assume first that $mode(k_2) = capture$ and consider the case $\left|\tilde{\boldsymbol{x}}_{k_2}\right| + \zeta_C\left\|\boldsymbol{w}^d\right\| \leq \frac{\mu_{k_2}}{\|C\|}$. Following the same arguments as in Lemma 2.4.6 which led to (2.73), we can write for $l \in \{1 \dots r\}$:

$$\|C\|\left|\tilde{\boldsymbol{x}}_{k_2+l}\right| \leq \|C\|\,\|A_d\|^l\left(\left|\tilde{\boldsymbol{x}}_{k_2}\right| + \zeta_C\left\|\boldsymbol{w}^d\right\|\right) \leq \mu_{k_2}\Omega^l_{out} = \mu_{k_2+l}.$$

This implies that if $mode(k_2) = capture$, then at $k_2 + r - p\,(k_2) \leq k_2 + r$ the controller will switch to the *measurement update* mode. If for some time step k the following holds

$$\left|\boldsymbol{y}_k - C\hat{\boldsymbol{x}}^-_k\right| \leq \|C\|\left|\tilde{\boldsymbol{x}}^-_k\right| \leq \mu_k \tag{2.74}$$

then the output from the quantizer will be such that $\boldsymbol{z}_k = \boldsymbol{c} = C\hat{\boldsymbol{x}}^-_k$. If for some time step k' (2.74) is true $\forall k \in \{k'-p \dots k'\}$, and $\hat{\boldsymbol{x}}^+_{k'}$ is updated with $G(\boldsymbol{z};\boldsymbol{u}^d;k')$, then we will have $\hat{\boldsymbol{x}}^+_{k'} = \hat{\boldsymbol{x}}^-_{k'}$. This implies

that $\tilde{\boldsymbol{x}}_{k'} = A_d \tilde{\boldsymbol{x}}_{k'-1} + \boldsymbol{w}^d_{k'-1}$. In turn, this means that if the estimation error is sufficiently small compared to the zoom factor, then (2.73) continues to hold for $l \in \{0 \dots k_3 - k_2\}$ even if we pick $k_3 > k_2$ such that $mode(k_3) \neq capture$.

Now define

$$\xi(\nu;\rho) \doteq \left(\frac{1}{\rho\varsigma}\right)^{\frac{\log\left(\|A_d\|^P\right)}{\log(\sigma)-\log\left(\|A_d\|^P\right)}} \left(\|C\|\,\nu\right)^{\frac{\log(\sigma)}{\log(\sigma)-\log\left(\|A_d\|^P\right)}}$$
$$T_2^*(\nu;\rho) \doteq P\left\lfloor \log_\sigma\left(\frac{\xi(\nu;\rho)}{\rho\varsigma}\right)\right\rfloor, \qquad \varsigma \doteq \min_{k\in\{r\dots r+P-1\}} \mu'(k) \leq \sigma.$$

Note that in the definition of ς we use the μ''s defined in (2.18) and we assume without loss of generality that $\varsigma > 0$. Assume also that $|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|$ is sufficiently small such that $T_2^*\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C \left\|\boldsymbol{w}^d\right\|; \mu_{k_2}\right) \geq r + P$. We defined ξ and T_2^* such that we will have for all $k \in \{k_2 \dots k_2 + T_2^*(\|\boldsymbol{w}\|)\}$

$$\mu_k \geq \mu_{k_2}\varsigma\sigma^{T_2^*\left(|\tilde{\boldsymbol{x}}_{k_2}|+\zeta_C\|\boldsymbol{w}^d\|;\mu_{k_2}\right)/P} > \xi\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right) \tag{2.75}$$

and

$$\begin{aligned}
&\|C\|\,\|\tilde{\boldsymbol{x}}\|_{\left\{k_2,\dots,k_2+T_2^*\left(|\tilde{\boldsymbol{x}}_{k_2}|+\zeta_C\|\boldsymbol{w}^d\|;\mu_{k_2}\right)\right\}} \leq \|A_d\|^{T_2^*\left(|\tilde{\boldsymbol{x}}_{k_2}|+\zeta_C\|\boldsymbol{w}^d\|;\mu_{k_2}\right)}\|C\|\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|\right)\\
&\leq \left(\frac{\xi\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right)}{\mu_{k_2}\varsigma}\right)^{\frac{\log\left(\|A_d\|^P\right)}{\log(\sigma)}} \|C\|\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|\right) = \xi\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right).
\end{aligned} \tag{2.76}$$

In deriving the first inequality in (2.76) we used (2.73) to bound the estimation error – even though it is not true that $mode(k_2 + l) = capture\ \forall l < T_2^*$, we can still use (2.73) since (2.75) and (2.76) imply that (2.74) holds. The proof is completed by setting

$$\tilde{\delta}_2(\nu;\rho) \doteq \frac{\xi(\nu;\rho)}{\|C\|}, \qquad \zeta_s \doteq \Omega^r_{out}/\varsigma \tag{2.77}$$

and letting $\varepsilon(\cdot) > 0$ be any class $\mathcal{K}$ function such that $\varepsilon(\rho) \leq \frac{\rho}{\|C\|}$ and $T_2^*(\varepsilon(\rho);\rho) \geq r + P$. Note that the function $\tilde{\delta}_2(\cdot;\rho)$ is a class $\mathcal{K}_\infty$ function for each fixed ρ, and that $\tilde{\delta}_2\left(|\tilde{\boldsymbol{x}}_{k_2}| + \zeta_C\left\|\boldsymbol{w}^d\right\|;\mu_{k_2}\right) \leq \mu_{k_2}\zeta_s\sigma^{T_2^*\left(|\tilde{\boldsymbol{x}}_{k_2}|+\zeta_C\|\boldsymbol{w}^d\|;\mu_{k_2}\right)/P}/\|C\|$. ∎

Proof of Lemma 2.7.2: First we have $\forall t \geq t_0 \geq \Delta$:

$$|\boldsymbol{x}_d(t)| \leq \beta_d\left(|\boldsymbol{x}_d(t_0)|, t - t_0\right) + \gamma_d\left(\|\boldsymbol{x}\|_{[t_0,t]}\right)$$

where $\beta_d(\nu,t) = 1_{t<\Delta}\nu + 1_{t\geq\Delta}e^{-1/\epsilon(t-\Delta)}$ with arbitrary $\epsilon > 0$ ($1_{t<\Delta}$ is the characteristic function whose value is 1 if $t < \Delta$ and 0 otherwise) and $\gamma_d(\nu) = \nu$. Note that $\beta_d \in \mathcal{KL}$ and $\gamma_d \in \mathcal{K}_\infty$. Defining $y(t) \doteq \gamma_x(|\boldsymbol{x}_d(t)|)$ we can have $\forall t \geq t_0 \geq \Delta$:

$$|y(t)| \leq \beta_y(|y(t_0)|, t-t_0) + \gamma_y\left(\|\boldsymbol{x}\|_{[t_0,t]}\right)$$

where $\beta_y(\nu,t) = \gamma_x\left(\beta_d\left(\gamma_x^{-1}(\nu), t\right)\right) \in \mathcal{KL}$ and $\gamma_y(\nu) = \gamma_x(\nu) \in \mathcal{K}_\infty$.

Invoking the small-gain theorem [25, Theorem 2.1], with $\beta_1(\nu,t) = \beta_x(\nu,t)$, $\gamma_1^y(\nu) = \nu$, $\gamma_1^u(\nu) = \gamma_w(\nu)$, $\beta_2(\nu,t) = \beta_d(\nu,t)$, $\gamma_2^y(\nu) = \gamma_x(\nu)$, $\gamma_2^u(\nu) = 0$, and $\rho_1 = \rho_2 = 1/\sqrt{\lambda} - 1$, we can get functions $\beta', \beta'' \in \mathcal{KL}$ such that $\forall t \geq t_0 \geq \Delta$:

$$|\boldsymbol{x}(t)| \leq \beta'\left(\left|\begin{array}{c}\boldsymbol{x}(t_0)\\ \gamma_x(|\boldsymbol{x}_d(t_0)|)\end{array}\right|, t-t_0\right) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right) \leq \beta''(|\boldsymbol{x}_d(t_0)|, t-t_0) + \gamma\left(\|\boldsymbol{w}\|_{[t_0,t]}\right)$$

for every $\gamma \in \mathcal{K}_\infty$ which satisfies (2.44). Because it must hold that $\beta''(\nu,0) \geq \nu$ we can arrive at (2.45) with $\beta(\nu,t) = \beta''(\nu, \max\{0, t-\Delta\})$. ■

Proof of Lemma 2.7.3: A standard result on ISS for linear systems is that the system defined by (2.46) follows

$$|\boldsymbol{x}(t)| \leq \tilde{\beta}_x(|\boldsymbol{x}(t_0)|, t-t_0) + \tilde{\gamma}_x\left(\|\boldsymbol{\theta}_x\|_{[t_0,t]}\right) + \tilde{\gamma}_x\left(\|\tilde{\boldsymbol{x}}\|_{[t_0,t]}\right) \tag{2.78}$$

where $\tilde{\beta}_x \in \mathcal{KL}$ and $\tilde{\gamma}_x \in \mathcal{K}_\infty$ is a linear function. For example one can take $\tilde{\beta}_x(\nu,t) = ce^{-\sigma t}\nu$ and $\tilde{\gamma}_x(\nu) = \frac{c\|BK\|}{\sigma}\nu$ where $c > 0$ and $\sigma > 0$ are such that $\left\|e^{(A+BK)t}\right\| \leq ce^{-\sigma t}\ \forall t \geq 0$.

We also have from the first line in (2.46), $\forall t \geq \Delta$:

$$\begin{aligned}
|\boldsymbol{\theta}_x(t)| =& \left|-\int_{t-\delta_{k(t)}}^{t} A\boldsymbol{x}(\tau) + BK\boldsymbol{x}\left(\tau - \delta_{k(\tau)}\right) + BK\tilde{\boldsymbol{x}}(\tau)\, d\tau\right| \\
\leq& \delta_{max}\left(\|A\| + \|BK\|\right)\|\boldsymbol{x}\|_{\left[t-\delta_{k(t)}-\delta_{k\left(t-\delta_{k(t)}\right)}, t\right]} + \delta_{max}\|BK\|\,\|\tilde{\boldsymbol{x}}\|_{\left[t-\delta_{k(t)}, t\right]} \\
\leq& \delta_{max}\left(\|A\| + \|BK\|\right)|\boldsymbol{x}_d(t)| + \delta_{max}\|BK\|\,|\tilde{\boldsymbol{x}}_d(t)|.
\end{aligned} \tag{2.79}$$

For the last inequality we used the fact that $\Delta \geq 2\delta_{\max}$. Substituting this into (2.78), we get (2.43) with

$$\begin{aligned}
\gamma_x(\nu) &= \tilde{\gamma}_x\left(\delta_{max}\left(\|A\| + \|BK\|\right)\nu\right) \\
\gamma_w(\nu) &= \tilde{\gamma}_x\left(\delta_{max}\|BK\|\,\nu\right) + \tilde{\gamma}_x(\nu)
\end{aligned}$$

(we used the fact that $\tilde{\gamma}_x$ is a linear function). Choosing $\bar{\delta}_{\max}$ such that $\tilde{\gamma}_x\left(\bar{\delta}_{\max}\left(\|A\| + \|BK\|\right)\nu\right) \leq \nu\ \forall\nu$,

(2.47) follows by Lemma 2.7.2. ■

Proofof Lemma 2.7.5: We can bound $\boldsymbol{w}_k^d$, $\forall k \geq 1$, as

$$\left|\boldsymbol{w}_k^d\right| \leq e^{T_s\|A\|}\left\|BK\right\| \int_{kT_s+\delta_k}^{(k+1)T_s+\delta_k} \left|\boldsymbol{\theta}_e\left(t\right)\right| dt + e^{T_s\|A\|}\left\|BK\right\| T_s \left\|\boldsymbol{\theta}_x\right\|_{[kT_s,(k+1)T_s]}.$$

We can also bound the estimation error between updates, $\forall k \geq 1$ and $\forall t \in [kT_s+\delta_k, (k+1)T_s+\delta_{k+1}]$:

$$\begin{aligned}\left|\tilde{\boldsymbol{x}}\left(t\right)\right| \leq & e^{(T_s+\delta_{\max})\|A\|}\left|\tilde{\boldsymbol{x}}_k\right| + e^{(T_s+\delta_{\max})\|A\|}\int_{kT_s+\delta_k}^{t}\left|\boldsymbol{w}\left(\tau\right)\right| d\tau \\ \leq & e^{(T_s+\delta_{\max})\|A\|}\left|\tilde{\boldsymbol{x}}_k\right| + e^{(T_s+\delta_{\max})\|A\|}\left\|BK\right\| \times \left(\int_{kT_s+\delta_k}^{t}\left|\boldsymbol{\theta}_e\left(\tau\right)\right| d\tau + \left(T_s+\delta_{\max}\right)\left\|\boldsymbol{\theta}_x\right\|_{[kT_s,t-\delta_k]}\right).\end{aligned}$$

Combining these two bounds with (2.51) and the first inequality in (2.79), we can arrive at, $\forall t \geq \Delta$:

$$\begin{aligned}\left|\tilde{\boldsymbol{x}}\left(t\right)\right| \leq & \beta_{e,e}\left(\left|\tilde{\boldsymbol{x}}_{k(t_0)}\right|, k\left(t\right)T_s - k\left(t_0\right)T_s; \mu_{k(t_0)}\right) + \gamma_{e,\theta}\left(\max_{k\in[k(t_0),k(t)]}\int_{kT_s+\delta_k}^{\min\{(k+1)T_s+\delta_{\max},t\}}\left|\boldsymbol{\theta}_e\left(\tau\right)\right| d\tau; \mu_{k(t_0)}\right) + \\ & \gamma_{e,e}\left(\delta_{\max}\left\|\tilde{\boldsymbol{x}}\right\|_{[k(t_0)T_s-\delta_{\max},t]}; \mu_{k(t_0)}\right) + \gamma_{e,x}\left(\delta_{\max}\left\|\boldsymbol{x}\right\|_{[k(t_0)T_s-2\delta_{\max},t]}; \mu_{k(t_0)}\right). \qquad (2.80)\end{aligned}$$

where $\beta_{e,e} \in \overline{\mathcal{KL}}$ and $\gamma_{e,\theta}, \gamma_{e,x}, \gamma_{e,e} \in \overline{\mathcal{K}}_\infty$.

From the definition of $\boldsymbol{\theta}_e$, $\forall t \geq \min\{2\delta_0, T_s+\delta_1\}$:

$$\boldsymbol{\theta}_e\left(t\right) = -\int_{t-\delta_{k(t)}}^{t}\dot{\tilde{\boldsymbol{x}}}\left(\tau\right)d\tau - \sum_{\tau\in\left(t-\delta_{k(t)},t\right]\cap\chi}\left(\tilde{\boldsymbol{x}}\left(\tau\right) - \tilde{\boldsymbol{x}}^-\left(\tau\right)\right) \qquad (2.81)$$

where $\chi \doteq \{t \geq 0 \,|\exists k \in \mathbb{N} \text{ such that } \tau = kT_s+\delta_k\}$. Each $t \in \chi$ affects $\boldsymbol{\theta}_e$ through the second term in (2.81) only in a time interval of length at most $\delta_{\max}$. The set $\left(kT_s+\delta_k-\delta_{k(kT_s+\delta_k)}, (k+1)T_s+\max\{\delta_k,\delta_{k+1}\}\right]\cap\chi$ contains at most two elements $\forall k \geq 1$. Using also (2.49), we can finally arrive at the bound: $\forall k \geq 2$ and $\forall t \in [kT_s+\delta_k, \max\{(k+1)T_s+\delta_k, (k+1)T_s+\delta_{k+1}\})$:

$$\begin{aligned}\int_{kT_s+\delta_k}^{t}\left|\boldsymbol{\theta}_e\left(\tau\right)\right| d\tau \leq & 4\delta_{\max}\left\|\tilde{\boldsymbol{x}}\right\|_{[kT_s,t]} + \\ & \delta_{\max}\left(T_s+\delta_{\max}\right)\left(\left\|A-BK\right\| + \left\|BK\right\|\right)\left\|\tilde{\boldsymbol{x}}\right\|_{[kT_s-\delta_{k-1},t]} + \\ & \delta_{\max}\left(T_s+\delta_{\max}\right)2\left\|BK\right\|\left\|\boldsymbol{x}\right\|_{[kT_s-\delta_{k-1}-\delta(kT_s-\delta_{k-1}),t-\delta_k]}. \qquad (2.82)\end{aligned}$$

Using (2.82) in (2.80) and the same argument we used at the end of the proof of Lemma 2.7.2 to move from a bound on $|\boldsymbol{x}\left(t\right)|$ to a bound on $|\boldsymbol{x}_d\left(t\right)|$, we can arrive at the result stated in the lemma. ■

Proof of Lemma 2.7.6: For any $x'_{\max} \geq 0$, $\bar{x}_{\max} \geq 0$, $\mu_{\max} \geq 0$, $r_1 > \max_{\mu\in[0,\mu_{\max}]}\tilde{\beta}_e\left(x'_{\max}, 0; \mu\right)$, and

$r_0 \in (0, r_1)$, one can find $\bar{\delta}_{\max} > 0$ and $\lambda < 1$ such that $\forall \mu \in [0, \mu_{\max}]$:

$$\tilde{\gamma}_e \left(\bar{\delta}_{\max} r; \mu \right) \leq \lambda r, \qquad \forall r \in [r_0, r_1] \tag{2.83}$$

and

$$\max \left\{ r_0, \frac{1}{1-\lambda} d \right\} \doteq \bar{\bar{x}}_{\max} < r_1 \tag{2.84}$$

where

$$d \doteq \max_{\mu \in [0, \mu_{\max}]} \tilde{\beta}_e \left(x'_{\max}, 0; \mu \right) + \max_{\mu \in [0, \mu_{\max}]} \tilde{\gamma}_w \left(\bar{\delta}_{\max} \bar{x}_{\max}; \mu \right).$$

Combining that with (2.52) and Lemma 2.8.2, we get $\| \tilde{\boldsymbol{x}}_d \|_{[\Delta, \infty]} \leq \bar{\bar{x}}_{\max}$, $\forall \left| \tilde{\boldsymbol{x}}_d \left(\Delta \right) \right| \leq x'_{\max}$, $\forall \| \boldsymbol{x}_d \|_{[\Delta, \infty]} \leq \bar{x}_{\max}$, and $\forall \mu_{k(\Delta)} \leq \mu_{\max}$ if $\delta_{\max} \leq \bar{\delta}_{\max}$. We remark that because $\tilde{\gamma}_e \left(r, \mu \right)$, for any fixed μ, grows faster than any linear function of r both at $r = 0$ and $r = \infty$, one cannot choose $r_0 = 0$ or $r_1 = \infty$ and still satisfy the assumptions in Lemma 2.8.2. We can now replace $\mu_{k(t_0)}$ in (2.52) with $\bar{\mu} = \max_{\mu \in [0, \mu_{\max}]} \psi \left(r_1; \mu \right)$ and write $\forall t \geq t_0 \geq \Delta$:

$$\begin{aligned} \left| \tilde{\boldsymbol{x}}_d \left(t \right) \right| \leq & \tilde{\beta}'_e \left(\left| \tilde{\boldsymbol{x}}_d \left(t_0 \right) \right|, t - t_0 \right) + \tilde{\gamma}'_e \left(\delta_{\max} \| \tilde{\boldsymbol{x}}_d \|_{[t_0, t]} \right) + \tilde{\gamma}'_w \left(\delta_{\max} \| \boldsymbol{x}_d \|_{[t_0, t]} \right), \\ \| \tilde{\boldsymbol{x}}_d \|_{[\Delta, \infty]} < & r_1 \end{aligned} \tag{2.85}$$

$\forall \left| \tilde{\boldsymbol{x}}_d \left(\Delta \right) \right| \leq x'_{\max}$, $\forall \| \boldsymbol{x}_d \|_{[\Delta, \infty]} \leq \bar{x}_{\max}$, $\forall \mu_{k(\Delta)} \leq \mu_{\max}$ and $\forall \delta_{\max} < \bar{\delta}_{\max}$ where $\tilde{\beta}'_e \in \mathcal{KL}$ and $\tilde{\gamma}'_e, \tilde{\gamma}'_w \in \mathcal{K}$. Taking $\bar{\delta}_{\max}$ to be smaller if necessary, we can also have

$$\tilde{\gamma}'_e \left(\delta_{\max} r \right) < r \qquad \forall r \in [r_0, r_1]. \tag{2.86}$$

We can now use the local version of the small-gain theorem (Corollary 2.8.3), similarly to how we used the small-gain theorem in Lemma 2.7.2, and arrive at (2.53). Note that when applying Corollary 2.8.3 to (2.85) and (2.86), we will have $d_1 = 0$ and $d_2 = 0$. Thus we get that $\lim_{r_0 \to 0} d' = 0$. And since we can choose r_0 to be arbitrarily small by reducing $\bar{\delta}_{\max}$, we can in turn make d' arbitrarily small.

Assume now that (2.83) and (2.84) hold for some $\delta_{\max} = \bar{\delta}_{\max}$. Then we can replace the constant λ with a function $\lambda \left(\delta_{\max} \right)$ and $\bar{\delta}_{\max}$ with $\delta_{\max}$ such that (2.83) and (2.84) continue to hold for every $\delta_{\max} \leq \bar{\delta}_{\max}$, and furthermore, $\lim_{\delta_{\max} \searrow 0} \lambda \left(\delta_{\max} \right) = 0$. We can then write $\| \tilde{\boldsymbol{x}}_d \|_{[\Delta, \infty)} < \bar{\bar{x}}_{\max}$ where

$$\bar{\bar{x}}_{\max} = \underbrace{\frac{1}{1 - \lambda \left(\delta_{\max} \right)} \max_{\mu \in [0, \mu_{\max}]} \tilde{\beta}_e \left(x'_{\max}, 0; \mu \right)}_{\doteq \bar{\gamma}_1 \left(\delta_{\max} \right)} + \underbrace{\frac{1}{1 - \lambda \left(\delta_{\max} \right)} \max_{\mu \in [0, \mu_{\max}]} \tilde{\gamma}_w \left(\delta_{\max} \bar{x}_{\max}; \mu \right)}_{\doteq \bar{\gamma}_e \left(\bar{x}_{\max}; \delta_{\max} \right)}$$

This gives us (2.54) where (2.55) follows from $\lim_{\delta_{\max} \searrow 0} \lambda(\delta_{\max}) = 0$ and $\tilde{\gamma}_w \in \mathcal{K}_\infty$. ■

Chapter 3

Minimum Sum of Distances Estimator: Robustness and Stability

3.1 Introduction

The problem of estimating a state $\boldsymbol{x}_0 \in \mathbb{R}^n$ from $m > n$ noisy linear measurements $\boldsymbol{y} \approx A\boldsymbol{x}_0 \in \mathbb{R}^m$, arises in a vast number of applications. In some applications one can assume that the difference between $\boldsymbol{y}$ and $A\boldsymbol{x}_0$ is a small i.i.d. Gaussian noise $\boldsymbol{z} \in \mathbb{R}^m$:

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{z}. \tag{3.1}$$

Given the model (3.1), the optimal estimate of $\boldsymbol{x}_0$ is the least-squares estimate: $\hat{\boldsymbol{x}}_2 = \left(A^T A\right)^{-1} A^T \boldsymbol{y} = \arg\min_x \|\boldsymbol{y} - A\boldsymbol{x}\|_2$. The least-squares estimate is known as *stable* in the sense that the estimation error $\|\hat{\boldsymbol{x}}_2 - \boldsymbol{x}_0\|_2$ is bounded by a continuous function of $\boldsymbol{z}$. Thus, small noise causes only small estimation error. Often, however, some of the measurements in $\boldsymbol{y}$ can be corrupted by arbitrarily large errors. In this case, we instead must solve $\boldsymbol{x}_0$ from the equation

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{z} + \boldsymbol{e} \tag{3.2}$$

where $\boldsymbol{e} \in \mathbb{R}^m$ has some arbitrarily large nonzero entries. One typical example is a GPS system, whose estimated position output can occasionally be considerably corrupted when the signals from the satellites are reflected off the surrounding terrain (i.e. multipath). Even one such corrupted measurement can cause arbitrarily large estimation error in the least-squares estimate.

When the state being estimated is a scalar ($n = 1$), the least-squares estimate $\hat{\boldsymbol{x}}_2$ is equivalent to taking a weighted average of the measurements. A known robust alternative to the average is the median. With the median, up to almost 50% of the measurements can be arbitrarily corrupted before the estimation error becomes unbounded. That is, the breakdown point of the median is 50%.

Taking the median, one essentially looks for the point which minimizes the sum of distances to all the measurements, whereas taking the average minimizes the sum of the squares of these distances. One

natural generalization of this concept to multivariate ($n > 1$) estimation[1] is to view the m measurements $\boldsymbol{y} \doteq [y_1, \ldots, y_m]^T$ as defining m hyperplanes:

$$H_i \doteq \left\{ \boldsymbol{x} \in \mathbb{R}^n \,\middle|\, y_i = \boldsymbol{a}_i^T \boldsymbol{x} \right\}$$

where $\boldsymbol{a}_i^T \in \mathbb{R}^n$ is the corresponding row of the matrix $A \doteq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^T$. Then the "median" estimate for $\boldsymbol{x}$ can be defined to be the point that minimizes the sum of distances to these hyperplanes:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{a}_i^T \boldsymbol{x} \right| = \arg\min_{\boldsymbol{x}} \left\| \boldsymbol{y} - A\boldsymbol{x} \right\|_1 . \tag{3.3}$$

To understand why this estimate can be robust to errors, let us assume the noise is zero for now: $\boldsymbol{z} = \boldsymbol{0}$. That is, we try to solve $\boldsymbol{x}_0$ from the equation $\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}$. If we could somehow compute $\boldsymbol{e}$, then $\boldsymbol{x}_0$ could be easily recovered from the clean system of equations $A\boldsymbol{x}_0 = \boldsymbol{y} - \boldsymbol{e}$. One approach to recovering $\boldsymbol{e}$ is to choose a matrix $B \in \mathbb{R}^{p \times m}$, $p = m - n$, with $BA = 0$, and define $\boldsymbol{w} = B\boldsymbol{y}$. Multiplying both sides of the measurement equation by B yields an underdetermined system of equations $\boldsymbol{w} = B\boldsymbol{e}$ in $\boldsymbol{e}$ alone. In the context of compressed sensing [6], it has recently been discovered that whenever $\boldsymbol{e}$ is sparse enough, it can be correctly recovered by solving the following ℓ^1-minimization problem:

$$\hat{\boldsymbol{e}} = \arg\min_{\boldsymbol{e}} \left\| \boldsymbol{e} \right\|_1 \quad \text{subject to} \quad \boldsymbol{w} = B\boldsymbol{e}. \tag{3.4}$$

So, in the noise free case, the two problems (3.3) and (3.4) are equivalent.

There is also a large literature analyzing the performance of (3.4) and related estimates in the presence of noise. The strongest available results ([5, 13], amongst others) have the following flavor: for some constants C and ρ, and almost all random matrices B, if one applies an ℓ^2-penalized version of (3.4) (i.e., the Lasso [72, 41]) and the number of errors $\left\| \boldsymbol{e} \right\|_0$ is less than $\rho \cdot n$, then the estimation error is bounded by $C \cdot \left\| \boldsymbol{z} \right\|$ for some $C > 0$. However, specific forms of the constants C and ρ are difficult to derive. A similar bound can be derived when B is known to be a *restricted isometry* [5]. However, it requires prior knowledge of the noise level, and the estimation error depends on the number of corrupted measurements, with the bound C diverging to infinity when the error fraction ρ approaches the breakdown point. Similar results have also been obtained for greedy alternatives to ℓ^1-minimization [47]. In this setting, one does not require a bound on the noise term. However, it does require that the number of corrupted measurements be considerably

[1] Another multivariate generalization of the median occurs in robust center-point estimation, where the observations are themselves points (rather than inner products). There, the estimator that minimizes the sum of distances to the observations, known as the Fermat-Weber point, achieves a breakdown point of 50% [36, Theorem 2.2]. Although the estimator studied here also generalizes the median, it addresses the more general problem of robust linear regression.

lower than the breakdown point for ℓ^1-minimization.

Whereas most of the existing stability results and bounds are derived for the underdetermined case (3.4), in this chapter, we directly study the stability of the ℓ^1 estimator for the overdetermined problem (3.3). Our bounds are weaker than those obtained in the asymptotic setting of large random matrices and small error fractions [13]. However, they hold for all matrices A, including the structured matrices arising in state estimation problems, and all error fractions ρ, up to the intrinsic breakdown point of the ℓ^1 estimator. Moreover, our bound has a very simple expression, whose derivation naturally suggests an algorithm for computing the intrinsic breakdown point of the ℓ^1 estimator. The complexity of our algorithm is exponentially lower than the existing alternative, and it is especially suitable for the kind of problems of interest to the system and control community – moderate-sized robust state estimation problems.

3.2 Preliminaries

Throughout, the 0-norm will denote the number of nonzero elements in a vector $\boldsymbol{v} \in \mathbb{R}^m$:

$$\|\boldsymbol{v}\|_0 \doteq \#\left\{i \middle| v_i \neq 0\right\}.$$

We will use $[m]$ to denote the set of indices $[m] \doteq \{1, 2, \ldots, m\}$. We will use the following notation for "positive" directional derivative of an arbitrary multivariate function $f : \mathbb{R}^m \to \mathbb{R}$:

$$D^+_{\boldsymbol{v}} f(\boldsymbol{x}) = \lim_{\varepsilon \searrow 0} \frac{f(\boldsymbol{x} + \varepsilon \boldsymbol{v}) - f(\boldsymbol{x})}{\varepsilon}.$$

Consider a general estimation problem, $\boldsymbol{y} = f(\boldsymbol{x}_0, \boldsymbol{z}, \boldsymbol{e})$, where $\boldsymbol{x}_0$ is the unknown state to be estimated, $\boldsymbol{z}$ is a noise term, $\boldsymbol{e}$ is a corruption term and $\boldsymbol{y}$ is the available measurements. Let $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$ be some estimate. We say that for given $\boldsymbol{x}_0$ and $\boldsymbol{z}$, the estimate is robust up to T corrupted measurements (or T-robust) if there exists a smooth function $\beta(\boldsymbol{x}_0, \boldsymbol{z}) \in \mathbb{R}$ such that

$$\forall \boldsymbol{e} : \text{if } \|\boldsymbol{e}\|_0 < T \text{ then } \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 \leq \beta(\boldsymbol{x}_0, \boldsymbol{z}). \tag{3.5}$$

The *breakdown point* of this scheme, $T^*(\boldsymbol{x}_0, \boldsymbol{z})$, is the minimum $T \in \mathbb{N}$ for which the estimation scheme is *not* T-robust. In other words,

$$T^*(\boldsymbol{x}_0, \boldsymbol{z}) \doteq \min\left\{T \in \mathbb{N} \,\middle|\, \sup_{\boldsymbol{e}, \|\boldsymbol{e}\|_0 \leq T} \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 = \infty\right\}.$$

We say T^* is a *stable* breakdown point if it does not depend on $\boldsymbol{x}_0$ and $\boldsymbol{z}$, i.e. $T^*(\boldsymbol{x}_0, \boldsymbol{z}) \equiv T^*$.

Throughout this chapter we consider the problem of estimating $\boldsymbol{x}_0$ from $\boldsymbol{y}$:

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{z} + \boldsymbol{e}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{z} \in \mathbb{R}^m$ and $\boldsymbol{e} \in \mathbb{R}^m$. For this problem we consider the minimum sum of distances (MSoD) estimation scheme

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} C_{\boldsymbol{y}}(\boldsymbol{x}) \tag{3.6}$$

with the cost function

$$C_{\boldsymbol{y}}(\boldsymbol{x}) \doteq \|\boldsymbol{y} - A\boldsymbol{x}\|_1 . \tag{3.7}$$

Our goal is to study whether the breakdown point of this estimate is stable and if so, how to compute it.

We start by giving results pertaining to the noiseless case, $\boldsymbol{z} = \boldsymbol{0}$. We assume T of the measurements can be corrupted. Geometrically, this means that the remaining $m - T$ measurement hyperplanes $H_i \doteq \left\{\boldsymbol{x} \in \mathbb{R}^n \mid y_i = \boldsymbol{a}_i^T\boldsymbol{x}\right\}$ pass through $\boldsymbol{x}_0$. We will let I denote the indices of these uncorrupted hyperplanes. The corrupted ones will be conveniently denoted by I^c. We ask whether these T hyperplanes can be positioned so that $\boldsymbol{x}_0$ no longer minimizes the cost function (3.7). Since $C_{\boldsymbol{y}}$ is convex, this will be true if and only if there exists a direction $\boldsymbol{v}$, from $\boldsymbol{x}_0$, along which the cost function does not increase, i.e. $D^+_{\boldsymbol{v}} C_{\boldsymbol{y}}(\boldsymbol{x}_0) \leq 0$. Since the uncorrupted hyperplanes pass through $\boldsymbol{x}_0$, moving in the direction of $\boldsymbol{v}$ from $\boldsymbol{x}_0$ will increase the distance to each of the uncorrupted hyperplanes at a rate of $\left|\boldsymbol{a}_i^T\boldsymbol{v}\right|$, $i \in I$. We have freedom in placing the corrupted hyperplanes, and so for each $\boldsymbol{v}$ we can position them so that moving in the direction of $\boldsymbol{v}$ will decrease the distance to each of the corrupted hyperplanes by a rate of $\left|\boldsymbol{a}_i^T\boldsymbol{v}\right|$, $i \in I^c$. In this case, which can be referred to as worst positioning of the corrupted hyperplanes given $\boldsymbol{v}$, the condition $D^+_{\boldsymbol{v}} C_{\boldsymbol{y}}(\boldsymbol{x}_0) \leq 0$ becomes

$$\sum_{i\in I} \left|\boldsymbol{a}_i^T\boldsymbol{v}\right| - \sum_{i\in I^c} \left|\boldsymbol{a}_i^T\boldsymbol{v}\right| \leq 0. \tag{3.8}$$

This is illustrated by the left diagram in Figure 3.1. Because (3.8) represents the worst case for a given $\boldsymbol{v}$, $\boldsymbol{x}_0$ fails to minimize the cost function if and only if (3.8) holds for some $\boldsymbol{v}$. Thus we arrive at a lemma following the next definition:

Definition 3.2.1. $\widetilde{T}(A)$ is defined as the minimal integer T for which there exists $I \subset [m]$, $|I| = m - T$ and $\boldsymbol{v} \in \mathbb{R}^n$ such that (3.8) holds.

Lemma 3.2.1. *Under the condition $\boldsymbol{z} = \boldsymbol{0}$, the breakdown point of the estimation scheme* (3.6) *is equal to*

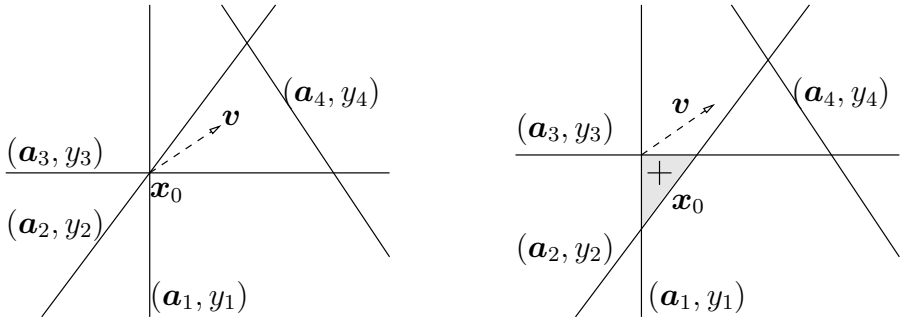

Figure 3.1: Low-dimensional ($n = 2$) illustration of the main ideas: here, the 1st, 2nd and 3rd lines are uncorrupted measurement hyperplanes, while the 4th has been corrupted. In the **left** diagram there is no noise. The cost function will not be minimized at $\boldsymbol{x}_0$ if there exists a direction $\boldsymbol{v}$ which increases the sum of distances to the uncorrupted hyperplanes at a rate slower the rate at which it decreases the sum of distances to the corrupted hyperplane(s). This condition is formulated in (3.8). In the **right** diagram there is also noise. The polytope P_I defined by the uncorrupted hyperplanes is shaded. The vector $\boldsymbol{v}$ illustrates a possible direction which reduces the cost function, even if (3.8) does not hold. This is because going in this direction will also reduce the distance to the 2nd hyperplane, which is uncorrupted. Note that the $\boldsymbol{a}_i$'s are vectors perpendicular to the hyperplanes they represent.

$\widetilde{T}$ as defined in Definition 3.2.1, i.e. $T^*(\boldsymbol{x}_0, \boldsymbol{0}) = \widetilde{T}(A)$, $\forall\, \boldsymbol{x}_0 \in \mathbb{R}^n$.

In the next section we will consider the noisy case and show that this breakdown point is stable and the estimation error is bounded by a linear function of the noise magnitude $\|\boldsymbol{z}\|_2$ that does not depend on $\boldsymbol{x}_0$. The difficulty that arises when dealing with the noisy case is illustrated by the right diagram in Figure 3.1

3.3 Proof of Robustness

We start with the following definition:

Definition 3.3.1. Given an arbitrary $T \in \mathbb{N}$ we call a set J' a *possibly extreme set* if there exists I, $I \supseteq J'$, $|I| = m - T$ such that the following holds:

$$\sum_{i \in J' \cup I^c} \left| \boldsymbol{a}_i^T \boldsymbol{\nu}_{J'} \right| \geq \sum_{i \in I \setminus J'} \left| \boldsymbol{a}_i^T \boldsymbol{\nu}_{J'} \right| \tag{3.9}$$

where $\boldsymbol{\nu}_{J'}$ is any of the singular vectors corresponding to the smallest singular value of the $|J'| \times n$ submatrix $A_{J'}$ of A containing those rows indexed by J': $\|A_{J'} \boldsymbol{\nu}_{J'}\|_2 = \sigma_{min}(A_{J'}) \|\boldsymbol{\nu}_{J'}\|_2$ with $\sigma_{min}(\cdot)$ being the smallest singular value. We define Q_T to be the set of all possibly extreme sets for a given T.

The following is our main result:

Theorem 3.3.1. *For any $T \in \{0, 1, \ldots, m\}$, if the number of corrupted measurements is not larger than T,*

then the estimation error is bounded as follows:

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 \leq \left(\max_{J' \in Q_T} \frac{1}{\sigma_{min}\left(A_{J'}\right)} \right) \|\boldsymbol{z}\|_2 . \tag{3.10}$$

Before proving Theorem 3.3.1 we emphasize a few observations. First note that if $T < \tilde{T}(A)$ then $\forall I \subset [m]$, $|I| = m - T$ the following holds:

$$\forall \boldsymbol{v} \in \mathbb{R}^n : \quad \sum_{i \in I} \left| \boldsymbol{a}_i^T \boldsymbol{v} \right| > \sum_{i \in I^c} \left| \boldsymbol{a}_i^T \boldsymbol{v} \right| . \tag{3.11}$$

Now, assume for some J' we have $\sigma_{min}\left(A_{J'}\right) = 0$. This implies that $\boldsymbol{a}_i^T \boldsymbol{\nu}_{J'} = 0 \; \forall i \in J'$. From (3.11) we see that in this case (3.9) cannot hold and thus $J' \notin Q_T$. From this we conclude that $\sigma_{min}\left(A_{J'}\right) > 0 \; \forall J' \in Q_T$ and thus the expression inside the brackets in (3.10) must be finite. The fact that we have established a finite bound (when $\|\boldsymbol{z}\|_2$ is finite) for all $T < \tilde{T}(A)$, and $\tilde{T}(A)$ is independent of $\boldsymbol{x}_0$ and $\boldsymbol{z}$, proves that the breakdown point $T^*(\boldsymbol{x}_0, \boldsymbol{z}) \equiv \tilde{T}(A)$ is stable.

The second observation is that for $T' < T$ we have $Q_{T'} \subseteq Q_T$ and possibly even $Q_{T'} \subset Q_T$ where some of the smaller sets in Q_T may not be in $Q_{T'}$. Since $J' \subseteq J$ implies $\sigma_{min}\left(A_{J'}\right) \leq \sigma_{min}\left(A_J\right)$, losing the smaller sets from Q_T (as we reduce the number of corrupted measurements) can produce a smaller bound in Theorem 3.3.1

Definition 3.3.2. We define the following sets:

$$\begin{aligned} J_+(\boldsymbol{x}, \boldsymbol{y}) &\doteq \left\{ i \in [m] \mid \boldsymbol{a}_i^T \boldsymbol{x} > y_i \right\} \\ J_0(\boldsymbol{x}, \boldsymbol{y}) &\doteq \left\{ i \in [m] \mid \boldsymbol{a}_i^T \boldsymbol{x} = y_i \right\} \\ J_-(\boldsymbol{x}, \boldsymbol{y}) &\doteq \left\{ i \in [m] \mid \boldsymbol{a}_i^T \boldsymbol{x} < y_i \right\} \end{aligned}$$

Also, for a point $\boldsymbol{x} \in \mathbb{R}^n$, $I_{\boldsymbol{x}}(\boldsymbol{y}) = J_0(\boldsymbol{x}, \boldsymbol{y}) \cap I$ is defined to be the set of uncorrupted hyperplanes passing through $\boldsymbol{x}$.

Proposition 3.3.1. *For any $\hat{\boldsymbol{x}} \in \mathbb{R}^n$:*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 \leq \frac{1}{\sigma_{min}\left(A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\right)} \|\boldsymbol{z}\|_2 .$$

Proof: Trivial since $\boldsymbol{z}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} = A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} (\hat{\boldsymbol{x}} - \boldsymbol{x}_0)$. ■

Our proof of Theorem 3.3.1 will go as follows. Assume $\boldsymbol{x}_0$, I, $\boldsymbol{z}$, $\boldsymbol{e}$ are given and let $\hat{\boldsymbol{x}}$ be the point minimizing the cost function. We will show that we can change only the noise and the corruption to $\boldsymbol{z}'$, $\boldsymbol{e}'$ such that $\|\boldsymbol{z}'\|_2 = \|\boldsymbol{z}\|_2$, $\boldsymbol{e}'_I = \boldsymbol{0}$, and the new corresponding minimizing point $\hat{\boldsymbol{x}}'$ achieves a larger estimation error, $\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_0\right\|_2 \geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2$. Furthermore, with the new $\boldsymbol{y}' = A\boldsymbol{x}_0 + \boldsymbol{z}' + \boldsymbol{e}'$ we will have $I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}') \in Q$. Applying then Proposition 3.3.1 on the new $\hat{\boldsymbol{x}}'$ and $\boldsymbol{y}'$, together with the fact that we did not decrease the estimation error, gives us (3.10). We will do this through several steps.

Proposition 3.3.2. *Let $\boldsymbol{y}$ and the corresponding point $\hat{\boldsymbol{x}}$ which minimizes the cost function be given. For a different $\boldsymbol{y}'$, if there exists a point $\boldsymbol{x}'$ such that*

$$\begin{aligned} J_+(\boldsymbol{x}', \boldsymbol{y}') &\subseteq J_+(\hat{\boldsymbol{x}}, \boldsymbol{y}) \\ J_-(\boldsymbol{x}', \boldsymbol{y}') &\subseteq J_-(\hat{\boldsymbol{x}}, \boldsymbol{y}) \\ J_0(\hat{\boldsymbol{x}}, \boldsymbol{y}) &\subseteq J_0(\boldsymbol{x}', \boldsymbol{y}'), \end{aligned} \tag{3.12}$$

then $\boldsymbol{x}'$ will minimize the cost function for $\boldsymbol{y}'$.

Proof: The rate of change of the cost function moving from $\boldsymbol{x}'$ in an arbitrary direction $\boldsymbol{v}$ is

$$\begin{aligned} D^+_{\boldsymbol{v}} C_{\boldsymbol{y}'}(\boldsymbol{x}') &= \sum_{i \in J_+(\boldsymbol{x}', \boldsymbol{y}')} \boldsymbol{a}_i^T \boldsymbol{v} + \sum_{i \in J_0(\boldsymbol{x}', \boldsymbol{y}')} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in J_-(\boldsymbol{x}', \boldsymbol{y}')} \boldsymbol{a}_i^T \boldsymbol{v} \\ &\geq \sum_{i \in J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \boldsymbol{a}_i^T \boldsymbol{v} + \sum_{i \in J_0(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \boldsymbol{a}_i^T \boldsymbol{v} = D^+_{\boldsymbol{v}} C_{\boldsymbol{y}}(\hat{\boldsymbol{x}}) > 0. \end{aligned}$$

■

Lemma 3.3.2. *Assume $\boldsymbol{x}_0$, I, $\boldsymbol{z}$, $\boldsymbol{e}$ are given and let $\hat{\boldsymbol{x}}$ be the point minimizing the cost function. There exists $\boldsymbol{z}'$, $\boldsymbol{e}'$ such that $\|\boldsymbol{z}'\| = \|\boldsymbol{z}\|$, $\boldsymbol{e}'_I = \boldsymbol{0}$, and the new corresponding minimizing point $\hat{\boldsymbol{x}}'$ achieves a larger estimation error, $\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_0\right\|_2 \geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2$. Furthermore, either $\boldsymbol{v}' \doteq \hat{\boldsymbol{x}}' - \boldsymbol{x}_0 \propto \boldsymbol{\nu}_{I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')}$ or $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \subsetneq I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')$.*

Proof: Define $\boldsymbol{v} \doteq \hat{\boldsymbol{x}} - \boldsymbol{x}_0$. If $\boldsymbol{v} \propto \boldsymbol{\nu}_{I_{\hat{x}}}$ then we are done. Otherwise set $\bar{\boldsymbol{v}} \propto \boldsymbol{\nu}_{I_{\hat{x}}}$, $\langle \bar{\boldsymbol{v}}, \boldsymbol{v} \rangle \geq 0$, $\|\bar{\boldsymbol{v}}\|_2 = 1$. Also, set $\bar{\boldsymbol{v}}^\perp$ to be the normalized vector perpendicular to $\bar{\boldsymbol{v}}$, in the span of $\boldsymbol{v}$ and $\bar{\boldsymbol{v}}$, and such that $\left\langle \bar{\boldsymbol{v}}^\perp, \boldsymbol{v} \right\rangle \geq 0$. Consider the vector function

$$f(\alpha) = \frac{\cos(\alpha)\, \bar{\boldsymbol{v}}^\perp + \sin(\alpha)\, \bar{\boldsymbol{v}}}{\left\| A_{I_{\hat{x}}(\boldsymbol{y})} \left(\cos(\alpha)\, \bar{\boldsymbol{v}}^\perp + \sin(\alpha)\, \bar{\boldsymbol{v}}\right) \right\|_2} \left\| \boldsymbol{z}_{I_{\hat{x}}(\boldsymbol{y})} \right\|_2.$$

Define $\alpha_0 = \sin^{-1}\left(\langle \bar{\boldsymbol{v}}, \boldsymbol{v} \rangle\right) \in [0, \pi/2]$. Note that if we set

$$\begin{aligned}
\bar{\boldsymbol{z}}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}(\alpha) &= A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} f(\alpha) \\
\bar{\boldsymbol{z}}_{[m]\setminus I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}(\alpha) &= \boldsymbol{z}_{[m]\setminus I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} \\
\bar{\boldsymbol{e}}_{I^c}(\alpha) &= \boldsymbol{e}_{I^c} + A_{I^c} f(\alpha) - A_{I^c} f(\alpha_0) \\
\bar{\boldsymbol{e}}_{I}(\alpha) &= \boldsymbol{0}
\end{aligned}$$

then $\bar{\boldsymbol{z}}(\alpha_0) = \boldsymbol{z}$ and for $\alpha \in [\alpha_0, \pi/2]$ we have $\|\bar{\boldsymbol{z}}\|_2 = \|\boldsymbol{z}\|_2$.

We will set $\boldsymbol{z}' = \bar{\boldsymbol{z}}(\alpha^*)$, $\boldsymbol{e}' = \bar{\boldsymbol{e}}(\alpha^*)$ where

$\alpha^* = \max\{\pi/2, \tilde{\alpha}\}$

$$\tilde{\alpha} = \sup\left\{ \alpha \;\middle|\; \begin{array}{c} \boldsymbol{a}_i f(\alpha) > z_i \,\forall i \in I \cap J_+(\hat{\boldsymbol{x}}, \boldsymbol{y}) \\ \textit{and} \\ \boldsymbol{a}_i f(\alpha) < z_i \,\forall i \in I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y}) \end{array} \right\}.$$

With this choice of $\boldsymbol{z}'$ and $\boldsymbol{e}'$ we guarantee that (3.12) holds with $\boldsymbol{x}' = \boldsymbol{x}_0 + f(\alpha^*)$, and therefore $\boldsymbol{v}' = f(\alpha^*)$ is the new estimation error. If $\alpha^* = \pi/2$ then $\boldsymbol{v}' \propto \boldsymbol{\nu}_{I_{\hat{\boldsymbol{x}}'}}$. Otherwise one of the strict inequalities in (3.13) must become an inequality with α^*, which implies $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \subsetneq I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')$. To complete the proof we are left to show that

$$\|f(\alpha)\|_2 = \frac{\left\|\cos(\alpha)\,\bar{\boldsymbol{v}}^{\perp} + \sin(\alpha)\,\bar{\boldsymbol{v}}\right\|_2}{\left\|A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\left(\cos(\alpha)\,\bar{\boldsymbol{v}}^{\perp} + \sin(\alpha)\,\bar{\boldsymbol{v}}\right)\right\|_2} \left\|\boldsymbol{z}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\right\|_2 \tag{3.13}$$

is monotonically non-decreasing.

The numerator in (3.13) as well as the $\left\|\boldsymbol{z}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\right\|_2$ term are constants. Because the singular vector $\bar{\boldsymbol{v}}$ is an eigenvector of $A^T_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}$ we have that $\left\langle A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\bar{\boldsymbol{v}}^{\perp}, A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\bar{\boldsymbol{v}}\right\rangle = 0$, thus the derivative of the denominator with respect to α is

$$\frac{\left(-\left\|A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\bar{\boldsymbol{v}}^{\perp}\right\|_2^2 + \left\|A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\bar{\boldsymbol{v}}\right\|_2^2\right)\sin(\alpha)\cos(\alpha)}{\left\|A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\left(\cos(\alpha)\,\bar{\boldsymbol{v}}^{\perp} + \sin(\alpha)\,\bar{\boldsymbol{v}}\right)\right\|_2}.$$

This is always non-positive because $\alpha \in [0, \pi/2]$ and $\bar{\boldsymbol{v}}$ is the singular vector corresponding to the smallest singular value. ∎

By iterating the procedure described in the last lemma, each time adding at least one more element to $I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')$, we arrive at the following corollary:

Corollary 3.3.3. *Assume $\boldsymbol{x}_0$, I, $\boldsymbol{z}$, $\boldsymbol{e}$ are given and let $\hat{\boldsymbol{x}}$ be the point minimizing the cost function. There exists $\boldsymbol{z}'$, $\boldsymbol{e}'$ such that $\|\boldsymbol{z}'\|_2 = \|\boldsymbol{z}\|_2$, $\boldsymbol{e}'_I = \boldsymbol{0}$, the new corresponding minimizing point $\hat{\boldsymbol{x}}'$ achieves a larger estimation error, $\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_0\right\|_2 \geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2$, and $\boldsymbol{v}' \doteq \hat{\boldsymbol{x}}' - \boldsymbol{x}_0 \propto \boldsymbol{\nu}_{I_{\hat{\boldsymbol{x}}'}}$.*

Remark 3.3.1. Without loss of generality, for a given $\boldsymbol{y} \in \mathbb{R}^m$ and an arbitrary direction $\boldsymbol{v} \in \mathbb{R}^n$, we can assume that $\boldsymbol{a}_i^T \boldsymbol{v} \geq 0 \ \forall\, i \in [m]$. This is because we can arbitrarily negate some of the $\boldsymbol{a}_i$'s and their corresponding y_i's without affecting the cost function (3.7).

Lemma 3.3.4. *Assume $\boldsymbol{x}_0$, I, $\boldsymbol{z}$, $\boldsymbol{e}$ are given. Let $\hat{\boldsymbol{x}}$ be the point minimizing the cost function and assume $\boldsymbol{v} \doteq \hat{\boldsymbol{x}} - \boldsymbol{x}_0 \propto \boldsymbol{\nu}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}$. If $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \notin Q$ then there exists $\boldsymbol{z}'$, $\boldsymbol{e}'$ such that $\|\boldsymbol{z}'\|_2 = \|\boldsymbol{z}\|_2$, $\boldsymbol{e}'_I = \boldsymbol{0}$, the new corresponding minimizing point $\hat{\boldsymbol{x}}'$ achieves a larger estimation error, $\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_0\right\|_2 \geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2$, and $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \subsetneq I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')$.*

Proof: WLOG (see Remark 3.3.1) assume $\boldsymbol{a}_i^T \boldsymbol{v} \geq 0 \ \forall i \in [m]$. The rate of change going in direction $-\boldsymbol{v}$ from $\hat{\boldsymbol{x}}$ is

$$-\sum_{i \in J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| + \sum_{i \in J_0(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| + \sum_{i \in J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right|. \tag{3.14}$$

Because $\hat{\boldsymbol{x}}$ minimizes the cost function, (3.14) must be nonnegative. If indeed $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \notin Q$ then from the fact that (3.9) is not satisfied for $J' = I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})$ we have

$$\sum_{i \in I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| > \sum_{i \in I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| + \sum_{i \in I^c} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in I \cap J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right|.$$

Now given that (3.14) is nonnegative we can write

$$\begin{aligned} &\sum_{i \in I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in I \cap J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| \\ &\quad \geq \sum_{i \in I^c \cap J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in I^c \cap J_0(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in I^c \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| - \sum_{i \in I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right|. \end{aligned}$$

Combining these last two inequalities we get

$$2 \sum_{i \in I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| > 2 \sum_{i \in I^c \cap J_+(\hat{\boldsymbol{x}}, \boldsymbol{y})} \left|\boldsymbol{a}_i^T \boldsymbol{v}\right| \geq 0$$

which implies that $I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})$ cannot be empty. Now, for every z_i, $i \in I \cap J_-(\hat{\boldsymbol{x}}, \boldsymbol{y})$, we have

$$z_i = y_i - \boldsymbol{a}_i^T \boldsymbol{x}_0 = y_i + \boldsymbol{a}_i^T \boldsymbol{v} - \boldsymbol{a}_i^T \hat{\boldsymbol{x}} > 0.$$

Arbitrarily choose $i' \in I \cap J_{-}(\hat{\boldsymbol{x}}, \boldsymbol{y})$ and consider the following:

$$\begin{aligned}
\bar{\boldsymbol{z}}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}(\alpha) &= A_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}(1+\alpha)\,\boldsymbol{v} \\
\bar{z}_{i'}(\alpha) &= \sqrt{z_{i'}^2 + \left(1-(1+\alpha)^2\right)\left\|\boldsymbol{z}_{I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})}\right\|_2^2} \\
\bar{\boldsymbol{z}}_{[m]\setminus\left(I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\cup\{i'\}\right)}(\alpha) &= \bar{\boldsymbol{z}}_{[m]\setminus\left(I_{\hat{\boldsymbol{x}}}(\boldsymbol{y})\cup\{i'\}\right)} \\
\bar{\boldsymbol{e}}_{I^c}(\alpha) &= \bar{\boldsymbol{e}}_{I^c}(\alpha) + \alpha A_{I^c}\boldsymbol{v} \\
\bar{\boldsymbol{e}}_I &= \boldsymbol{0}
\end{aligned}$$

with $\alpha \geq 0$. Note that $\|\bar{\boldsymbol{z}}(\alpha)\|_2$ is constant, and $\bar{\boldsymbol{z}}(0) = \boldsymbol{z}$. We will set $\boldsymbol{z}' = \boldsymbol{z}(\alpha^*)$ and $\boldsymbol{e}' = \bar{\boldsymbol{e}}(\alpha^*)$ where

$$\alpha^* = \sup\left\{\alpha \,\middle|\, \boldsymbol{a}_i^T(1+\alpha)\,\boldsymbol{v} < \bar{z}_i(\alpha)\ \forall i \in I \cap J_{-}(\hat{\boldsymbol{x}}, \boldsymbol{y})\right\}.$$

For every $\alpha \in [0, \alpha^*)$ we have that (3.12) holds with $\boldsymbol{x}' = \boldsymbol{x}_0 + (1+\alpha)\,\boldsymbol{v}$ and therefore $(1+\alpha)\,\boldsymbol{v}$ is the new estimation error. With $\alpha = \alpha^*$ we also have $\boldsymbol{a}_i^T(1+\alpha)\,\boldsymbol{v} = z_i' \Leftrightarrow \boldsymbol{a}_i^T\boldsymbol{x}' = y_i'$ for some $i \in I \cap J_{-}(\hat{\boldsymbol{x}}, \boldsymbol{y})$. This implies $I_{\hat{\boldsymbol{x}}}(\boldsymbol{y}) \subsetneq I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}')$. ■

By iterating the procedures described in (3.3.2) and (3.3.4) several times as necessary we arrive at the final corollary:

Corollary 3.3.5. *Assume $\boldsymbol{x}_0$, I, $\boldsymbol{z}$, $\boldsymbol{e}$ are given and let $\hat{\boldsymbol{x}}$ be the point minimizing the cost function. There exists $\boldsymbol{z}'$, $\boldsymbol{e}'$ such that $\|\boldsymbol{z}'\|_2 = \|\boldsymbol{z}\|_2$, $\boldsymbol{e}'_I = \boldsymbol{0}$, and the new corresponding minimizing point $\hat{\boldsymbol{x}}'$ achieves a larger estimation error, $\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_0\right\|_2 \geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2$. Furthermore, $I_{\hat{\boldsymbol{x}}'}(\boldsymbol{y}') \in Q$.*

The last corollary, together with Proposition 3.3.1, proves Theorem 3.3.1.

3.4 Computing the Breakdown Point

Definition 3.2.1 does not immediately suggest an algorithm for computing $\widetilde{T} = T^*$, because it requires checking condition (3.8) for all $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_2 = 1$, and there are infinitely many such $\boldsymbol{v}$. The following Lemma 3.4.1, however, states that it is sufficient to check only a finite subset of $\mathbb{R}^n$:

Lemma 3.4.1. *Condition* (3.8) *holds for some $I \subset [m]$ and $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ if and only if there exist $J \subset [m]$ and $\boldsymbol{v}' \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ with the following properties: $|J| = n-1$; $\{\boldsymbol{a}_i\}_{i\in J}$ is a set of $n-1$ linearly independent vectors; $\boldsymbol{a}_i^T\boldsymbol{v}' = 0\ \forall i \in J$; and* (3.8) *holds for $\boldsymbol{v}'$.*

The *if* direction in 3.4.1 is trivial. In the degenerate case where $\dim \operatorname{span}\{\boldsymbol{a}_i\}_{i\in I} \leq n-1$ the *only if*

is also trivial since (3.8) will hold for any nonzero vector which is not in the span of $\{\boldsymbol{a}_i\}_{i\in I}$. The *only if* direction in the non-degenerate case is an immediate corollary of the following proposition:

Proposition 3.4.1. *Assume I and $\boldsymbol{v}$ are given, and* $\dim \operatorname{span}\{\boldsymbol{a}_i\}_{i\in I} = n$. *Define $J(\boldsymbol{v}) \doteq \{i \in I \,|\boldsymbol{a}_i\boldsymbol{v} = 0\}$ and $d(J) \doteq \dim \operatorname{span}\{\boldsymbol{a}_i\}_{i\in J}$. If Condition* (3.8) *holds for $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ and $d(J(\boldsymbol{v})) < n-1$ then there exists $\boldsymbol{v}' \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ for which* (3.8) *also holds but in addition $d(J(\boldsymbol{v}')) > d(J(\boldsymbol{v}))$.*

Proof: WLOG we can assume $\boldsymbol{a}_i^T\boldsymbol{v} \geq 0\ \forall i \in [m]$. Consider the following set of equations in $\boldsymbol{z} \in \mathbb{R}^n$:

$$\sum_{i\in I^c} \boldsymbol{a}_i^T\boldsymbol{z} = 0 \tag{3.15}$$

$$\boldsymbol{a}_i^T\boldsymbol{z} = 0 \quad \forall i \in J(\boldsymbol{v}). \tag{3.16}$$

In the case that $d(J(\boldsymbol{v})) = \dim \operatorname{span}\{\boldsymbol{a}_i\}_{i\in J(\boldsymbol{v})} < n-1$, there is a nontrivial solution $\tilde{\boldsymbol{z}} \neq 0$ to (3.15) and (3.16). By changing the sign of $\tilde{\boldsymbol{z}}$ if necessary, we can assume

$$\sum_{i\in I} \boldsymbol{a}_i^T\tilde{\boldsymbol{z}} \leq 0. \tag{3.17}$$

Define the set $P \doteq \left\{i \in I \,\middle|\boldsymbol{a}_i^T\tilde{\boldsymbol{z}} < 0\right\}$ and $\alpha \doteq \min_{i\in P} \frac{\boldsymbol{a}_i^T\boldsymbol{v}}{-\boldsymbol{a}_i^T\tilde{\boldsymbol{z}}}$. Note that from (3.17) and the assumption that $d(I) = n$, P cannot be empty and thus α is well defined and positive. Also note that P contains only the indices of vectors from I which are linearly independent of $\{\boldsymbol{a}_i\}_{i\in J(\boldsymbol{v})}$. Set $\boldsymbol{v}' = \boldsymbol{v} + \alpha\tilde{\boldsymbol{z}}$. By our choice of α we have for some $i' \in P \subset I \setminus J$ that $\boldsymbol{a}_{i'}^T\boldsymbol{v}' = 0$. Since $\tilde{\boldsymbol{z}}$ satisfies (3.16) this gives us $J(\boldsymbol{v}') \supsetneq J(\boldsymbol{v})$ and $d(J(\boldsymbol{v}')) > d(J(\boldsymbol{v}))$. From (3.15) we have

$$\sum_{i\in I^c} \left|\boldsymbol{a}_i^T\boldsymbol{v}'\right| \geq \sum_{i\in I^c} \boldsymbol{a}_i^T(\boldsymbol{v} + \alpha\tilde{\boldsymbol{z}}) = \sum_{i\in I^c} \boldsymbol{a}_i^T\boldsymbol{v} = \sum_{i\in I^c} \left|\boldsymbol{a}_i^T\boldsymbol{v}\right|. \tag{3.18}$$

By our choice of α we also have $\boldsymbol{a}_i^T\boldsymbol{v}' \geq 0\ \forall i \in I$. Together with (3.17) this gives us

$$\sum_{i\in I} \left|\boldsymbol{a}_i^T\boldsymbol{v}'\right| = \sum_{i\in I} \boldsymbol{a}_i^T\boldsymbol{v}' = \sum_{i\in I} \boldsymbol{a}_i^T\boldsymbol{v} + \alpha\sum_{i\in I} \boldsymbol{a}_i^T\tilde{\boldsymbol{z}} \leq \sum_{i\in I} \left|\boldsymbol{a}_i^T\boldsymbol{v}\right|. \tag{3.19}$$

Combining (3.18), (3.19) and the fact that (3.8) holds for $\boldsymbol{v}$ implies that (3.8) also holds for $\boldsymbol{v}'$. ∎

Given $J \subset I$, $|J| = n-1$, $d(J) = n-1$, the condition $A_J\boldsymbol{v}' = \boldsymbol{0}$ determines $\boldsymbol{v}'$ uniquely up to scale. The validity of condition (3.8) is unchanged by scaling $\boldsymbol{v}'$. Thus, we could equivalently define $T^*(A)$ to be the minimal integer T such that there exists $J \subset [m]$ of size $|J| = n-1$, $d(J) = n-1$, and $I \subset [m]$ of size $|I| = m - T$ for which condition (3.8) holds for $\boldsymbol{v}'$ satisfying $A_J\boldsymbol{v}' = \boldsymbol{0}$. Fix J (and a corresponding $\boldsymbol{v}$),

and sort the $|\boldsymbol{a}_i^T\boldsymbol{v}|$ such that $\left|\boldsymbol{a}_{r_1}^T\boldsymbol{v}\right| \geq \left|\boldsymbol{a}_{r_2}^T\boldsymbol{v}\right| \geq \ldots \geq \left|\boldsymbol{a}_{r_m}^T\boldsymbol{v}\right|$. Then, condition (3.8) holds for some I of size $m-T$ if and only if it holds for $I \doteq \{r_{T+1} \ldots r_m\}$. We can therefore compute $T^*(A)$ by checking this condition for every subset J of size $n-1$. This idea is formalized as Algorithm 5.

Algorithm 5 Computing $T^*(A)$

Require: $A \in \mathbb{R}^{m\times n}$.

Set $T \leftarrow m$ and let $J_1, \ldots, J_N$, $N = \binom{m}{n-1}$, be all the subsets of $[m] \doteq \{1 \ldots m\}$ containing $n-1$ indices.
for $k = 1 : N$ **do**
  if $\dim \operatorname{span}\{\boldsymbol{a}_i\}_{i\in J_k} = n-1$ **then**
    Find a nontrivial solution $\boldsymbol{v} \in \mathbb{R}^n$ such that
  $\boldsymbol{a}_i^T\boldsymbol{v} = 0\ \forall i \in J_k$.
    Find the order $r_1 \ldots r_m$ such that
  $\left|\boldsymbol{a}_{r_1}^T\boldsymbol{v}\right| \geq \left|\boldsymbol{a}_{r_2}^T\boldsymbol{v}\right| \geq \ldots \geq \left|\boldsymbol{a}_{r_m}^T\boldsymbol{v}\right|$.
    Find the smallest integer, s, such that

$$\sum_{i=1}^{s} \left|\boldsymbol{a}_{r_i}^T\boldsymbol{v}\right| \geq \sum_{i=s+1}^{m} \left|\boldsymbol{a}_{r_i}^T\boldsymbol{v}\right|.$$

    Set $T \leftarrow \min\{T, s\}$.
  end if
end for

Ensure: T.

The computation time of Algorithm 5 is

$$\binom{m}{n}\left(t_{sle}(n-1) + t_{mv}(m) + t_{sort}(m)\right) \tag{3.20}$$

where $t_{sle}(n) = \mathrm{O}\left(n^3\right)$, $t_{mv}(n) = \mathrm{O}\left(n^2\right)$ and $t_{sort}(n) = \mathrm{O}(n\log n)$ are the times it takes to solve a system of linear equations, to compute a matrix-vector multiplication, and to sort, respectively. When both m and n grow, $\binom{m}{n}$, and thus the computation time of our algorithm, grows exponentially. In many control applications, however, the number of variables describing the state of the system, n, is fixed, while the number of measurements, m, is flexible. In this case, where n is fixed, our algorithm's computation time is polynomial in m. We further note that, while the running time of the algorithm might still be relatively large in practice, from the engineering design point of view it needs to be executed only once during the design of the system to analyze its performance. In real time only (3.6) needs to be evaluated, which can be done very efficiently using linear programming.

The algorithm described above is different from the existing algorithm in the literature for computing the breakdown point. In the introduction we have mentioned that in the absence of noise, (3.3) and (3.4) are equivalent problems when $B \in \mathbb{R}^{p\times m}$, $p = m-n$, $BA = 0$. The following result, proved in [12] and in [6, §II], states that the ability of (3.4) to recover $\boldsymbol{e}$ from the underdetermined linear system $\boldsymbol{w} = B\boldsymbol{e}$ depends only on the sign pattern of $\boldsymbol{e}$:

Theorem 3.4.2. *If for some $\boldsymbol{e}' \in \mathbb{R}^n$, we have*

$$\boldsymbol{e}' = \arg\min_{\boldsymbol{e}} \|\boldsymbol{e}\|_1 \quad \textit{subject to} \quad B\boldsymbol{e} = B\boldsymbol{e}', \tag{3.21}$$

then for all $\tilde{\boldsymbol{e}}$ such that $sign(\tilde{e}_i) = sign(e'_i)$, $i = 1 \dots n$,

$$\tilde{\boldsymbol{e}} = \arg\min_{\boldsymbol{e}} \|\boldsymbol{e}\|_1 \quad \textit{subject to} \quad B\boldsymbol{e} = B\tilde{\boldsymbol{e}}.$$

From this result, to determine whether we can recover any T-sparse signal $\boldsymbol{e}$ (i.e. $\|\boldsymbol{e}\|_0 = T$), we only need to check one $\boldsymbol{e}$ for each T-sparse sign pattern. Specifically:

$$T^* = \min \left\{ T \in \mathbb{N} \,\middle|\, \exists \boldsymbol{e}' \in E_T : \boldsymbol{e}' \neq \arg\min_{\boldsymbol{e} | B\boldsymbol{e} = B\boldsymbol{e}'} \|\boldsymbol{e}\|_1 \right\} \tag{3.22}$$

where $E_T \doteq \{\boldsymbol{e} \in \mathbb{R}^m \mid \forall i : e_i \in \{-1, 0, 1\},\ \|\boldsymbol{e}\|_0 = T\}$.

Since $|E_T| = 2^T \binom{m}{T}$, a straightforward algorithm for computing (3.22) requires time

$$\sum_{T=1}^{T^*} 2^T \binom{m}{T} t_{lp}(m \times p) \tag{3.23}$$

where t_{lp} is the time it takes to solve the linear programming problem (3.21). We note that instead of actually solving for the right-hand side of (3.21), one can check if $\boldsymbol{e}'$ minimizes the right-hand side by looking for appropriate sub-gradients (see [6, §II]). This alternative approach, however, still requires solving a linear programming problem of similar size.

It is easy to see that the running time of our algorithm (3.20) is exponentially faster than the alternative (3.23) when n/m is small compared to T^*/m (i.e. A is very tall) or when n/m is very close to one (i.e. A is almost square). The first case is precisely the interest of robust estimation – the number of measurements needs to be large so as to tolerate more errors. This is the case for the robust state estimation problem one often encounters in control systems.

3.5 Comparison to Other Robust Estimators

In this section we compare the Minimum Sum of Distances (MSoD) estimator to other typical robust estimation schemes in the literature.

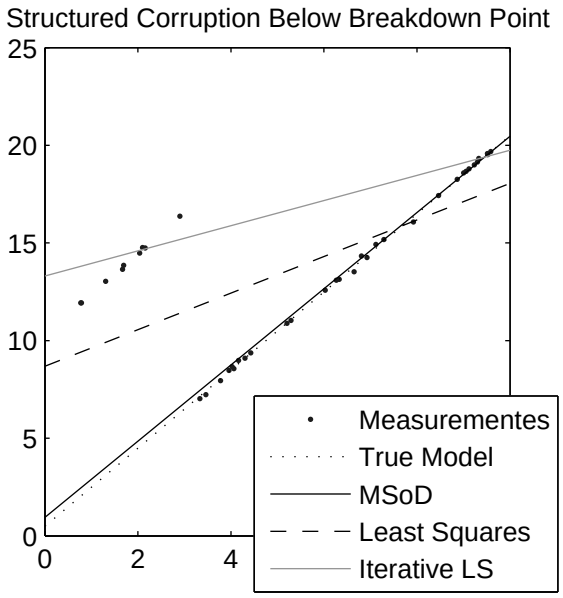

Figure 3.2: We attempt to estimate a line model from which 40 noisy and corrupted points are drawn. The breakdown point of the MSoD estimator is 10 points. Corrupting the 10 leftmost points corresponds to the worst case in which the MSoD will fail. In the example shown here we corrupted only the 9 leftmost points. Shown in the plot are the initial model estimated using least-squares for all the points, the model estimated by the iterative least-squares method, and that estimated by the MSoD. We can see that the MSoD works well, but the iterative trimming method, labeled "iterative LS," fails to converge to a good model.

3.5.1 Iterative Trimming

Arguably, this is the simplest robust estimator. Its application involves calculating an estimate using all (noisy and corrupted) measurements, say by least squares in our case. After discarding a certain number of measurements which are most inconsistent with the estimate, one recomputes the estimate using the remaining measurements. One may iterate the above process until only a predefined number of measurements remains, or until the residual error of the remaining measurements drops below some predefined level.

The main drawback of this method is that for certain corruption, the initial estimate from all the data can be made to favor some of the corrupted measurements over the uncorrupted measurements. We are not aware of any work that carefully analyzes the breakdown point of such an iterative method. However, we found that we can make this method fail using far fewer corrupted measurements than the breakdown point calculated for the MSoD estimator. Figure 3.2 shows a simple example in which the iterative least squares method fails but MSoD succeeds.

3.5.2 Random Sampling

Another very popular approach to obtain robust estimate is through the Random Sampling Consensus (RANSAC) method [15]. In our context, this corresponds to randomly selecting n of the m measurements (equations) and solving $\boldsymbol{x}$. One then checks how many other measurements are consistent with this estimate; say error incurred is below some level. The algorithm repeatedly select sets of n measurements until an estimate with high consensus is obtained. In theory, this approach has a breakdown point of 50%.

With p randomly selected sets of n measurements, the probably that at least one set contains no corrupted measurements at all is $1-(1-q^n)^p$ where q is the percentage of uncorrupted points. When n is small, this probability of success can be very high with relatively small number of selections – the reason why RANSAC has been very popular among practitioners. However, ensuring a fixed probability of success requires that the number of selections p grows *exponentially* in n, making it utterly inefficient when the dimension n is high. Linear programming solvers which minimize the MSoD cost function, on the other hand, require time polynomial in the size of the matrix A. Hence, MSoD is more scalable than RANSAC in dimension n, despite a lower breakdown point.[2]

3.6 Application - Vehicle Position Estimation

In this subsection we present a "real-life" application that demonstrates the potential benefits of the Minimum Sum of Distances Estimator (MSoD). The problem which we address is estimating the position, orientation and velocity of a vehicle moving in 2D. The vehicle has inertial navigation sensors (gyroscopes) that generate noisy measurements of its velocity v and its rate of orientation change $\dot{\theta}$. In addition, the vehicle receives noisy measurements of its east, e, and north, n, positions. A typical source for such measurements is a GPS system, which may produce corrupted or erroneous measurements due to multi-paths. The inertial measurements are generated every t_s seconds, while the position measurements are generated every T_s seconds, with $t_s \ll T_s$.

Given the car state at time t_0, its position at time t_1 is

$$\begin{aligned} e(t_1) &= e(t_0) + \int_{t_0}^{t_1} \cos\theta(\tau)\, v(\tau)\, d\tau \\ n(t_1) &= n(t_0) + \int_{t_0}^{t_1} \sin\theta(\tau)\, v(\tau)\, d\tau. \end{aligned}$$

Denote by $\hat{\cdot}$ our estimate of the car state and by $\boldsymbol{x} = \left(e-\hat{e}, n-\hat{n}, \theta-\hat{\theta}, v-\hat{v}\right)^T$ our (presumably small)

[2] It has been shown in the literature that for randomly generated A, the breakdown point of MSoD grows linearly in m [6, 14]. However, the fraction is normally bounded from above by 1/3.

estimation error. Denote by g_e, g_n the position measurements and by $\boldsymbol{y}(t) \doteq \left(y_0^T(t), \ldots, y_d(t)^T\right)^T$ the measurement residuals over a dT_s-time period, where

$$y_k(t) \doteq \begin{pmatrix} g_e(t+kT_S) - \hat{e}(t+kT_S) \\ g_n(t+kT_S) - \hat{n}(t+kT_S) \end{pmatrix} \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \boldsymbol{x}(t+kT_s) \doteq C\boldsymbol{x}_k(t).$$

Based on our assumptions, we can write

$$\begin{aligned} \boldsymbol{x}(t+T_s) &\approx \boldsymbol{x}(t) + \begin{pmatrix} \int_t^{t+T_s} \cos\theta(\tau)\, v(\tau)\, d\tau - \int_t^{t+T_s} \cos\hat{\theta}(\tau)\, \hat{v}(\tau)\, d\tau \\ \int_t^{t+T_s} \sin\theta(\tau)\, v(\tau)\, d\tau - \int_t^{t+T_s} \sin\hat{\theta}(\tau)\, \hat{v}(\tau)\, d\tau \\ 0 \\ 0 \end{pmatrix} \\ &\approx \begin{pmatrix} 1 & 0 & \int_t^{t+T_s} -\sin\theta(\tau)\, v(\tau)\, d\tau & \int_t^{t+T_s} \cos\theta(\tau)\, d\tau \\ 0 & 1 & \int_t^{t+T_s} \cos\theta(\tau)\, v(\tau)\, d\tau & \int_t^{t+T_s} \sin\theta(\tau)\, d\tau \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \boldsymbol{x}(t) \doteq F(t)\, \boldsymbol{x}(t) \end{aligned}$$

so that

$$\boldsymbol{y}(t) \approx \begin{pmatrix} C \\ CF(t) \\ \vdots \\ CF(t+(d-1)T_s) \ldots F(t+T_s) F(t) \end{pmatrix} \boldsymbol{x}(t) \doteq A(t)\, \boldsymbol{x}(t). \tag{3.24}$$

The approximations are due to the linearization of the nonlinear relation between the presumably small estimation error and the measurement residuals, and due to the noise and corruptions of the measurements.

Equation (3.24) is the linear model on which we apply our estimation scheme. Every time a new position measurement is generated we use it together with the last d position measurements to correct the vehicle estimated state. The matrix $A(t)$ and the estimated expected positions in the $\boldsymbol{y}$ vector are regenerated every time a new position measurement arrives to reflect our best estimate so far.

Simulation results are given in Figure 3.3. In this simulation, the breakdown point, calculated by Algorithm 5, ranges from 4 to 6, depending on the vehicle maneuvers. While the number of corrupted measurements occasionally exceeded the breakdown point, the results were still remarkably good. This is because the breakdown point represents a worst case scenario whose probability is relatively low. For comparison we also show in Figure 3.3 simulation results when a standard nonlinear Kalman filter was used for this system.

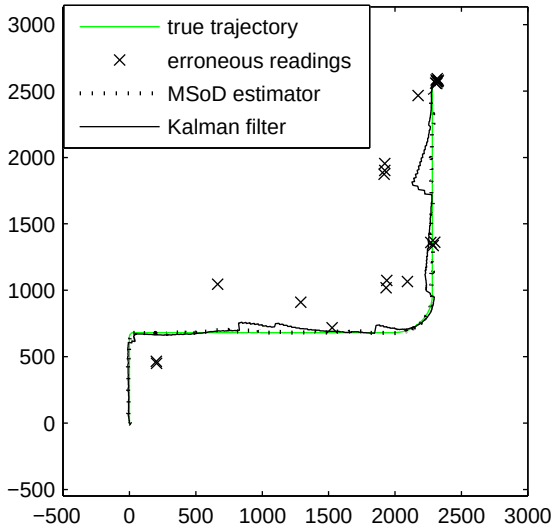

Figure 3.3: Estimating a vehicle position which is moving in a 2D plane from noisy and corrupted measurements. The MSoD estimation scheme was applied on the linear model (3.24) using $d = 19$. The units in the plot are meters. The car average velocity is $85_{km/h}$. New position measurements are generated every $T_s = 1$ seconds. Uncorrupted position measurements have noise with 10_m standard deviation. Corrupted measurements have errors which are uniformly distributed up to 400_m. The system has a 0.06 (6%) probability of switching from an uncorrupted to a corrupted mode, and a 0.5 probability of switching from a corrupted mode to an uncorrupted mode. The maximum and the average magnitude of the position errors were 55_m and 9_m, respectively. For comparison we also show the results of using standard nonlinear Kalman filter. The standard deviation of the position errors, used to calculate the Kalman gains, was 200_m. The maximum and the average magnitude of the position errors were 157_m and 30_m, respectively.

Chapter 4

Adaptive Control using Quantized Measurements with Application to Vision-only Landing Control

4.1 Introduction

In this chapter we focus on fixed output quantization, for example due to sensors of limited resolution, when some of the plant model parameters are unknown. When the plant dynamics are unknown and system identification, but not stabilization, is desired, we mention [68], [23], and [73] among others which address the issue of quantization. Surprisingly, very few papers deal with the problem of stabilization when the plant dynamics are unknown, despite the prevalence of this problem in many control applications. Two papers, [21] and [67], consider input quantization. The assumption taken by these papers, of input quantization with deterministic quantizers, makes a fundamental difference from the settings of output quantization: With the controller knowing the input acting on the plant, a certainty equivalence principle separates the estimation of the plant dynamics from the effects of quantization. In Chapter 2 we proved robustness to variations in the plant dynamics using a specific dynamic quantization scheme. As the actual plant dynamics are not estimated, this requires the variation of the plant dynamics from some nominal model to be sufficiently small. Using supervisory control, [74] switches between several controllers, finds the one that best approximates the actual plant dynamics, and uses that one to stabilize the system. Finally, [66] achieves reference tracking for open-loop stable systems with input and state quantization, where the unknown plant dynamics enter as a linear feedback gain.

To generate the stabilizing control input, or to identify system parameters, an estimate of the state needs to be extracted from the quantized measurements. The basic approach is to select the center of the quantization regions as the state estimate. This was the approach followed by all the references cited above, including proofs of convergence for example in [68]. However, as we show in this chapter, the convergence may be too slow and a more careful treatment of the quantized measurements, with their special characteristics, needs to be employed in selecting the state estimate. Here we follow up on the approached proposed in [49] in the context of system identification. The first novelty in this chapter is an alternative computational approach for solving the optimization problem that selects the state estimate.

The second novelty is the development of a simulation, on which this new approach can be tested, of a vision-based control problem: controlling a fixed-wing airplane to follow a gliding path on approach to landing. The only available feedback for the landing controller is a camera mounted on the airplane and focused on the runway. The source of quantization in this problem is the pixelization of the image. We believe that the development of a simulation platform for this problem has merits of its own in exposing the issue of quantization in vision-based control, and in the ability to compare different approaches for addressing this issue. As an application, we expect the settings we consider here to be applicable to a small unmanned aerial vehicle (UAV) where one may want to avoid installing gyroscopes due to cost and weight considerations. For a complete flight control system, though, the addition of at least an airspeed indicator to our settings would be required to measure this critical quantity that is not observable using camera measurements.

For the sake of completeness we cite [3], [42] and [51] as some of the other works which address the problem of landing by vision only. As each of these studies uses different settings and a different set of assumptions, we cannot compare their results to ours. We do not cite other studies where vision is only used for guidance and the aircraft stabilization is accomplished using inertial sensors (gyroscopes).

The outline of this chapter is as follows. In §4.2 we formulate the general problem we address in this chapter and propose an algorithm for solving the problem. In §4.3 we provide a convergence result for the proposed algorithm. Starting from §4.4 we focus on the specific application. In §4.4 we recall the longitudinal dynamics of an airplane and derive a reduced order model. In §4.5 we make the connection between the camera input and the state of the airplane. In §4.6 we derive the linear model for the airplane which is consistent with the problem we formulated in §4.2. In §4.7 we provide details regarding the implementation of the controller including, in particular, the control law. And finally, in §4.8 we present simulation results. Additional details regarding the simulation are provided in §4.9 and §4.10.

4.2 Problem Formulation and Estimation Method

Consider a control system which consists of a plant, a quantizer, and a controller. The plant is assumed to be linear time-varying (LTV):

$$\begin{aligned} \dot{x}(t) &= A(t,a)\,x(t) + B(a)\,u(t) + D(a) \\ y(t) &= Cx(t) \qquad\qquad z(t) = E(t,a) \end{aligned} \tag{4.1}$$

where $x(t) \in \mathbb{R}^n$ is the state of the plant, $u(t) \in \mathbb{R}^m$ is the control input, $y(t) \in \mathbb{R}^p$ $(p \leq n)$ is the output signal which is sampled by the quantizer, $z(t) \in \mathbb{R}^q$ is an uncontrollable signal that is sampled by the quantizer and

assists in estimating the model parameters, and $a \in \mathbb{R}^l$ is the vector of (the constant) model parameters. The matrices $A(t,a)$, $B(a)$, C, $D(a)$ and $E(t,a)$ are of appropriate dimensions. Their structure, as a function of the model parameters, a, and possibly of the time, is known, but the model parameters themselves are unknown.

The quantizer samples the signals y and z once every τ seconds, and for each individual component, y_i and z_j, sends the controller an interval which contains the value of this component. Thus up to time t, assuming the sampling started at $t = 0$, the information available to the controller is the lower and upper bounds, $\underline{y}_i(k\tau), \overline{y}_i(k\tau)$, $i = 1,\ldots,p$, $k = 0,\ldots,\lfloor t/\tau \rfloor$ as well as $\underline{z}_j(k\tau), \overline{z}_j(k\tau)$, $j = 1,\ldots,q-p$, $k = 0,\ldots,\lfloor t/\tau \rfloor$, such that for each i, j, and k: $\underline{y}_i(k\tau) \le y_i(k\tau) \le \overline{y}_i(k\tau)$ and $\underline{z}_i(k\tau) \le z_i(k\tau) \le \overline{z}_i(k\tau)$. The controller estimates the state and the model parameters based on this information, and then uses that estimate to generate the control input that will drive the system to some desired steady-state value. The estimation, independent of the law by which the control input is generated, will be discussed in this section. In §4.7 we will provide an example of a control law.

By restricting to piecewise constant control input such that $\forall k \in \mathbb{N}$: $u(t) = u(k\tau)$ $\forall t \in [k\tau, (k+1)\tau)$, the continuous plant dynamics (4.1) can be written in discrete form as

$$\begin{aligned} x((k+1)\tau) &= A_k(a)\,x(k\tau) + B_k(a)\,u(k\tau) + D_k(a) \\ y(k\tau) &= Cx(k\tau) \qquad\qquad z(k\tau) = E_k(a) \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} A_k(a) &= \Phi(k\tau+\tau, k\tau), \qquad B_k(a) = \int_0^\tau \Phi(k\tau+\tau, k\tau+s')\,ds' B(a), \\ D_k(a) &= \int_0^\tau \Phi(k\tau+\tau, k\tau+s')\,ds' D(a), \end{aligned}$$

and $\Phi(t,t_0)$ is the state transition matrix for $\dot{x} = A(t,a)\,x$. To avoid the computation of $\Phi(k\tau+\tau, k\tau)$, in many cases an approximation as the one we use in §4.6 will be sufficient.

Define

$$\begin{aligned} U_{k+1} &\doteq \left[u^T(0), u^T(\tau), \ldots, u^T((k-1)\tau)\right]^T \\ Y_{k+1} &\doteq \left[y^T(0), z^T(0), y^T(\tau), \ldots, y^T(k\tau), z^T(k\tau)\right]^T, \end{aligned}$$

and consider

$$\mathscr{E}\left(k, Y_k, U_k, a\right) = \mathfrak{A}\left(k\tau, a\right) \begin{bmatrix} y\left(\left(k-r\right)\tau\right) \\ \vdots \\ y\left(\left(k-1\right)\tau\right) \end{bmatrix} + \mathfrak{B}\left(k\tau, a\right) \begin{bmatrix} u\left(\left(k-r\right)\tau\right) \\ \vdots \\ u\left(\left(k-1\right)\tau\right) \end{bmatrix} + \mathfrak{D}\left(k\tau, a\right) - y\left(k\tau\right)$$

where r is the observability index (assumed to be constant) for the system $(A_k,\, C)$, and

$$\mathcal{O}_{k-r}\left(a\right) = \begin{bmatrix} C \\ CA_{k-r}\left(a\right) \\ \cdots \\ C\prod_{i=k-r}^{k-2} A_i\left(a\right) \end{bmatrix} \qquad \mathcal{O}^{\dagger} = \left(\mathcal{O}^T\mathcal{O}\right)^{-1}\mathcal{O}^T$$

$$\mathfrak{A}\left(k\tau, a\right) = C\prod_{i=k-r}^{k-1} A_i\left(a\right)\mathcal{O}_{k-r}\left(a\right)^{\dagger}$$

$$\mathfrak{B}\left(k\tau, a\right) = \left[C\prod_{i=k-r+1}^{k-1} A_i\left(a\right)B_{k-r}\left(a\right)\ \ldots\ CB_{k-1}\left(a\right)\right] -$$

$$C\prod_{i=k-r}^{k-1} A_i\left(a\right)\mathcal{O}_{k-r}\left(a\right)^{\dagger} \times \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ CB_{k-r}\left(a\right) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ C\prod_{i=k-r+1}^{k-2} A_i\left(a\right)B_{k-r}\left(a\right) & \cdots & CB_{k-2}\left(a\right) & 0 \end{bmatrix}$$

$$\mathfrak{D}\left(k\tau, a\right) = \sum_{i=k-r+1}^{k} C\prod_{j=i}^{k-1} A_j\left(a\right)D\left(a\right) -$$

$$C\prod_{i=k-r}^{k-1} A_i\left(a\right)\mathcal{O}_{k-r}\left(a\right)^{\dagger} \times \begin{bmatrix} 0 \\ CD\left(a\right) \\ \vdots \\ \sum_{i=k-r+1}^{k-1} C\prod_{j=i}^{k-2} A_j\left(a\right)D\left(a\right) \end{bmatrix}.$$

Note that when the values in Y_k and U_k follow the dynamics in (4.2), $\mathscr{E}\left(k, Y_k, U_k, a\right) = 0$. We refer to $\mathscr{E}\left(k, Y_k, U_k, a\right)$ as the modeling error.

We define the cost function,

$$f\left(Y_N, a\right) \doteq \sum_{k=r}^{N} \left\|\mathscr{E}\left(k, Y_N, U_N, a\right)\right\|_2^2 + \sum_{k=0}^{N-1} \left\|E\left(k\tau, a\right) - z\left(k\tau\right)\right\|_2^2,$$

and propose the following minimization problem in order to estimate both the state and the model param-

eters:

$$\begin{aligned}
&\min_{a,Y_N} f\left(Y_N, a\right) \\
&\text{subject to } \forall k \in [0, \ldots, N-1]: \\
&\qquad \underline{y}_i\left(k\tau\right) \leq y_i\left(k\tau\right) \leq \overline{y}_i\left(k\tau\right) \quad \forall i \in [1, \ldots, p] \\
&\qquad \underline{z}_i\left(k\tau\right) \leq z_i\left(k\tau\right) \leq \overline{z}_i\left(k\tau\right) \quad \forall j \in [1, \ldots, q-p]
\end{aligned} \tag{4.3}$$

where N is the number of quantized output measurements we collected. Problem (4.3) is a constrained nonlinear minimization problem. We solve it using the following iterative algorithm:

1. Initialize the algorithm by selecting an arbitrary Y_N that satisfies the inequality constraints in (4.3).
2. Fix Y_N and find a that minimizes the cost function $f\left(Y_N, a\right)$.
3. Increase N if more measurements are available.
4. Fix a and find Y_N that minimizes the cost function $f\left(Y_N, a\right)$ and satisfies the inequality constraints in (4.3).
5. Repeat from step 2.

When a is fixed, minimizing over Y_N becomes a constrained quadratic programming problem for which there exist computationally efficient solvers. Minimizing over a when Y_N is fixed can still be a nonlinear minimization problem, but it is now unconstrained, and it has a fixed (small) number of variables that does not grow with N. In many cases, as in the problem we address below, we can derive the second derivative, the Hessian, explicitly and solve the minimization problem efficiently using general purpose nonlinear solvers.

4.3 Proof of Convergence

Because $f\left(Y, a\right)$ is quadratic in Y, we can rewrite (4.3), for fixed N and U_N, as

$$\begin{aligned}
&\min_{a,Y} \quad \left\|Q\left(a\right)Y - r\left(a\right)\right\|_2^2 \\
&\text{subject to} \quad \underline{Y}_i \leq Y_i \leq \overline{Y}_i \quad \forall i \in [1, \ldots, n]
\end{aligned} \tag{4.4}$$

where $Y \in \mathbb{R}^n$, $Q : \mathbb{R}^l \to \mathbb{R}^{m \times n}$, $r : \mathbb{R}^l \to \mathbb{R}^m$ (the n and m defined in this section are different from the n and m defined in the previous section). Note that $m < n$. In proving convergence of our proposed iterative

algorithm, we will refer to (4.4) as the problem being minimized. We note that the proof below is applicable to a fixed number of measurements. More work is needed to derive results applicable to a growing number of measurements.

We define $\mathcal{Y}$ to be the set of Y's satisfying the inequality constraints in (4.4). We say that a is a *critical point of f for a given Y* if

$$\frac{\partial f(Y,a)}{\partial a_i} = 0, \qquad \forall i \in \{1,\ldots,l\}. \tag{4.5}$$

We say that Y is a *critical point of f for a given a* if

$$\begin{aligned} \frac{\partial f(Y,a)}{\partial Y_i} &\geq 0 \text{ if } Y_i = \underline{Y}_i, \\ \frac{\partial f(Y,a)}{\partial Y_i} &= 0 \text{ if } \underline{Y}_i < Y_i < \overline{Y}_i \\ \frac{\partial f(Y,a)}{\partial Y_i} &\leq 0 \text{ if } Y_i = \overline{Y}_i \end{aligned} \qquad \forall i \in \{1,\ldots,n\}. \tag{4.6}$$

And we say that (Y,a) is a *critical point of f* if both (4.5) and (4.6) hold. We define σ as the function that maps each $Y \in \mathcal{Y}$ to the set of critical points of f given Y.

Consider the following assumptions:

1. The functions $Q(\cdot)$ and $r(\cdot)$ are continuous.
2. Let (Y_k, a_k) and (Y_{k+1}, a_{k+1}) be the estimated values before and after iteration k of the algorithm. Then for every k: $a_{k+1} \in \sigma(Y_k)$, $f(Y_k, a_{k+1}) < f(Y_k, a_k)$ if $a_{k+1} \neq a_k$, Y_{k+1} is a critical point of f given a_{k+1}, and $f(Y_{k+1}, a_{k+1}) < f(Y_k, a_{k+1})$ if $Y_{k+1} \neq Y_k$.
3. There exists $K \in \mathbb{N}$ such that the number of critical points of f given Y, $\sigma(Y)$, is smaller than K for every $Y \in \mathcal{Y}$, and furthermore, $\sigma(\cdot)$ is continuous.

We now can state the following convergence result:

Theorem 4.3.1. *Given that assumptions 1-3 hold, and following the iterative algorithm described above, the series (Y_k, a_k) converges to a set M of critical points of f.*

To prove Theorem 4.3.1, we need the following results and definitions.

Definition 4.3.1. A set valued map $\Gamma(\cdot) : \mathbb{R}^k \to 2^{\mathbb{R}^n}$ is said to be:

1. *closed* at t_0 if, for any sequence $\{t_i\}$ and $\{x_i\}$, $t_i \to t_0$, $x_i \in \Gamma(t_i)$, $x_i \to x_0$ imply $x_0 \in \Gamma(t_0)$

2. *upper semicontinuous* at t_0 if, for any open set Ω containing $\Gamma(t_0)$, there exists an $\epsilon = \epsilon(\Omega) > 0$ such that $\Gamma(t) \subset \Omega$ for any $t \in \left\{ t \in \mathbb{R}^k \,|\, \|t - t_0\| \leq \epsilon \right\}$.

Lemma 4.3.2 ([2, Theorem 2.2])**.** *Consider the following set valued map:*

$$S(t) \doteq \begin{array}{ll} \arg\min\limits_{x} & c(t)^T x + \frac{1}{2} x^T \bar{Q}(t) x \\ s.t. & Ax \leq b \end{array} \tag{4.7}$$

where $t \in \mathbb{R}^k$, $c(t)$ and $\bar{Q}(t)$ are continuous at $t = 0$, and $c(0) = c$, $\bar{Q}(0) = \bar{Q}$. We suppose that $\bar{Q}(t)$ is a symmetric, positive-semidefinite $n \times n$ matrix for each $t \in \mathbb{R}^k$. Assume that the following two conditions hold: (1) $\nexists s \in \mathbb{R}^n \setminus 0$ such that $As \leq 0$, $c^T s \leq 0$ and $\bar{Q}s = 0$; (2) $\nexists w \in \mathbb{R}^n \setminus 0$ such that $A^T w = 0$, $b^T w \leq 0$ and $w \geq 0$. Then S is closed and upper semicontinuous at $t = 0$.

Remark 4.3.1. The quadratic problem (4.7) relates to (4.4) with $c(t) = r(a)^T Q(a)$, $\bar{Q}(t) = Q(a)^T Q(a)$, and with A consisting of pairs of rows, ϵ_i^T, $-\epsilon_i^T$, $i = 1, \ldots, n$, where we use ϵ_i to denote the i'th canonical vector: $(\epsilon_i)_j = 0 \; \forall j \neq i$ and $(\epsilon_i)_i = 1$.

Definition 4.3.2. An algorithm, or a map, $T : \mathbb{R}^n \to \mathbb{R}^n$, is said to be closed at $P \in \mathbb{R}^n$ if for all convergent sequences $P_k \to P$, $P'_k \to P'$ such that $P'_k \in T(P_k)$, one has that $P' \in T(P)$. An algorithm is said to be closed on $W \subset \mathbb{R}^n$ if it is closed at P, for all $P \in W$.

Definition 4.3.3. A function $V : W \to \mathbb{R}_{\geq 0}$ is a Lyapunov function for an algorithm (a map) $T : \mathbb{R}^n \to \mathbb{R}^n$ on $W \subset \mathbb{R}^n$ if V is continuous on W and $V(P') \leq V(P)$ for all $P' \in T(P)$ and all $P \in W$.

Definition 4.3.4. A set C is said to be weakly positively invariant with respect to an algorithm T if for any $P_0 \in C$ there exists $P \in T(P_0)$ such that $P \in C$.

Lemma 4.3.3 ([8, Theorem C.1])**.** *Let T be a closed algorithm on $W \subset \mathbb{R}^N$ and let V be a Lyapunov function for T on W. Let $x_0 \in W$ and assume the sequence $\{x_n \,|n \in \mathbb{N} \cup \{0\}\}$ defined via $x_{n+1} \in T(x_n)$ is in W and bounded. Then there exists $c \in \mathbb{R}$ such that*

$$x_n \to M \cap V^{-1}(c), \tag{4.8}$$

where M is the largest weakly positively invariant set contained in

$$\left\{ x \in \mathbb{R}^N \,|\exists y \in T(x) \;\; \text{such that } V(y) = V(x) \right\} \cap \overline{W}. \tag{4.9}$$

Proof of Theorem 4.3.1: We start by showing that the conditions stated in Lemma 4.3.2 hold for (4.4). Writing (4.4) as (4.7), the functions $c(a)$ and $\bar{Q}(a)$ are continuous $\forall a$ due to Assumption 1. From Remark 4.3.1 $\bar{Q}$ is symmetric and positive-semidefinite. Also from Remark 4.3.1, the only vector s satisfying $As \leq 0$ is the zero vector, so condition (1) holds. From the structure of A, a vector w satisfies $A^T w = 0$ if and only if $w_{2i-1} = w_{2i}$, $\forall i \in \{1, \ldots, n\}$. For each $i \in \{1, \ldots, n\}$, $-b_{2i-1}$ and b_{2i} correspond to one quantized measurement, where $-b_{2i-1}$ is the lower bound and b_{2i} is the upper bound of that measurement. Thus $-b_{2i-1} < b_{2i}$. Multiplying $b^T w$, if $w \geq 0$ and $w \neq 0$ then we must have $b^T w > 0$. Therefore condition (2) also holds. Applying Lemma 4.3.2 we conclude that the second step in each iteration of our algorithm is a closed function. The first step is continuous due to Assumption 3. Let $Y_k \to Y$, $a_k \to a$ and $Y'_k \to Y$, $a'_k \to a$ be two convergent sequences such that $a'_k \in \arg\min_a f(Y_k, a)$ and $Y'_k \in \arg\min_{Y \in \mathcal{Y}} f(Y, a'_k)$. Then $a' \in \arg\min_a f(Y, a)$ and $Y' \in \arg\min_{Y \in \mathcal{Y}} f(Y, a')$ and we conclude that our algorithm is closed.

Since in each step that the algorithm computes, the value of $f(Y, a)$ does not increase, f serves as a Lyapunov function for the algorithm. As $\sigma(\mathcal{Y})$ is compact due to Assumption 3 and the compactness of $\mathcal{Y}$, the sequence (Y_k, a_k) is in the compact set $W \doteq \mathcal{Y} \times \sigma(\mathcal{Y})$. Thus we can apply Lemma 4.3.3 and conclude the existence of $c \in \mathbb{R}$ such that $(Y_k, a_k) \to M \cap f^{-1}(c)$ where M is the largest weakly invariant set contained in $\overline{W}$ as defined in Lemma 4.3.3. Let $(Y, a) \in M$ and assume it is not a critical point. Then $\forall (Y', a') \in T(Y, a)$, $f(Y', a')$ will be strictly smaller than $f(Y, a)$, contradicting the definition of M. ∎

We find it appropriate to report here an additional result, even though its implication is still under investigation. Set $n_Q = \operatorname{rank} Q$. We say that $Q \in \mathbb{R}^{m \times n}$ is *in general directions* if every n_Q columns of Q are linearly independent. We also define $V(a) \doteq \min_{Y \in \mathcal{Y}} f(Y, a)$, and $\mathcal{P}_0 \doteq \left\{ a \in \mathbb{R}^l \,|\, V(a) = 0 \right\}$ ($\mathcal{P}_0$ might be an empty set). Finally we recall the definition of B(ouligand)-derivative:

Definition 4.3.5.

$$F'(x, v): \quad \frac{F(y) - F(x) - F'(x, y - x)}{\|y - x\|} \to 0 \quad as \quad y \to x,\ y \neq x. \tag{4.10}$$

When it exists, the B-derivative equals the directional derivative $F'(a; v) \doteq \lim_{\varepsilon \searrow 0^+} \frac{F(a + \varepsilon v) - F(a)}{\varepsilon}$. The importance of the B-derivative, in our context, is that it satisfies the chain rule, [53, Corollary A.4], whereas the direction derivative does not. It is a strictly weaker notion than the Fréchet derivative, but its existence does not imply, nor is it implied by, the existence of the Gâteaux derivative.

Proposition 4.3.1. *With the additional assumption that the functions $Q(\cdot)$ and $r(\cdot)$ are C^2 (twice continuously differentiable), the vector of model parameters a in the iterative algorithms described above converges to a set $M \subset \mathbb{R}^l$ for which one of the following holds: (1) $M \subseteq \mathcal{P}_0$; (2) $Q(M)$ contains matrices not in general*

directions; (3) for every $a \in M$, the B-derivative $F'(a;v)$ exists and satisfies $F'(a;v) \geq 0$, $\forall v \in \mathbb{R}^l \setminus 0$.

Due to disturbances and nonlinearity in the system, and the zero measure of the set of matrices not in general directions in $\mathbb{R}^{m\times n}$, we expect the third case to be the prevailing one.

For the proof of Proposition 4.3.1, we recall the following result:

Lemma 4.3.4 ([52, Theorem 2])**.** *Considering* (4.7) *and a pair t and $x \in S(t)$, define P to be the set of i's such that $A_i x = b_i$ and assume that the following hold:*

1. *$\bar{Q}(t)$ and $c(t)$ are twice continuously differentiable.*
2. *There exists $v \in \mathbb{R}^n$ such that $A_i v < 0$ for any $i \in P$.*
3. *For every $v \in \mathbb{R} \setminus 0$ such that $\epsilon_P^T v = 0$, $v^T \bar{Q} v > 0$.*

Then for some neighborhoods U of t and V of x, there is a locally Lipschitz and B-differentiable function $y(\cdot)$ mapping U to V such that for each $t \in U$, $y(t)$ is the unique local solution of $S(t)$.

Proof of Proposition 4.3.1: By proposition 4.3.1 the algorithm converges to a set M which is contained in $f^{-1}(c)$ for some $c \in \mathbb{R}_{\geq 0}$. If $c = 0$ then we get that $M \subseteq \mathcal{P}_0$. Otherwise $c > 0$ and assume that $Q(M)$ does not contain matrices not in general directions. Let $(x,t) = (Y,a) \in M$ such that $x \in S(t)$ as defined in (4.7) and $Q(a)$ is in general directions. We will show that in this case the assumptions in Lemma 4.3.4 hold. By the assumption taken in Proposition 4.3.1, assumption 1 holds. Noting again that A in (4.7) consists of pairs of rows, ϵ_i^T, $-\epsilon_i^T$, $i = 1,\ldots,n$, and that for any such pair, only one row is in P since the lower and upper quantization bounds are never equal, it is straightforward to see that assumption 2 also holds.

Now to assumption 3. We set $Q = Q(a)$ and define $P' = \{i \in \{1,\ldots,n\} \,|\, 2i-1 \in P \text{ or } 2i \in P\}$. Note that the number of elements in P and in P' is the same, again due to the nonzero difference between each pair of lower and upper quantization bounds. First we show that P' contains at least $n - n_Q + 1$ constraints where $n_Q = \operatorname{rank} Q$. By projecting Q and r, if necessary, to a lower subspace, we can assume without loss of generality that Q has full row rank and $Q \in \mathbb{R}^{n_Q \times n}$. Let $P'^c \doteq \{1,\ldots,n\} \setminus P'$ be the complement of P'. We recognize that if $\frac{\partial f(x,a)}{\partial x}\epsilon_i \neq 0$ for some i, then we must have that $i \in P'$ (otherwise we could have decreased the cost function without violating any of the constraints). Therefore $\frac{\partial f(x,a)}{\partial x}\epsilon_{P'^c} = 0$, which implies that $0 \in \operatorname{argmin}_{v\in\mathbb{R}^{|P'^c|}} \|Qx + Q\epsilon_{P'^c} v - r\|_2^2$. By the projection theorem this implies that the vector $u = r - Qx$, which is nonzero since we assumed $f(x,t) \neq 0$, satisfies $u^T Q \epsilon_{P'^c} = 0$. Or in other words, it implies that $\operatorname{rank} Q\epsilon_{P'^c} < n_Q$. By the assumption that Q is in general directions, this implies that $|P'^c| < n_Q$ or $|P'| \geq n - n_Q + 1$.

Let $v \in \mathbb{R}^n \setminus 0$ such that $\epsilon_{P'}^T v = 0$. This implies we can write $v = \epsilon_{P'^c} u$ for some $u \in \mathbb{R}^{|S'^c|} \setminus 0$. Since $|P'^c| < n_Q$, and Q is in general directions, $Q\epsilon_{P'^c}$ has full column rank so that $Qv \neq 0$. Thus we get that assumption 3 from Lemma 4.3.4 holds. We note that in the original statement of [52, Theorem 2] there was an additional assumption (Assumption A4). However, given that the inequality constraints are linear in x and independent of t, this assumption holds trivially.

Applying Lemma 4.3.4 we now conclude that the B-derivative $S'(a; v)$ exists $\forall v \in \mathbb{R}^l$. We can then write $V'(a, v) = \frac{df(S(a),a)}{da} = \frac{\partial f(x,a)}{\partial x} S'(a; v) + \frac{\partial f(x,a)}{\partial a} v$, where $S'(a, v)$ is a feasible direction which does not violate the constraints. Since we already proved that $(x, a) \in M$ is a critical point of f, we get that $V'(a, v) \geq 0$, $\forall v \in \mathbb{R}^l \setminus 0$. ■

We now address the assumptions we made. Assumption 1 holds with many models. Most optimization tools satisfy assumption 2 (ignoring numerical errors). For assumption 3 to hold we need that the number of locally minimizing a's be finite for any $Y \in \mathcal{Y}$. This requires Y to be sufficiently exciting in some sense (depending on the specific model and the quantization). We are still investigating what other conditions need to be satisfied in order to guarantee that assumption 3 holds.

4.4 Airplane Dynamics

We now demonstrate the applicability of the approach we developed in the previous sections to vision-based landing control for a fixed-wing airplane. We consider only the longitudinal dynamics in the vertical plane, and we make the following assumptions. There is no difference in elevation between the two ends of the runway. The aircraft has static stability — the control system consisting of only the pitch and the pitch rate, with all other signals considered as external input, is open-loop stable. The unknown wind velocity has only a fixed (independent of height) horizontal component. There is no thrust (power-off landing). The lift, drag and gravitational forces, associated with the pitch angle required to follow the desired glide slope at the initial velocity, are in balance such that the velocity remains relatively steady. And finally, we assume the airplane starts relatively close to the desired glide slope angle. In the last section we will test our method on the dynamics of a Cessna 172.

This section is divided into two subsections. In the first subsection we state the true dynamics of the airplane. In the second subsection we approximate the true dynamics using an LTV model. We emphasize that while we use the LTV model to design the implementation of our method, we use the true dynamics to test it in simulation.

4.4.1 True Dynamics

By considering only longitudinal dynamics in the vertical plane, we are left with six degrees of freedom describing the motion of the airplane[1]:

$$\left.\begin{matrix} p_x \\ p_z \end{matrix}\right\} \text{— position} \qquad \left.\begin{matrix} v_x \\ v_z \end{matrix}\right\} \text{— velocity} \qquad \left.\begin{matrix} \theta \\ \dot{\theta} \end{matrix}\right\} \text{— pitch angle and pitch rate.} \tag{4.11}$$

All the quantities above are defined in the frame of reference whose origin is fixed at the beginning of the runway. The x-axis positive direction is defined such that the runway is on the positive side of this axis. The z-axis positive direction is up. The control input is the elevator deflection, δ_e, which measures the angular displacement of the elevator from its trim position.

Using the six quantities in (4.11) and several aerodynamics constants, we can derive the equations of motion. These equations can be found in any textbook on flight dynamics, [65] for example. We will use the following rotation matrix in deriving the equations: $R(\phi) \doteq \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$. A standard way of computing the forces acting on the airplane is to first compute the lift and the drag. The lift is the aerodynamic force perpendicular to the relative wind, and the drag is the aerodynamic force parallel to the relative wind. Both forces are assumed to act on the center of lift.

The complete derivation of the equations of motion is as follows:

1. Angle of relative wind, $\varphi = \tan^{-1}\left(v_z / \left(v_x + v_{\text{wind}}\right)\right)$.
2. Angle of attack, $\alpha = \theta - \varphi$.
3. Airspeed, $V_T = \left\| \left[v_x + v_{\text{wind}}, v_z\right]^T \right\|$.
4. Lift coefficient, $C_L = C_{L_\alpha}(\alpha) + C_{L_q}\dot{\theta}\bar{c}/(2V_T) + C_{L_{\delta_e}}\delta_e$.
5. Drag coefficient, $C_D = C_{D_\alpha}(\alpha) + \left|C_{D_{\delta_e}}\delta_e\right|$.
6. Pitch moment coefficient, $C_m = C_{m_\alpha}(\alpha) + C_{m_{\dot{\alpha}}}\dot{\alpha}\bar{c}/(2V_T) + C_{m_q}\dot{\theta}\bar{c}/(2V_T) + C_{m_{\delta_e}}\delta_e$.

Remark: Above C_{L_q}, $C_{L_{\delta_e}}$, $C_{D_{\delta_e}}$, $C_{m_{\dot{\alpha}}}$, C_{m_q} and $C_{m_{\delta_e}}$ are airframe dependent empirically obtained constants; C_{L_α}, C_{D_α}, C_{m_α} are airframe dependent functions of the angle of attack, and are derived from tables of empirically obtained values. Below ρ is the air density which we assume to be constant, and S is the wing area.

[1] The standard state used in the literature on automatic landing includes: true airspeed v_T, angle of attack α, pitch θ, pitch rate $q = \dot{\theta}$, height $h = p_z$, and deviation from the glide slope d [65, p.341]. Knowing the wind, one can translate between this standard state and (4.11).

7. Lift, $L = 0.5 C_L \rho V_T^2 S$.

8. Drag, $D = 0.5 C_D \rho V_T^2 S$.

9. Linear aerodynamic forces, $F = R(\varphi)\left[-D, L\right]^T$.

10. Pitch moment, $M = 0.5 C_m \rho V_T^2 S \bar{c} + \left(R(\theta) D_{cg}\right) \times_y F$ where $\bar{c}$ is the mean aerodynamic chord, D_{cg} is the displacement of the center of gravity from the center of lift, and $\left[x_1, z_1\right]^T \times_y \left[x_2, z_2\right]^T \doteq z_1 x_2 - x_1 z_2$.

11. Linear accelerations, $\dot{v}_x = F_x/W$, $\dot{v}_z = F_z/W - g$, where W is the weight of the aircraft, and g is the acceleration due to gravity.

12. Angular acceleration, $\ddot{\theta} = M/I_{yy}$ where I_{yy} is the moment of inertia around the pitch axis.

13. Linear velocity, $\dot{p}_x = v_x$, $\dot{p}_z = v_z$.

4.4.2 Reduced Order Model

We define the reduced order state: $x \doteq \left(\theta_p, \theta_v, \theta, \dot{\theta}\right)^T$, where $\theta_p \doteq \tan^{-1}\left(p_z/p_x\right) - \gamma_R$, $\theta_v \doteq \tan^{-1}\left(v_z/v_x\right) - \gamma_R$, and γ_R is the (negative) desired glide slope angle. In the literature on airplane dynamics, $\gamma_R + \theta_v$ is referred to as the flight path angle, γ. The motivation for this choice of state is that our main goal is to drive θ_p to zero. The dynamics of θ_p depend strongly on all the signals in x, so we also need to drive these signals to some appropriate values. As the dynamics of θ_p depend on the remaining signals from the full order model, $d = \sqrt{p_x^2 + p_y^2}$ and $V_E = \sqrt{v_x^2 + v_z^2}$, only through multiplications with the states already in x, we exclude explicit reference to these states in the system model that we attempt to control. Note that by our convention,

$$\begin{pmatrix} p_x \\ p_z \end{pmatrix} = d \begin{pmatrix} -\cos\left(\theta_p + \gamma_R\right) \\ -\sin\left(\theta_p + \gamma_R\right) \end{pmatrix}, \qquad \begin{pmatrix} v_x \\ v_z \end{pmatrix} = V_E \begin{pmatrix} \cos\left(\theta_v + \gamma_R\right) \\ \sin\left(\theta_v + \gamma_R\right) \end{pmatrix}.$$

We now rewrite the dynamics for this reduced order state. We start with the θ_p dynamics:

$$\begin{aligned} \dot{\theta}_p &= \frac{1}{1 + \left(p_z/p_x\right)^2} \frac{v_z p_x - v_x p_z}{p_x^2} \\ &= \frac{V_E}{d}\left(\sin\left(\theta_x + \theta_{gs}\right)\cos\left(\theta_v + \theta_{gs}\right) - \sin\left(\theta_v + \theta_{gs}\right)\cos\left(\theta_x + \theta_{gs}\right)\right) \\ &= \frac{V_E}{d}\left(\sin\left(\theta_x\right)\cos\left(\theta_v\right) - \sin\left(\theta_v\right)\cos\left(\theta_x\right)\right) \end{aligned}$$

(the second equality is a trigonometric identity). If we assume that throughout the approach maneuver, θ_v and θ_p remain relatively small and V_E stays relatively fixed, then we can approximate the θ_p dynamics with

$$\dot{\theta}_p \approx \frac{1}{t_f - t}\theta_p - \frac{1}{t_f - t}\theta_v \tag{4.12}$$

where t_f is the time we expect to reach the beginning of the runway.

We continue with the θ_v dynamics. We approximate the factors which depend on the angle of attack, α, in the lift and drag coefficients with linear functions, and neglect the remaining factors due to their relatively small contribution. This results in

$$C_L \approx C_{L,1}\alpha + C_{L,0} \qquad\qquad C_D \approx C_{D,1}\alpha + C_{D,0}.$$

With that we have (we use $\phi \doteq \theta_v + \gamma_R - \varphi$)

$$\begin{aligned}\dot{\theta}_v =& \frac{1}{V_E}\left[-\sin\left(\theta_v + \gamma_R\right), \cos\left(\theta_v + \gamma_R\right)\right] \times \left(R\left(\varphi\right) \frac{\rho V_T^2 S}{2W} \begin{bmatrix} -C_D \\ C_L \end{bmatrix} - \begin{bmatrix} 0 \\ g \end{bmatrix} \right) \\ \approx& \frac{\rho V_T^2 S}{2V_E W}\left(-\sin\left(\phi\right) C_{D,1} + \cos\left(\phi\right) C_{L,1}\right)\left(\theta - \varphi\right) + \frac{\rho V_T^2 S}{2V_E W}\left(-\sin\left(\phi\right) C_{D,0} + \cos\left(\phi\right) C_{L,0}\right) - \frac{g\cos\left(\gamma_R\right)}{V_E}\end{aligned}$$

where we also used the approximation $\cos\left(\theta_v + \gamma_R\right) \approx \cos\left(\gamma_R\right)$ assuming θ_v is relatively small. In the windless case, $\phi = 0$ as $\varphi = \theta_v + \gamma_R$, and it is easy to see how a linear model can be derived:

$$\dot{\theta}_v \approx -C_{v\to v}\theta_v + C_{p\to v}\left(\theta - \theta_0\right) \tag{4.13}$$

where $C_{v\to v}$, $C_{p\to v}$ and θ_0 are considered constants. They do in fact depend on V_T as well as on other environmental variables such as the air pressure and the weight of the aircraft, but we assume that all of these variables (including in particular V_T) change only slightly throughout the approach maneuver. We claim that the linear model (4.13) is still a good approximation even when there is a wind, where now the three constants just mentioned also depend on the wind speed.

We finish with the angular acceleration, $\ddot{\theta}$, dynamics. We see that M is the sum of two terms, one which depends on the pitch moment coefficient, and one due to the linear aerodynamic forces. Since we found that the second term is small compared to the first, we approximate the angular acceleration dynamics without it. If we further approximate $C_{m/\alpha}\left(\alpha\right) \approx C_{m,1}\alpha + C_{m,0}$, then it is not hard to see that in the windless case

the angular acceleration can be written as

$$\ddot{\theta} \approx C_{v\to\dot{\theta}}\theta_v + C_{\theta\to\dot{\theta}}\theta + C_{\dot{\theta}\to\dot{\theta}}\dot{\theta} + C_{\delta_e\to\dot{\theta}}\left(\delta_e - \delta_0\right). \tag{4.14}$$

And again, we claim that the linear model (4.14), with different constants, is still a good approximation even when there is a wind.

4.5 Camera Feedback

We assume a runway recognition algorithm provides the information about the rows corresponding to different points on the runway, and the width of the runway at these points. All the information is given in pixels, but knowing the parameters of the camera and the angle at which it is installed on the aircraft, we can easily translate the rows into angles in the vertical plane from any reference axis fixed to the aircraft. The reference axis we use is the longitudinal axis, which is also used to define the pitch angle as the angle between this axis and the plane tangent to Earth's surface. Points below the reference axis will be associated with a negative angle. Summarizing, we assume we have the following information (all the widths are given in pixel units):

$\underline{\phi}_b$	$\overline{\phi}_b$	—	the angle in the vertical plane at which the runway begins
$\underline{w}_b$	$\overline{w}_b$	—	the width of the runway where it begins
$\underline{w}_{b'}$	$\overline{w}_{b'}$	—	the width of the runway in the first row of pixels above ϕ_b
$\underline{\phi}_e$	$\overline{\phi}_e$	—	the angle in the vertical plane to an arbitrary point on the runway
$\underline{w}_e$	$\overline{w}_e$	—	the width of the runway at ϕ_e

Each quantity comes with a lower and upper bound, denoted by the underline and the overline respectively, which is the result of the pixelization.

We now discuss the physical quantities we can derive from these measured quantities. First, the angle at which the runway begins, ϕ_b, relates to our state variables as

$$\underline{\phi}_b \le \gamma_R + \theta_p - \theta \le \overline{\phi}_b. \tag{4.15}$$

Second, the width in pixels of the runway where it begins, w_b, relates to the distance to the runway as $w_b = \mu/d$, where

$$\mu \doteq \frac{\text{runway width (meters)}}{\tan\left(\frac{\text{horizontal field of view (degrees)}}{2}\right)} \begin{pmatrix}\text{number of pixels on}\\ \text{the horizontal axis}\end{pmatrix}.$$

Last, a pixel at a vertical angle ϕ corresponds to a point on the surface which is at an angle of $-\phi-\theta$ below the horizon from the aircraft point of view. The distance to that point, assuming a planar terrain, is $\frac{\sin(-\phi_0-\theta)d_0}{\sin(-\phi-\theta)}$ where d_0 is the distance to another point on the surface which appears at an angle ϕ_0. We just showed that the distance to any object on the surface is inversely proportional to the width in pixels of that object. Thus we have that

$$\frac{\underline{w}_{b'}}{\overline{w}_e} \leq \frac{\sin\left(-\overline{\phi}_b-\theta\right)}{\sin\left(-\underline{\phi}_e-\theta\right)} \leq \frac{\overline{w}_{b'}}{\underline{w}_e} \tag{4.16}$$

from which we can derive bounds for possible values of θ. There is not a closed form solution to derive these bounds, but they can be easily calculated using simple iterative methods.

Although w_b is not controllable, estimating it helps us to estimate the time we expect to reach the runway, since

$$w_b = \frac{\mu}{V_E\left(t_f-t\right)} \doteq \frac{C_w}{t_f-t}. \tag{4.17}$$

4.6 Discretization and Linearization

In (4.12), (4.13), (4.14), (4.17) we have established that (4.1) is applicable to our system, where $u\left(t\right)=\delta_e\left(t\right)$, $z\left(t\right)=w_b\left(t\right)$, $a \doteq \left[t_f, C_{v\to v}, C_{v\to\dot{\theta}}, C_{\theta\to v}, \theta_0, C_{\theta\to\dot{\theta}}, C_{\dot{\theta}\to\dot{\theta}}, C_{\delta_e\to\dot{\theta}}, \delta_0, C_w\right]$,

$$A\left(t,a\right) \doteq \begin{bmatrix} \frac{1}{t_f-t} & -\frac{1}{t_f-t} & 0 & 0 \\ 0 & -C_{v\to v} & C_{\theta\to v} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & C_{v\to\dot{\theta}} & C_{\theta\to\dot{\theta}} & C_{\dot{\theta}\to\dot{\theta}} \end{bmatrix} \qquad B\left(a\right) \doteq \begin{bmatrix} 0 \\ 0 \\ 0 \\ C_{\delta_e\to\dot{\theta}} \end{bmatrix} \qquad D\left(a\right) \doteq \begin{bmatrix} 0 \\ -C_{\theta\to v}\theta_0 \\ 0 \\ -C_{\delta_e\to\dot{\theta}}\delta_0 \end{bmatrix}$$

$$C \doteq \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad E\left(t,a\right) \doteq \frac{C_w}{t_f-t}.$$

Defining $x'\left(k\right)=x\left(k\tau\right)$ we approximate the continuous dynamics with the following discrete version: $x'\left(k+1\right)=x'\left(k\right)+\tau\left(A\left(k\tau\right)x'\left(k\right)+Bu\left(k\right)+D\right)$. We now desire to derive a linear model for x_1 which depends only the observable states, x_1 and x_3. We see that

$$\left(t_f-\left(k-1\right)\tau\right)x_1'\left(k\right)=\left(t_f-\left(k-1\right)\tau+\tau\right)x_1'\left(k-1\right)-\tau x_2'\left(k-1\right),$$

so we can write

$$x_2'(k-1) = \left(\frac{t_f}{\tau} - k + 2\right) x_1'(k-1) - \left(\frac{t_f}{\tau} - k + 1\right) x_1'(k)$$

from which

$$\begin{aligned}(t_f - k\tau)\, x_1'(k+1) = &\left(t_f - k\tau + \tau + \tau\left(1 - \tau C_{v\to v}\right)\left(\frac{t_f}{\tau} - k + 1\right)\right) x_1'(k) \\ &- \tau\left(1 - \tau C_{v\to v}\right)\left(\frac{t_f}{\tau} - k + 2\right) x_1'(k-1) - \tau^2 C_{\theta\to v} x_3'(k-1) + \tau^2 C_{\theta\to v}\theta_0. \quad (4.18)\end{aligned}$$

Rearranging (4.18) we can also get

$$\begin{aligned}&k\tau x_1'(k+1) + (k-2)\,\tau x_1'(k-1) - 2(k-1)\,\tau x_1'(k) = \\ &\quad t_f\left(x_1'(k-1) - 2x_1'(k) + x_1'(k+1)\right) + \\ &\quad C_{v\to v}\left((k-2)\,\tau^2 x_1'(k-1) - (k-1)\,\tau^2 x_1'(k)\right) + \\ &\quad t_f C_{v\to v}\left(\tau x_1'(k) - \tau x_1'(k-1)\right) + \\ &\quad C_{\theta\to v}\tau^2 x_3'(k-1) - C_{\theta\to v} p_0 \tau^2. \quad (4.19)\end{aligned}$$

We now derive a linear model for the second observable state, x_3, which depends only on the observable states. As before

$$x_4'(k-1) = \frac{x_3'(k) - x_3'(k-1)}{\tau},$$

so that

$$\begin{aligned}x_3'(k+1) = &\tau^2 C_{v\to\dot\theta}\left(\frac{t_f}{\tau} - k + 2\right) x_1'(k-1) - \tau^2 C_{v\to\dot\theta}\left(\frac{t_f}{\tau} - k + 1\right) x_1'(k) + \left(2 + \tau C_{\dot\theta\to\dot\theta}\right) x_3'(k) + \\ &\left(\tau^2 C_{\theta\to\dot\theta} - 1 - \tau C_{\dot\theta\to\dot\theta}\right) x_3'(k-1) + \tau^2 C_{\delta_e\to\dot\theta} u(k-1) - \tau^2 C_{\delta_e\to\dot\theta} e_0. \quad (4.20)\end{aligned}$$

Rearranging (4.20) we can also get

$$\begin{aligned} & x_3'(k+1) - 2x_3'(k) + x_3'(k-1) = \\ & \quad C_{v\to\dot{\theta}}\left((2-k)\tau^2 x_1'(k-1) - (k-1)\tau^2 x_1'(k)\right) + \\ & \quad t_f C_{v\to\dot{\theta}}\left(\tau x_1'(k-1) - \tau x_1'(k)\right) + \\ & \quad C_{\theta\to\dot{\theta}}\tau^2 x_3'(k-1) + C_{\dot{\theta}\to\dot{\theta}}\left(\tau x_3'(k) - \tau x_3'(k-1)\right) \\ & \quad C_{\delta_e\to\dot{\theta}}\tau^2 u(k-1) - C_{\delta_e\to\dot{\theta}}\tau^2 e_0. \end{aligned} \tag{4.21}$$

Finally, we derive a linear model for the fifth state from (4.17):

$$(t_f - k\tau)\, z'(k) = C_w. \tag{4.22}$$

4.7 Implementation Details

Our model consists of 10 parameters. Looking at (4.19), (4.21) and (4.22) we see that to find the model parameters, a, which minimize the cost function from (4.3) given the output values at N samples, we need to solve a problem of the form

$$\min_a \|y - [a_1a_2, a_1a_3, a_1, \ldots, a_{10}]\, X\|_2^2$$

where $y \in \mathbb{R}^{1\times 3(N-2)}$, $X \in \mathbb{R}^{10\times 3(N-2)}$. We refer to §4.9 where we detail how we solve this nonlinear minimization problem efficiently. To find the output values which minimize (4.3) given the model parameters, we use (4.18), (4.20) and (4.22), and the constraints on $x_1 - x_3$, x_3 and z from §4.5 to generate a constrained quadratic programming formulation. We iterate these two steps, as described in §4.2, while adding more and more measurements as the simulation time advances. To initialize the process we set the output values to the center of each quantization range.

Since we estimate a discrete linear model, and we have constraints on the control input, we chose to use model predictive control (MPC) [40]. In order to use the MPC in the standard settings, we transform our LTV system to an LTI (linear time invariant) system by using the following state variables:

$$\tilde{x}_k(k) = \begin{bmatrix} (t_f - (k-1)\tau)\, x_1(k) \\ x_3(k) - p_0 \\ (t_f - (k-2)\tau)\, x_1(k-1) \\ x_3(k-1) - p_0 \end{bmatrix}.$$

The transformed state variable follows $\tilde{x}_{k+1} = \tilde{A}\tilde{x}_k + \tilde{B}\left(u - \tilde{e}_0\right)\ \forall k$ for some constant matrices $\tilde{A}$ and $\tilde{B}$, where $\tilde{e}_0 = e_0 - C_{\theta\to\dot{\theta}} p_0 / C_{\delta_e\to\dot{\theta}}$. We use H to denote the control horizon we use in the MPC. In order to satisfy conditions A1–A4 in [40, §3.3], which ensure closed-loop asymptotic stability in the non-adaptive case, we use the terminal constraint $\tilde{x}_{k+H} = 0$ and $u_{k+H} = \tilde{e}_0$. We use a quadratic cost where we only penalize the change in the control action: $\sum_{i=k+1}^{k+H} \left(u\left(i\right) - u\left(i-1\right)\right)^2$.

4.8 Simulation

We used a Cessna 172 model for our simulation. The numerical values we used are listed in §4.10. We positioned the airplane 250 m from the runway on the desired 1:9 ratio glide slope. The initial velocity was set to 36 m/s (∼70 knots) true air speed, moving parallel to the ground, no flaps configuration. The pitch and pitch rate were initialized to zero. Wind was set to 5 m/s headwind, ISA atmospheric conditions (ρ =1.2250 kg/m^3). The airplane weight was set to 2405 lb. The camera's field of view was 32° (horizontal) by 24° (vertical), its resolution was 640 by 480, and it took 50 frames per seconds (fps). The width of the runway was 20 m and ϕ_e was associated with a point on the runway that was distanced 200 m from the beginning of the runway. The simulation time constant (the time interval between each update of the dynamics) was 0.01 s. We started the simulation using an open-loop control consisting of several step functions. After 2 seconds we started running the estimator, and for every new measurement received we ran two iterations of the estimator. After 3 seconds we closed the loop by engaging the model predictive controller. The control input was limited to ±12°. The control horizon was set to $H = 150$ (or 3 s). Once the field of view was too small to cover the whole width of the runway where it begins, the controller was disengaged and the elevator deflection remained constant. We did not simulate ground effect.

We show here two runs of our simulation. In the first run we used the approximated dynamics which we derived in §4.4.2. The reason is that this way we know exactly to which model parameters the estimator should converge. In the second run we used the true airplane dynamics from §4.4.1. In the first run, in addition to the quantized estimator from §4.2, we also used a basic estimator for comparison. The basic estimator does not take into account the special characteristics of quantized measurements, and attempts to fit a model to measurements which are the center of each quantization range. Essentially the basic estimator minimizes the same cost function from (4.3) but only over the model parameters.

The simulation output is shown in Figures 4.1, 4.2, and 4.3. As predicted by Theorem 4.3.1, the estimated parameters for the quantized estimator do converge. Note that while the theorem only predicts convergence to a set, in the simulation the estimated parameters actually converge to a point corresponding to the true

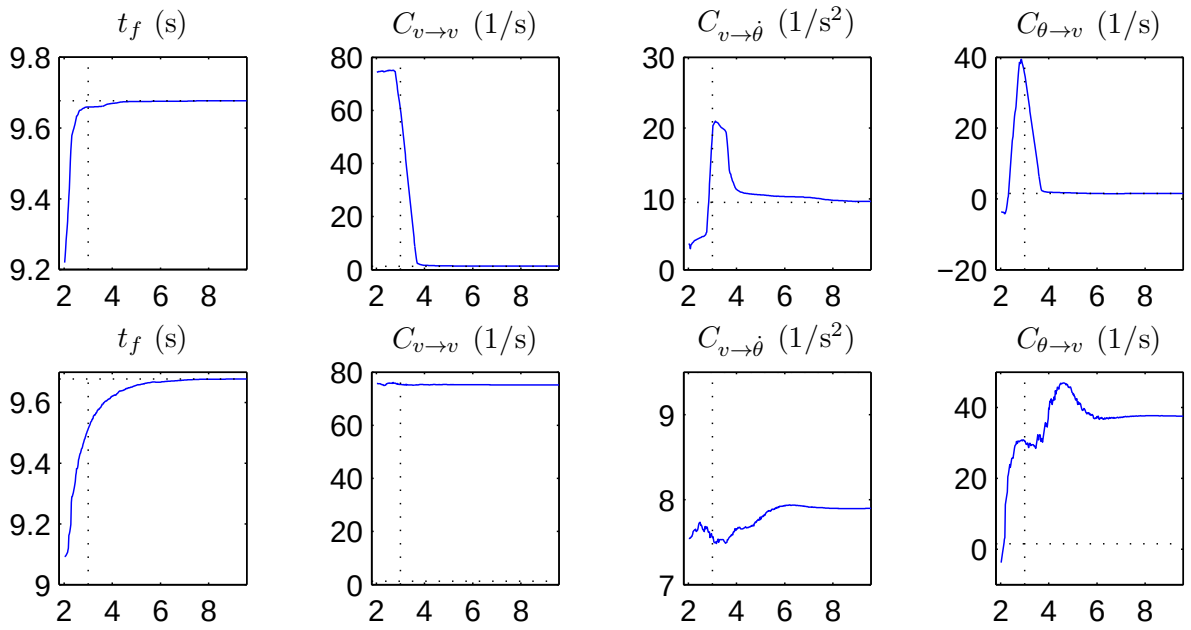

Figure 4.1: Comparison between the quantized estimator and the basic estimator using simulated linearized airplane dynamics. The first row of figures shows some of the estimates of the quantized estimator. The second row of figures shows the same estimates of the basic estimator. The x-axis in all the figures represents time (seconds). The horizontal dotted black lines in both sets of figures represent the true model parameters. Note the estimation error of the basic estimator compared to the quantized estimator especially in the estimation of $C_{v\to v}$ and of $C_{\theta\to v}$. The vertical dotted black lines in all the figures represent the time when the model predictive controller was engaged.

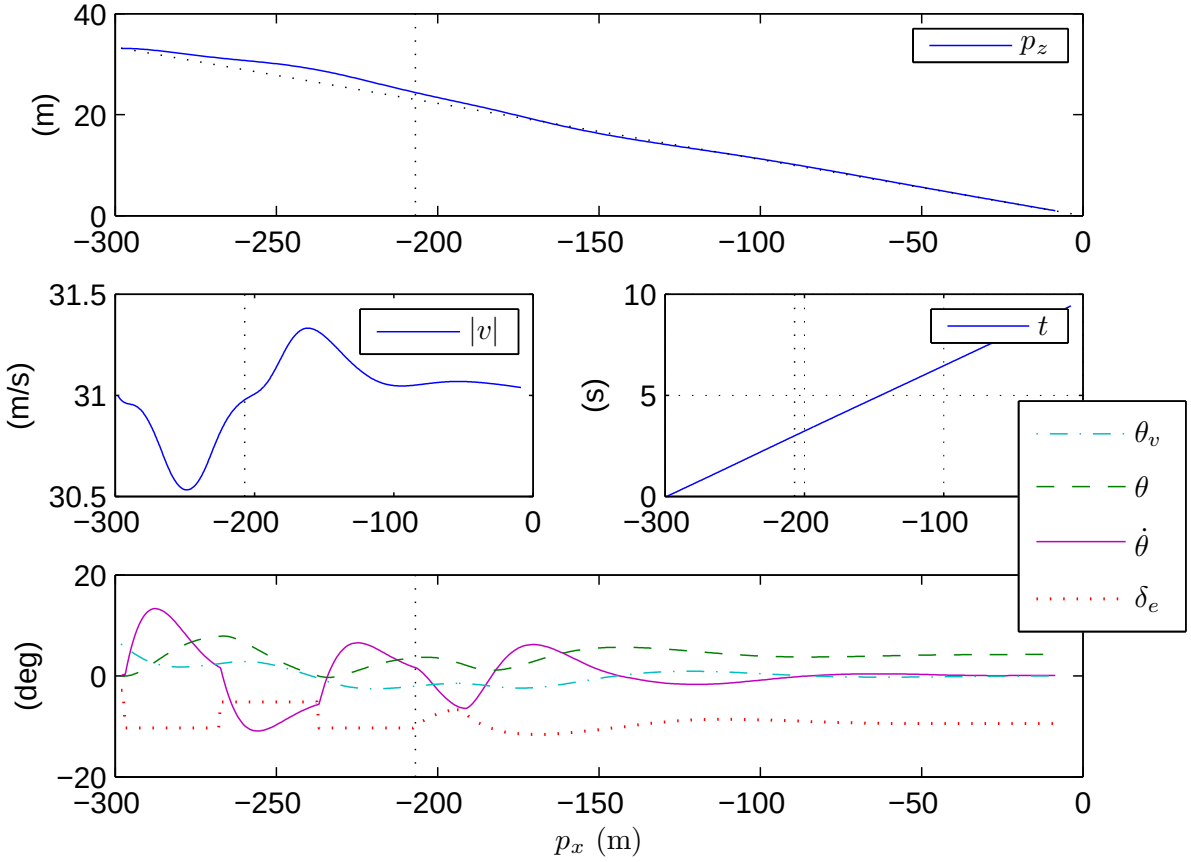

Figure 4.2: Simulation of true airplane dynamics using the quantized estimator. The four figures show the six degrees of freedom state of the airplane and the control input. The slanted dotted black line in the top figure indicates the desired glide slope. The vertical dotted black lines in all the figures represent the time when the model predictive controller was engaged.

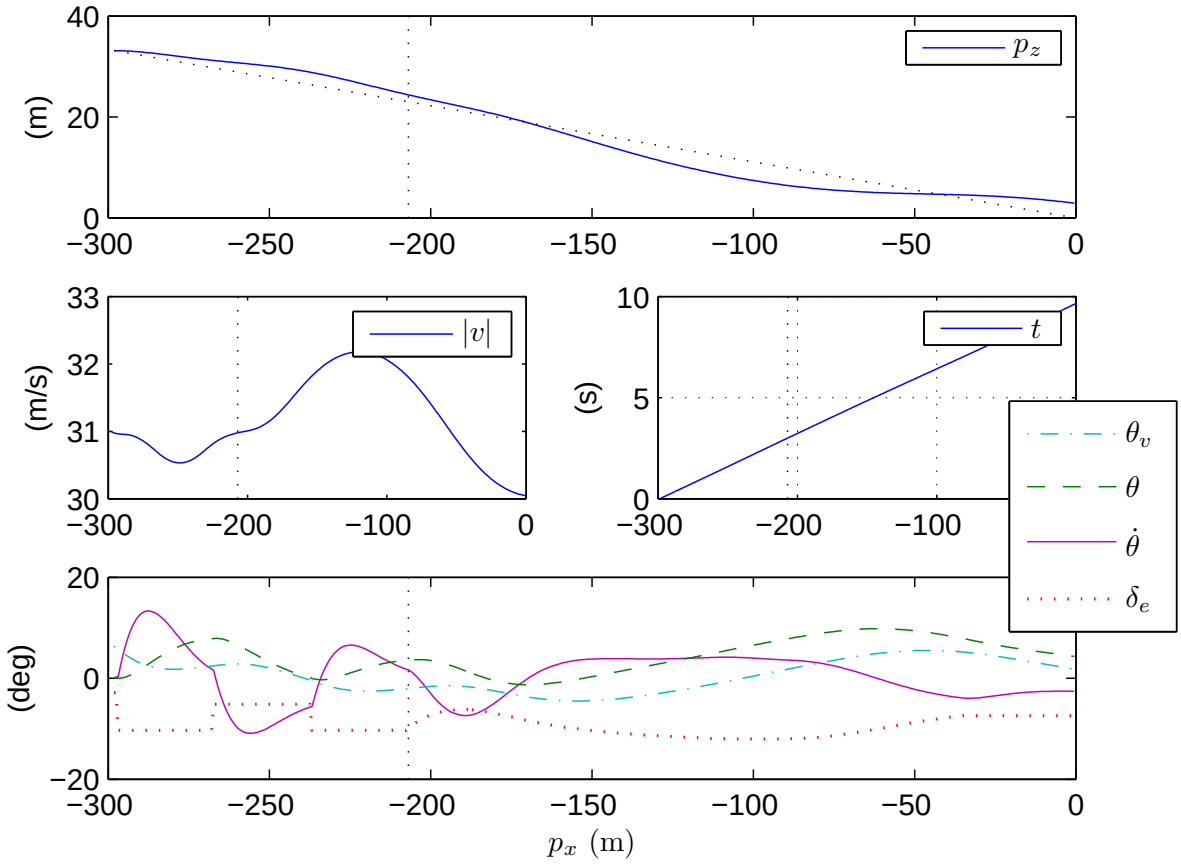

Figure 4.3: Simulation of true airplane dynamics using the basic estimator. The four figures show the six degrees of freedom state of the airplane and the control input. The slanted dotted black line in the top figure indicates the desired glide slope. The vertical dotted black lines in all the figures represent the time when the model predictive controller was engaged. Note the inability of the controller to converge to the gliding slope (compare to Figure 4.2).

value of the parameters. This gives hope that the results of Theorem 4.3.1 can be strengthened. The simulation also shows significant improvement in the estimation error between the basic estimator and the quantized estimator. Finally, the simulation shows that with the quantized estimator we were able to stabilize the system, whereas with the basic estimator the system failed to stabilize.

4.9 Nonlinear Minimization to Find the Model Parameters

Consider the following minimization problem:

$$\min_a \|y - [a_1 \cdot a_2, a_1 \cdot a_3, a_1, \ldots, a_k]\, X\|_2^2$$

where $y \in \mathbb{R}^{1\times N}$ and $X \in \mathbb{R}^{(k+2)\times N}$. Taking the derivative with respect to each a_i and equating to zero,

$$\begin{aligned}
a_2 v^T Q_1 + a_3 v^T Q_2 + v^T Q_3 - a_2 r_1 - a_3 r_2 - r_3 &= 0 \\
a_1 v^T Q_1 + v^T Q_4 - a_1 r_1 - r_4 &= 0 \\
a_1 v^T Q_2 + v^T Q_5 - a_1 r_2 - r_5 &= 0 \\
v^T Q_6 - r_6 &= 0 \\
&\vdots \\
v^T Q_{k+2} - r_{k+2} &= 0
\end{aligned}$$

where $v^T \doteq [a_1 a_2, a_1 a_3, a_1, \ldots, a_k]$, $Q = XX^T$ and $r = Xy^T$. The important thing to note is that whereas the original minimization problem involves $(k+3)\,N$ constants, the root finding problem above involves only $(k+3)\,(k+2)$ constants. To solve this root finding problem, which is still nonlinear, we derived its Jacobian and then used MATLAB's general purpose nonlinear solver. The Jacobian of the LHS of the above set of equations is

$$\begin{array}{cccc}
J_{11} & J_{12} & J_{13} & [a_2, a_3, 1]\, Q_{1:3,6:k+2} \\
J_{21} & J_{22} & J_{23} & [a_1, 1]\, Q_{[1,4],6:k+2} \\
J_{31} & J_{32} & J_{33} & [a_1, 1]\, Q_{[1,5],6:k+2} \\
a_2 Q_{61} + a_3 Q_{62} + Q_{63} & a_1 Q_{61} + Q_{64} & a_1 Q_{62} + Q_{65} & Q_{6,6:k-2} \\
\vdots & \vdots & \vdots & \vdots \\
a_2 Q_{(k+2)1} + a_3 Q_{(k+2)2} + Q_{(k+2)3} & a_1 Q_{(k+2)1} + Q_{(k+2)4} & a_1 Q_{(k+2)2} + Q_{(k+2)5} & Q_{(k+2),6:k-2}
\end{array}$$

where

$$
\begin{aligned}
J_{11} =& a_2\left(a_2Q_{11}+a_3Q_{12}+Q_{13}\right)+a_3\left(a_2Q_{21}+a_3Q_{22}+Q_{23}\right)+a_2Q_{31}+a_3Q_{32}+Q_{33}\\
J_{12} =& v^TQ_1+a_2\left(a_1Q_{11}+Q_{14}\right)+a_3\left(a_1Q_{21}+Q_{24}\right)+a_1Q_{31}+Q_{34}-r_1\\
J_{13} =& a_2\left(a_1Q_{12}+Q_{15}\right)+v^TQ_2+a_3\left(a_1Q_{22}+Q_{25}\right)+a_1Q_{32}+Q_{35}-r_2\\
J_{21} =& v^TQ_1+a_1\left(a_2Q_{11}+a_3Q_{12}+Q_{13}\right)+\left(a_2Q_{41}+a_3Q_{42}+Q_{43}\right)-r_1\\
J_{22} =& a_1\left(a_1Q_{11}+Q_{14}\right)+a_1Q_{41}+Q_{44} \qquad J_{23}=a_1\left(a_1Q_{12}+Q_{15}\right)+a_1Q_{42}+Q_{45}\\
J_{31} =& v^TQ_2+a_1\left(a_2Q_{21}+a_3Q_{22}+Q_{23}\right)+\left(a_2Q_{51}+a_3Q_{52}+Q_{53}\right)-r_2\\
J_{32} =& a_1\left(a_1Q_{21}+Q_{24}\right)+a_1Q_{51}+Q_{54} \qquad J_{33}=a_1\left(a_1Q_{22}+Q_{25}\right)+a_1Q_{52}+Q_{55}
\end{aligned}
$$

4.10 Aerodynamic Constants of Cessna 172

All the numerical values listed in Tables 4.1–4.4 were taken from [56], which lists several dynamical models to be used with the FlightGear Flight Simulator [16].

Table 4.1: Aerodynamic constants of Cessna 172

Constant Symbol	Value	Unit
C_{L_q}	3.9	s[/rad]
$C_{L_{\delta_e}}$	0.43	[/rad]
$C_{D_{\delta_e}}$	0	[/rad]
$C_{m_{\dot{\alpha}}}$	-5.2	s[/rad]
C_{m_q}	-12.4	s[/rad]
$C_{m_{\delta_e}}$	-1.28	[/rad]
$\bar{c}$	1.4935	m
S	16.1651	m^2
D_{cg}	$[0, -0.4572]^T$	m
I_{yy}	1825	kg-m^2

Table 4.2: Lift coefficient, C_{L_α}, as a function of angle of attack, α, obtained from [56]

Angle of attack (degrees)	Lift coefficient
-167.9998199	0.3
-149.9997778	0.15
-119.9997077	0.1
-89.9996375	0.
-3.999818368	-0.153
-2.999434058	-0.076
-1.999622705	0.001
-0.999811353	0.078
0.	0.155
0.999811353	0.232
2.000195663	0.309
3.000007015	0.386
3.999818368	0.463
5.000202678	0.54
6.000014031	0.617
6.999825383	0.694
8.000209693	0.771
9.000021046	0.848
9.999832398	0.925
11.00021671	1.002
12.00002806	1.079
12.99983941	1.156
14.00022372	1.233
15.00003508	1.28
15.99984643	1.3
17.00023074	1.3
18.00004209	1.23
18.99985344	0.9
20.00023775	0.6
21.00004911	0.575
21.99986046	0.55
30.00007015	0.3
60.00014031	0.2
90.00021046	0.
119.9997077	-0.1
149.9997778	-0.15
175.0002182	-0.3
179.999848	-0.1

Table 4.3: Drag coefficient, C_{D_α}, as a function of angle of attack, α, obtained from [56]

Angle of attack (degrees)	Drag coefficient
-179.999848	0.15
-167.9998199	0.2
-149.9997778	0.4
-119.9997077	0.65
-89.9996375	1.
-59.99956735	0.65
-29.9994972	0.4
-14.99946212	0.08
-5.999441073	0.035
-4.99962972	0.033
-3.999818368	0.031
-2.999434058	0.03
-1.999622705	0.03
-0.999811353	0.031
0.	0.033
0.999811353	0.035
2.000195663	0.038
3.000007015	0.042
3.999818368	0.046
5.000202678	0.051
6.000014031	0.056
8.000209693	0.069
9.000021046	0.076
9.999832398	0.084
11.00021671	0.093
12.00002806	0.102
12.99983941	0.112
14.00022372	0.115
15.00003508	0.12
15.99984643	0.13
17.00023074	0.138
18.00004209	0.145
18.99985344	0.15
20.00023775	0.165
21.00004911	0.175
21.99986046	0.35
30.00007015	0.65
60.00014031	1.
90.00021046	0.65
119.9997077	0.4
149.9997778	0.2
175.0002182	0.15
179.999848	0.15

Table 4.4: Pitch moment coefficient, C_{m_α}, as a function of angle of attack, α, obtained from [56]

Angle of attack (degrees)	Pitch moment coefficient
-179.999848	0.
-160.0001832	0.37
-139.9999454	0.59
-119.9997077	0.73
-100.0000429	0.79
-90.00021046	0.8
-79.9998051	0.7863
-60.00014031	0.72469
-40.00001714	0.60145
-20.00000857	0.3892
-10.00000429	0.115334
-4.999973495	0.037667
0.	-0.04
4.999973495	-0.117667
10.00000429	-0.195334
14.9816947	-0.273001
16.00001832	-0.29
17.00000156	-0.315
19.00002533	-0.375
20.00000857	-0.405
20.99999181	-0.42
21.99997505	-0.42
23.99999883	-0.39
27.00000584	-0.315
30.00001286	-0.215
34.99998635	-0.01
36.00002689	0.015
38.99997661	0.065
41.00000038	0.085
44.99999064	0.06
48.99998089	0.005
54.99999492	-0.15
69.9999727	-0.595
75.00017538	-0.72
79.9998051	-0.815
85.00000778	-0.875
90.00021046	-0.9
100.0000429	-0.89
119.9997077	-0.8
139.9999454	-0.64
160.0001832	-0.36
179.999848	0.

Chapter 5

Additional Results

This chapter consists of miscellaneous results which were obtained as part of the investigations reported in the previous chapters, but which have not reached sufficient maturity to be included in those chapters or as separate chapters.

5.1 Control Input Generation for Quantized Measurements

Most works on quantization make a separation between estimating the state and setting the control input. The prevailing approach is to select a control law as a function of the state estimate that makes the closed-loop system stable to estimation error, and then use the quantized measurements to minimize the estimation error. With quantized measurements, our true estimate of the state is a region, not a point, in the state space. Choosing the center point of that region and applying the control law using that point as a state estimate may not be optimal in terms of the control objective we seek to achieve. The question we try to answer here is as follows. Given a discrete linear system with an associated quadratic infinite horizon cost function, and given a region in the state space known to contain the state of the system, what is the control input that is guaranteed to decrease the cost function the most in the worst case?

More precisely, we formulate the following problem: Consider the discrete control system,

$$x\left(k+1\right) = Ax\left(k\right) + Bu\left(k\right) \tag{5.1}$$

with state $x(k) \in \mathbb{R}^n$ and control input $u(k) \in \mathbb{R}^m$. Assume we want to minimize the quadratic infinite horizon cost,

$$\sum_{k=0}^{\infty} x\left(k\right)^T Qx\left(k\right) + u\left(k\right)^T Ru\left(k\right) \tag{5.2}$$

where Q and R are positive definite matrices. Solving the discrete-time algebraic Riccati equation associated

with this optimal control problem, we can find a positive definite matrix P such that

$$\min_{\substack{u(k)\in\mathbb{R}^m \\ k=0,\ldots,\infty}} \sum_{k=0}^{\infty} x\left(k\right)^T Qx\left(k\right) + u\left(k\right)^T Ru\left(k\right) = x\left(0\right)^T Px\left(0\right). \tag{5.3}$$

We define $V\left(x\right) = x^T Px$, which is a Lyapunov function for this system. Define $[n] \doteq \{1, \ldots, n\}$. Assume at every time step we are given $\underline{x}(k)$, $\overline{x}(k)$ such that

$$\underline{x}_i\left(k\right) \leq x_i\left(k\right) \leq \overline{x}_i\left(k\right), \qquad \forall i \in [n]. \tag{5.4}$$

We will use the notation X_k to denote the set of all x_k that satisfy (5.4). From (5.3), the optimal u at time step k is $u\left(k\right) = \arg\min_{u\in\mathbb{R}^m} V\left(Ax\left(k\right) + Bu\right) - V\left(x\left(k\right)\right) + u^T Ru$. However, since we do not know what $x\left(k\right)$ is exactly, we want to solve for

$$u_k = \arg\min_u \max_{x\in X_k} V\left(Ax + Bu\right) - V\left(x\right) + u^T Ru. \tag{5.5}$$

Note that

$$\max_{x\in X_k} V\left(Ax + Bu\right) - V\left(x\right) \tag{5.6}$$

is not a convex optimization problem. However, we can use the algorithm described in §5.1.1 to solve it efficiently. Once we can solve (5.6), we can use nonlinear solvers to solve the unconstrained minimization problem (5.5) that will provide us with a local minimum. In §5.1.2 we will demonstrate the benefits of this approach through a simulation.

5.1.1 Solving for the Worst State

For fixed u we can write $V\left(Ax + Bu\right) - V\left(x\right)$ as

$$Q\left(x; M, v\right) = x^T Mx + c^T x \tag{5.7}$$

where M is symmetric but not necessarily definite.

Lemma 5.1.1. $\sup_{x\in\mathrm{int}(X)} Q\left(x; M, c\right) > \max_{X\setminus\mathrm{int}(X)} Q\left(x; M, c\right)$ *if and only if* M *is negative definite and* $-M^{-1}c \in \mathrm{int}\left(X\right)$.

Proof: If M is negative definite, then $Q\left(x; M, c\right)$ is strictly concave and thus it has a unique maximizer in $\mathbb{R}^n$. Differentiating $Q\left(x; M, c\right)$ reveals that this maximizer is $-M^{-1}v$. If indeed $-M^{-1}v \in \mathrm{int}\left(X\right)$,

then $\sup_{x\in\text{int}(X)} Q(x;M,c) = Q\left(-M^{-1}v;M,c\right) > \max_{X\setminus\text{int}(X)} Q(x;M,c)$. If $-M^{-1}v \notin \text{int}(X)$, choose an arbitrary point $x \in \text{int}(X)$. Because of concavity, $\forall\varepsilon \in (0,1]$ the point $y(x,\varepsilon) \doteq (1-\varepsilon)x + \varepsilon\left(-M^{-1}v\right)$ satisfies $Q(y;M,c) < Q(x;M,c)$. And for every $x \in \text{int}(X)$ we can find $\varepsilon_x \in (0,1]$ such that $y(x,\varepsilon_x) \in X \setminus \text{int}(X)$. This implies that $\sup_{x\in\text{int}(X)} Q(x;M,c) \leq \max_{X\setminus\text{int}(X)} Q(x;M,c)$.

Now assume M is not negative definite. Then there exists $v \in \mathbb{R}^n \setminus 0$ such that $v^T Mv \geq 0$. Writing for some $x \in \mathbb{R}^n$

$$(x+\varepsilon v)^T M(x+\varepsilon v) + c^T(x+\varepsilon v) = x^T Mx + c^T x + \varepsilon^2 v^T Mv + \varepsilon\left(2xM + c^T\right)v, \tag{5.8}$$

we can see that by changing the sign of v if necessary to make $\left(2xM + c^T\right)v \geq 0$, we can have $Q(x;M,c) \leq Q(x+\varepsilon v;M,c), \forall\varepsilon \geq 0$. This again means that for every point $x \in \text{int}(X)$ we can find a point $y \in X\setminus\text{int}(X)$ such that $Q(x;M,c) \leq Q(y;M,c)$, and conclude that $\sup_{x\in\text{int}(X)} Q(x;M,c) \leq \max_{X\setminus\text{int}(X)} Q(x;M,c)$. ■

By Lemma 5.1.1, if M is not negative definite, or $-M^{-1}c \notin \text{int}(X)$, then we only need to look at the boundaries of X for the point that maximizes $Q(x;M,c)$ over X. We can then use the recursive algorithm below to find this point.

We use the notation $In_{*,[n]\setminus i}$, $i \in [n]$, to denote an n by $n-1$ matrix which contains all the columns of the n by n identity matrix except for the i'th column.

Algorithm 6

Require: n, $M \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}$, $\underline{x} \in \mathbb{R}^n$, $\overline{x} \in \mathbb{R}^n$
  if M is negative definite and $\underline{x} \leq -M^{-1}c \leq \overline{x}$ **then**
    set $x = -M^{-1}c$
  else
    set $x = (\underline{x} + \overline{x})/2$
    for $i = 1,\ldots,n$ **do**
      set $H = In_{*,[n]\setminus i}$
      set $x' = 0 \in \mathbb{R}^n$
      set $x'_i = \underline{x}_i$
      for $j = 1,2$ **do**
        if $n \geq 1$ **then**
          use Algorithm 6 to find $x'' = \arg\max_{\underline{x}_{[n]\setminus i} \leq x \leq \overline{x}_{[n]\setminus i}} Q\left(x; H^T MH, H^T Mx' + H^T c\right) \in \mathbb{R}^{n-1}$
        end if
        if $Q(x' + Hx''; M, c) > Q(x; M, c)$ **then**
          set $x = x' + Hx''$
        end if
        set $x'_i = \overline{x}_i$
      end for
    end for
  end if
Ensure: $x = \arg\max_{\underline{x}\leq x\leq\overline{x}} x^T Mx + c^T x$

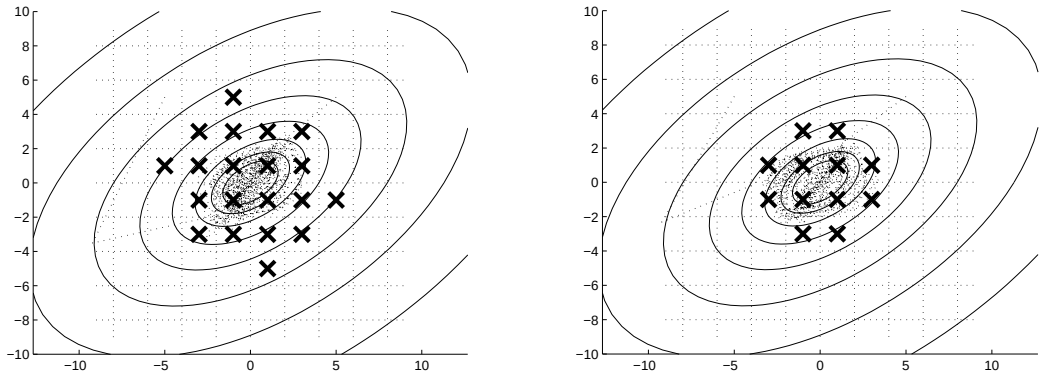

Figure 5.1: Results of the simulations described in §5.1.2. The horizontal and vertical dotted lines depict the boundaries of the quantization regions. The elliptical lines correspond to level sets of the Lyapunov function. All the quantization regions where the Lyapunov function is not guaranteed to decrease are marked with $\times$. Finally, two sample trajectories, starting from $[-5, 5]$ and $[5, 5]$, are plotted by dotted lines.

5.1.2 Simulation

We simulated a control system where the state evolves according to (5.1) and $A = \begin{bmatrix} 0.71 & -1.14 \\ -0.05 & 1.75 \end{bmatrix}$, $B = [0, 1]^T$. We set the quantization such that the lower bound is $\underline{x}_i(k) = 2\max\{z \in \mathbb{Z} \,|2z \leq x_i(k)\}$ and the upper bound is $\bar{x}_i(k) = 2\min\{z \in \mathbb{Z} \,|2z > x_i(k)\}$. We considered the cost function (5.2) where $Q = I_2$, $R = 0.001$. We ran two simulations. In the first simulation we set the control input to $u(k) = -\left(R + B^T P B\right)^{-1} B^T P A \hat{x}(k)$ where $\hat{x}(k) = (\underline{x}(k) + \bar{x}(k))/2$. This is the solution to (5.5) if $X(k) = \{\hat{x}(k)\}$. In the second simulation we set the control input to the solution of (5.5) when X_k is the set of all x_k that satisfy (5.4). The results are displayed in Figure 5.1.

In both simulations each quantization region is associated with a fixed control input. We marked every quantization region where the Lyapunov function is not strictly decreasing for every state in that quantization region using the corresponding control input. Using standard Lyapunov analysis, it is easy to prove that the system converges to within the smallest level set of the Lyapunov function encompassing all the marked quantization regions. Since there are fewer marked quantization regions in the second simulation, with our approach of setting the control input we can prove convergence to a smaller set.

5.1.3 Discussion

In this work we assume to have a Lyapunov function represented by the matrix P, and we look for an appropriate control input for every quantization region. In some sense this is complimentary to [24], where one assumes the control inputs are given and then looks for an appropriate P matrix using a randomized

algorithm. It would be interesting to explore whether the results here can be used to derive a deterministic algorithm for the problem that was posed in [24]. The Lyapunov function resulting from the algorithm in [24] is only guaranteed to decrease at the set of points that are randomly sampled by the algorithm. Extending the approach presented here may allow to guarantee the decrease of the Lyapunov function over the whole set from which quadratic stability can be established. A more long-term research project would be to simultaneously find a Lyapunov function and the control inputs that will minimize the region into which all trajectories of the closed-loop system converge.

5.2 Stability Analysis for Disturbed Systems with Deterministic and Stochastic Information

In this section we want to predict the behavior of a disturbed control system using either deterministic or probabilistic information on the disturbance. In particular we consider the discrete system

$$\boldsymbol{x}\left(k+1\right)=\bar{A}\boldsymbol{x}\left(k\right)+\bar{B}\boldsymbol{w}\left(k\right) \tag{5.9}$$

where $\boldsymbol{x}$ is the state, $\boldsymbol{w}$ is the disturbance, and $\bar{A}$ is Schur. The motivation for studying this system in the context of limited information feedback is as follows. Assume we have the following control system:

$$\boldsymbol{x}(k+1)=A_d\boldsymbol{x}(k)+B_d\boldsymbol{u}(k),$$

with $\boldsymbol{u}\left(k\right)$ chosen to be $K\hat{\boldsymbol{x}}(k)$ for some K such that A_d+B_dK is Schur. Then denoting the estimation error as $\tilde{\boldsymbol{x}}(k)\doteq\hat{\boldsymbol{x}}(k)-\boldsymbol{x}(k)$, the control system can be rewritten as

$$\boldsymbol{x}(k+1)=A_d\boldsymbol{x}(k)+B_dK\hat{\boldsymbol{x}}(k)=\left(A_d+B_dK\right)\boldsymbol{x}(k)+B_dK\tilde{\boldsymbol{x}}(k)\doteq\bar{A}\boldsymbol{x}(k)+\bar{B}\tilde{\boldsymbol{x}}(k),$$

which corresponds to (5.9). In the next subsection we provide a deterministic analysis for reference, but remark that the results in this subsection are not new. In the subsection after that, we provide a probabilistic analysis that we believe to be novel.

5.2.1 Deterministic Analysis

We assume here that a deterministic bound on the disturbance, $\boldsymbol{w}$, is given. Because we assume $\bar{A}$ is stable (Schur), for every positive definite Q there exists P such that $\bar{A}^TP\bar{A}-P=-Q$. Since $\boldsymbol{w}$ is bounded, we

can derive

$$\alpha = 1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}, \quad \beta = 2\frac{\max \left\| \bar{A}^T P \bar{B} \boldsymbol{w}(k) \right\|_2}{\sqrt{\lambda_{min}(P)}}, \quad \gamma = \max \left\| \boldsymbol{w}(k)^T \bar{B}^T P \bar{B} \boldsymbol{w}(k) \right\|_2.$$

Proposition 5.2.1. *With the parameters defined above, if* $\|\boldsymbol{x}(t_0)\|_2 \leq c$ *for some* $c \in \mathbb{R}^n_+$ *and* $t_0 \in \mathbb{R}$*, then* $\forall k \geq T$, $\|\boldsymbol{x}(k)\|_2 \leq d$ *where*

$$d = \frac{1}{\sqrt{\lambda_{min}(P)}} \left(\frac{\beta + \sqrt{\beta^2 + \gamma(\delta - \alpha)}}{(\delta - \alpha)} \right)$$

and

$$T = t_0 + \frac{1}{\log(\delta)} \log \left(\frac{1}{\lambda_{max}(P) c^2} \left(\frac{\beta + \sqrt{\beta^2 + \gamma(\delta - \alpha)}}{(\delta - \alpha)} \right)^2 \right).$$

Proof: The proof follows standard Lyapunov analysis. Define $V(k) = \boldsymbol{x}(k)^T P \boldsymbol{x}(k)$. Then,

$$V(k+1) = V(k) + \boldsymbol{x}(k)^T \left(\bar{A}^T P \bar{A} - P \right) \boldsymbol{x}(k) + 2\boldsymbol{x}(k)^T \bar{A}^T P \bar{B} \tilde{\boldsymbol{x}}(k) + \tilde{\boldsymbol{x}}(k)^T \bar{B}^T P \bar{B} \tilde{\boldsymbol{x}}(k) \tag{5.10}$$

from which we can write

$$V(k+1) \leq \alpha V(k) + \beta \sqrt{V(k)} + \gamma.$$

For all $\delta \in (\alpha, 1)$ we get that

$$V(k) \geq \left(\frac{\beta + \sqrt{\beta^2 + \gamma(\delta - \alpha)}}{(\delta - \alpha)} \right)^2 \Rightarrow V(k+1) \leq \delta V(k)$$

so we can write

$$V(k) \leq \max \left\{ \delta^k V(0), \left(\frac{\beta + \sqrt{\beta^2 + \gamma(\delta - \alpha)}}{(\delta - \alpha)} \right)^2 \right\}. \tag{5.11}$$

The rest of the proof follows easily. ■

Since (5.11) is true for all $\delta \in (\alpha, 1)$, in particular for $\delta \to 1$ we get

$$\lim_{k \to \infty} V(k) = \left(\frac{\beta + \sqrt{\beta^2 + \gamma(1 - \alpha)}}{(1 - \alpha)} \right)^2. \tag{5.12}$$

We note that the approach presented in the proof of Proposition 5.2.1 is not the only approach available in the literature for predicting the limit to which the norm of the state converges. Following [26, Example 2.4], we can also write for (5.9):

$$|\boldsymbol{x}(k)| \leq \beta(|\boldsymbol{x}(0)|, k) + \gamma(\|\boldsymbol{w}\|), \qquad \beta(r, k) \doteq c\sigma^k r, \quad \gamma(r) \doteq \frac{c \left\| \bar{B} \right\| r}{1 - \sigma} \tag{5.13}$$

where $c > 0$ and $0 \leq \sigma < 1$ are constants such that $\left\|A^k\right\| \leq c\sigma^k \ \forall k \in \mathbb{N}$.

5.2.2 Probabilistic Analysis

In the previous subsection we assumed a deterministic bound on the disturbance. In this subsection we assume the disturbance has zero mean, and we know its covariance, $\Lambda = \mathrm{E}\, \boldsymbol{w}\boldsymbol{w}^T$. Note that this does not imply that we assume that the disturbance follows a Gaussian distribution. We further assume that the disturbance at each time step is uncorrelated with the state at previous and current time steps, as well as with the disturbances at previous time steps.

Again, we define a Lyapunov function $V(k) = \boldsymbol{x}(k)^T P \boldsymbol{x}(k)$ such that $\bar{A}^T P \bar{A} - P = -Q$ where both P and Q are positive definite. We can now evaluate the expected change in the Lyapunov function:

$$\begin{aligned} \mathrm{E}\left(V(k+1) \,|V(k)\right) &\leq \sup_{\boldsymbol{x} | \boldsymbol{x}^T P \boldsymbol{x} \leq V(k)} \boldsymbol{x}(k)^T \left(\bar{A}^T P \bar{A} - P + P\right) \boldsymbol{x}(k) + \mathrm{E}\, \boldsymbol{w}(k)^T \bar{B}^T P \bar{B} \boldsymbol{w}(k) \\ &= - \sup_{\boldsymbol{x} | \boldsymbol{x}^T P \boldsymbol{x} \leq V(k)} \boldsymbol{x}(k)^T Q \boldsymbol{x}(k) + V(k) + \mathrm{trace}\left(\Lambda \bar{B}^T P \bar{B}\right). \end{aligned} \tag{5.14}$$

Note the disappearance of the cross terms involving $\boldsymbol{x}$ and $\boldsymbol{w}$, which did appear in (5.10), due to the assumption that these two random variables are uncorrelated with each other. From this we can write

$$\mathrm{E}\left(V(k+1) \,|V(k)\right) \leq \alpha V(k) + \beta \tag{5.15}$$

where

$$\alpha \doteq \left(1 - \frac{\lambda_{min}(Q)}{\lambda_{max}(P)}\right), \qquad \beta \doteq \mathrm{trace}\left(\Lambda \bar{B}^T P \bar{B}\right).$$

We can then get that

$$\lim_{k \to \infty} \mathrm{E}\, V(k) \leq \frac{\beta}{1 - \alpha}. \tag{5.16}$$

The preceding analysis follows similar lines as the analysis appearing in [20, 30], with necessary modifications for discrete systems. Because we are using a stochastic analysis, in addition to evaluating the expected value of the Lyapunov function, we also need to estimate or bound the probability of deviating substantially from the expected value. In the cited literature, this probability was bounded using Chebyshev's inequality, utilizing the fact that the Lyapunov function is nonnegative. The disadvantage of using Chebyshev's inequality is that it assumes a worst-case probability distribution given the expected value. In the following discussion, we will try to better bound the probability of deviating from the expected value using a more realistic probability distribution. In addition, we will also try to bound the probability that the supremum

of the Lyapunov function over several time steps deviates from its expected value. This is in contrast with existing literature which only considers one time step.

Theorem 5.2.1. *Given a Lyapunov function $V(k) = \boldsymbol{x}(k)^T P \boldsymbol{x}(k)$ such that $\bar{A}^T P \bar{A} - P = -Q$, its expected value evolves according to* (5.15)*. In addition, given the value of the Lyapunov function at k_0, and given $\gamma > 0$, $k_2 > k_1 > k_0$ and $\theta \geq 0$, the probability*

$$\operatorname{Prob}\{\exists k \in \{k_1, \ldots, k_2\} \,|V(k) > \theta\} \tag{5.17}$$

can be evaluated numerically using the algorithm below.

Currently we will proceed assuming the following conjecture is correct:

Conjecture 5.2.2. *Assume $V(k)$ evolves according to* (5.15)*, and define the random process $y(k)$, $k > k_0$, to evolve according to*

$$\begin{aligned} y(k) \,|y(k-1) &\sim \mathcal{N}\left(\sqrt{\alpha} y(k-1), \beta\right) \\ y(k_0) &\sim \mathcal{N}(0, V(k_0)). \end{aligned} \tag{5.18}$$

Then $\operatorname{Prob}\{\exists k \in \{k_1, \ldots, k_2\} \,|V(k) > \theta\} \leq \operatorname{Prob}\{\exists k \in \{k_1, \ldots, k_2\} \,|z(k) > \theta\}$ *where $z(k) \doteq y^2(k)$.*

Note that if the state $\boldsymbol{x}$ has single dimension ($n = 1$), the disturbance $\boldsymbol{w}$ is Gaussian, and $P = 1$, then V evolves exactly as z.

Proof of Theorem 5.2.1 (based on Conjecture 5.2.2): Calculate

$$\operatorname{E} z(k_1) = \alpha^{k_1 - k_0} V(k_0) + \left(\sum_{i=1}^{k_1 - k_0} \alpha^{i-1} \right) \beta.$$

We want to calculate the probability of $z(k) \geq \theta$ for at least one $k \in \{k_1 \ldots k_2\}$. We will use the notation $\|z\|_{\{k_1 \ldots k_1 + m\}} \doteq \max_{k \in \{k_1 \ldots k_1 + m\}} z(k)$. From (5.18) and using symmetry we can write

$$\begin{aligned} z(k) \,|z(k-1) &\sim \mathcal{N}^2\left(\sqrt{\alpha}\sqrt{z(k-1)}, \beta\right) \\ z(k_1) &\sim \mathcal{N}(0, \operatorname{E} z(k_1)) \end{aligned} \tag{5.19}$$

where $\mathcal{N}^2$ stands for the square of a Gaussian random variable with the given expectation and variance.

The distribution $\mathcal{N}^2$ has the following pdf:

$$f_{\mathcal{N}^2}\left(x;\mu,\sigma^2\right)=\frac{1}{2\sqrt{2\pi\sigma^2 x}}e^{-\frac{\left(\sqrt{x}-\mu\right)^2}{2\sigma^2}}+\frac{1}{2\sqrt{2\pi\sigma^2 x}}e^{-\frac{\left(-\sqrt{x}-\mu\right)^2}{2\sigma^2}}. \tag{5.20}$$

Obviously,

$$\operatorname{Prob}\left(z\left(k_1\right)\geq\theta\right)=\int_{\theta}^{\inf}f_{\mathcal{N}^2}\left(x;0,\operatorname{E}z\left(k_1\right)\right)\operatorname{d}x. \tag{5.21}$$

Define recursively the pdf of $z\left(k_1+m\right)$ given that $z\left(k\right)<\theta\ \forall k\in\left\{k_1\ldots k_1+m-1\right\}$:

$$\begin{aligned}g\left(x;\theta,m\right)&\doteq f_{z(k+m)\left|\|z\|_{\{k_1\ldots k_1+m-1\}}<\theta\right.}\\&=\int_0^{\theta}f_{\mathcal{N}^2}\left(x;\sqrt{\alpha w},\beta\right)\frac{g\left(w;\theta,m-1\right)}{\int_0^{\theta}g\left(w';\theta,m-1\right)\operatorname{d}w'}\operatorname{d}w\\g\left(x;\theta,0\right)&=f_{\mathcal{N}^2}\left(x;0,\operatorname{E}z\left(k_1\right)\right).\end{aligned} \tag{5.22}$$

We can now calculate iteratively:

$$\begin{aligned}\operatorname{Prob}\left(\|z\|_{\{k_1\ldots k_1+m\}}\geq\theta\right)=&\operatorname{Prob}\left(\|z\|_{\{k_1\ldots k_1+m-1\}}\geq\theta\right)+\\&\operatorname{Prob}\left(\|z\|_{\{k_1\ldots k_1+m-1\}}<\theta\right)\times\operatorname{Prob}\left(z\left(k_1+m\right)\geq\theta\left|\|z\|_{\{k_1\ldots k_1+m-1\}}<\theta\right.\right)\\=&\operatorname{Prob}\left(\|z\|_{\{k_1\ldots k_1+m-1\}}\geq\theta\right)+\\&\left(1-\operatorname{Prob}\left(\|z\|_{\{k_1\ldots k_1+m-1\}}\geq\theta\right)\right)\int_{\theta}^{\infty}g\left(x;\theta,m\right)\operatorname{d}x.\end{aligned} \tag{5.23}$$

Computing $\int_{\theta}^{\infty}g\left(x;\theta,m\right)\operatorname{d}x$ using (5.22) can be done numerically, which in the general case has a complexity linear with m. However, if we assume that $\operatorname{E}z\left(k_1\right)\approx\frac{\beta}{1-\alpha}$, the steady-state expectation of (5.19), and that θ is at least a few times larger than $\frac{\beta}{1-\alpha}$, then $\int_{\theta}^{\infty}g\left(x;\theta,m\right)\operatorname{d}x$ converges after only a few iterations. With these assumptions, (5.23) can be computed instantly for any m. ■

5.2.3 Discussion

The results we derived hold for scalar systems with independent Gaussian disturbance. They may also apply for higher dimensions if Conjecture 5.2.2 can be proved to hold in that setting. We note that this work originated from the desire to explain and predict the behavior of control systems with quantized state or output feedback. When we applied the deterministic analysis to simulations of such systems, the performance of the system was considerably better than the analysis predicted. The probabilistic analysis presented here, on the other hand, predicted a better performance than what was actually observed. We

believe that this discrepancy stems mainly from the non-correlation assumption between the disturbance and the state of the system, which does not hold when the disturbance is due to quantization errors.

5.3 Change in Entropy As a Condition for Convergence of State Estimate under Quantization

In Chapter 2 we used a geometric approach to address quantization. This implies that after some finite time we are able to construct a compact containment region known to contain the state. Consecutive measurements allow us to reduce the size of that containment region until, in the absence of external disturbances, its size converges to zero.

In this section we seek to address quantization using an information approach. Looking at the distribution function of the state, measuring how it changes due to the instability of the system and when new measurements arrive, we want to derive necessary and sufficient conditions for the convergence of the estimation error. This may allow us to derive convergence results even when the geometry approach is not applicable. We show for example that, for linear systems with equiprobable quantization, we can get convergence without switching between a zoom-in and a zoom-out mode. Potentially, it could lead to results applicable to more general quantizers where the quantization indices do not necessarily correspond to uniform distributions over mutually disjoint set.

Notable results following the information approach include [50, 22, 70, 69, 45, 39, 55, 46]. All these results, however, develop a specific quantization scheme with which they prove convergence of the estimation error. In contrast, here we look for general conditions, not related to a specific quantization scheme, that will guarantee convergence.

A related work, which also does not consider a specific quantization scheme, is [46]. There, using the notion of topological feedback entropy, a necessary and sufficient condition in terms of the average number of quantization regions was derived for general nonlinear systems. Yet given a quantization scheme, that work also does not show how to answer whether this quantization scheme produces a converging state estimate.

5.3.1 Definitions

Consider the following single dimension dynamical system with $x_k \in \mathcal{X} \subseteq \mathbb{R}$, $u_k \in \mathcal{U} \subseteq \mathbb{R}$:

$$x_{k+1} = F\left(x_k, u_k\right). \tag{5.24}$$

We assume F is an injective function. We consider x_k to be the realization of a random variable X_k. It is important to note that we do not assume that $\mathcal{X}$ is compact.

The only information that can be transmitted to the controller at every step k is an integer between 1 and N. For each k let $Q_1^k, \ldots, Q_N^k$ be a partition of $\mathcal{X}$,

$$Q_i^k \cap Q_j^k = \emptyset, \forall i \neq j, \quad \cup_{i=1}^N Q_i^k = \mathcal{X}. \tag{5.25}$$

Let q_k be the random variable such that $q_k = i$ if and only if $x_k \in Q_i^k$. We use $d_k \in \{1, \ldots, N\}$ to denote a possible realization of q_k. We also define $\vec{q}_k \doteq [q_0, \ldots, q_k]$ and use $\vec{d}_k \in \{1, \ldots, N\}^k$ to denote a possible realization of $\vec{q}_k$. Note that we consider the partitions of $\mathcal{X}$ at step k to be a function of $\vec{q}_{k-1}$ (we do not assume they are given a priori).

Let $f_{X_k} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be the probability density function (pdf) of the state location at time step k — $\forall t \in \mathbb{R}$: $\text{Prob}\left(X_k \leq t\right) = \int_{-\infty}^{t} f_{X_k}\left(x\right) \mathrm{d}\, x$. Let $f_{X_k|\vec{q}_{k'}=\vec{d}_{k'}}$ be the pdf of the state's location at time step k given that $x_i \in Q_{d_i}^i$ $\forall i \in \{0, \ldots, k'\}$. We assume the initial pdf f_{X_0} is given. Again we note that we do not assume that f_{X_0} is nonzero only inside a compact set. We also assume that there is at most a countable number of points where f_{X_0} is not differentiable. Between the time steps, the controller updates the pdf according to the Frobenius-Perron operator:

$$f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\left(x\right) = \frac{f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\left(y\right)}{\left|\partial F\left(z, u_{k-1}\right)/\partial z_{|z=y}\right|} \tag{5.26}$$

where y is such that $F\left(y, u_{k-1}\right) = x$. When new information arrives to the controller, it updates the pdf according to

$$f_{X_k|\vec{q}_k=\vec{d}_k}\left(x\right) = \begin{cases} \frac{f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}(x)}{p_i^k} & x \in Q_{d_k}^k \\ 0 & x \notin Q_{d_k}^k \end{cases} \tag{5.27}$$

where $p_i^k \doteq \int_{Q_i^k} f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\left(x\right) \mathrm{d}\, x$, the probability that Q_i^k will be active given that $\vec{q}_{k-1} = \vec{d}_{k-1}$.

The estimated state location will be set to equal the expected value of X_k given the measurements:

$$\hat{x}_k\left(\vec{d}_k\right) = \mathrm{E}\left(X_k \middle| \vec{q}_k = \vec{d}_k\right) \doteq \int_{\mathcal{X}} x f_{X_k|\vec{q}_k=\vec{d}_k}\left(x\right) \mathrm{d}\, x.$$

Define $\tilde{x}_k = \hat{x}_k - x_k$ to be the estimation error and let $\tilde{X}_k$ be the random variable for which $\tilde{x}_k$ is the

realization. We define $\mathrm{Cov}\left(\tilde{X}_k\right)$ to be the covariance of $\tilde{X}_k$:

$$\mathrm{Cov}\left(\tilde{X}_k\right) \doteq \sum_{\vec{d}_k \in \{1 \ldots N\}^k} \mathrm{Prob}\left(\vec{q}_k = \vec{d}_k\right) \int_{\tilde{\mathcal{X}}} \left(\tilde{x}_k - \mathrm{E}\,\tilde{X}_k\right)^2 f_{X_k|\vec{q}_k=\vec{d}_k}\,(x)\,\mathrm{d}\,x$$

$$\mathrm{E}\,\tilde{X}_k \doteq \mathrm{E}_{\vec{q}_k}\,\mathrm{E}\,(X_k\,|\vec{q}_k\,),$$

where we used the following notation for expectation of an arbitrary function h on $\{1 \ldots N\}^k$:

$$\mathrm{E}_{\vec{q}_k}\,h\,(\vec{q}_k) \doteq \sum_{\vec{d}_k \in \{1 \ldots N\}^k} \mathrm{Prob}\left(\vec{q}_k = \vec{d}_k\right) h\left(\vec{d}_k\right).$$

We say that the state estimate is converging in mean square if

$$\lim_{k \to \infty} \mathrm{Cov}\left(\tilde{X}_k\right) = 0. \tag{5.28}$$

For a given $\sigma^2 > 0$ and a given $k \in \mathbb{N}$ we define $D\left(k, \sigma^2\right) \doteq \left\{\vec{d}_k \in \{1 \ldots N\}^k \,\middle|\, \mathrm{Cov}\left(\tilde{X}_k \,\middle|\vec{q}_k = \vec{d}_k\right) > \sigma^2\right\}$ to be the set of realizations of $\vec{q}_k$ for which the covariance of the estimation error at time k is larger than σ^2. We say that the state estimate is converging in probability if $\forall \sigma^2 > 0$:

$$\lim_{k \to \infty} \mathrm{Prob}\left(\vec{q}_k \in D\left(k, \sigma^2\right)\right) = 0.$$

We say that the state estimate is converging with probability $\rho \leq 1$ if $\forall \sigma^2 > 0$:

$$\lim_{k \to \infty} \mathrm{Prob}\left(\vec{q}_k \in D\left(k, \sigma^2\right)\right) \leq 1 - \rho.$$

Note that these definitions are given in decreasing order of strength.

Finally, we will use the standard definitions for entropy,

$$\mathrm{H}\left(p^{\vec{k}}\right) \doteq -\sum_{i=1}^{n} p_i \log p_i, \tag{5.29}$$

and for differential entropy,

$$\mathrm{H}\,(f) \doteq -\int_{\mathcal{X}} f\,(x) \log f\,(x)\,\mathrm{d}\,x. \tag{5.30}$$

All the logarithms in this paper are taken over the same base, b. We intentionally do not specify the base as our results hold with any choice of base.

5.3.2 Evaluating the Change in Entropy As a Necessary Condition

We start by stating and proving the following intuitive lemma:

Lemma 5.3.1. *The following is a necessary condition for having the state estimate converge in mean square:*

$$\lim_{k\to\infty} \mathrm{E}_{\vec{q}_k}\left(\mathrm{H}\left(f_{X_k|\vec{q}_k}\right)\right) = -\infty. \tag{5.31}$$

Proof: It is well known that the Gaussian distribution maximizes the differential entropy for a given covariance. Thus for any random variable X with the corresponding pdf f_X we have $\mathrm{H}(f_X) \leq \log\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2}$ where $\sigma^2 = \mathrm{Cov}(X)$. Now,

$$\begin{aligned}
\mathrm{Cov}\left(\tilde{X}_k\right) &= \int_{\mathcal{X}} x^2 \sum_{\vec{d}_k\in\{1,\ldots,N\}^k} \mathrm{Prob}\left(\vec{q}_k = \vec{d}_k\right) f_{\tilde{X}_k|\vec{q}_k=\vec{d}_k}(x)\,\mathrm{d}\,x \\
&= \sum_{\vec{d}_k\in\{1,\ldots,N\}^k} \mathrm{Prob}\left(\vec{q}_k = \vec{d}_k\right) \int_{\mathcal{X}} x^2 f_{\tilde{X}_k|\vec{q}_k=\vec{d}_k}(x)\,\mathrm{d}\,x \\
&= \mathrm{E}_{\vec{q}_k}\,\mathrm{Cov}\left(\tilde{X}_k \middle|\, \vec{q}_k\right) = \mathrm{E}_{\vec{q}_k}\,\mathrm{Cov}\left(X_k|\,\vec{q}_k\right).
\end{aligned} \tag{5.32}$$

Note that we used the fact that $\mathrm{E}\left(\tilde{X}_k\right) = 0$ as well as that $\mathrm{E}\left(\tilde{X}_k \middle| \vec{q}_k = \vec{d}_k\right) = 0\ \forall \vec{d}_k \in \{1,\ldots,N\}^k$. Continuing,

$$\mathrm{E}_{\vec{q}_k}\left(\mathrm{H}\left(f_{X_k|\vec{q}_k}\right)\right) \leq \mathrm{E}_{\vec{q}_k}\left(\frac{1}{2}\log\left(2\pi\,\mathrm{Cov}\left(X_k|\vec{q}_k\right)\right) + \frac{1}{2}\right) \leq \frac{1}{2}\log\left(2\pi\,\mathrm{E}_{\vec{q}_k}\left(\mathrm{Cov}\left(X_k|\vec{q}_k\right)\right)\right) + \frac{1}{2} \tag{5.33}$$

where the last inequality is due to Jensen's inequality and the concavity of the log function. If indeed the state estimate is converging in mean square, then from (5.28) and (5.32) the RHS of (5.33) must converge to $-\infty$. Therefore (5.31) must hold. ∎

The following is an immediate corollary of Lemma 5.3.1.

Corollary 5.3.2. *Assume that when using* (5.26) *the average entropy is increased by at least $\alpha > 0$ for every k:*

$$\mathrm{H}\left(f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) \geq \mathrm{H}\left(f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) + \alpha.$$

Assuming further that the number of bits available for the transmission of the information on the state location, $r_{bits/step}$, is fixed, then having the state estimate converge in mean square requires that $r_{bits/step} \geq \frac{\alpha}{\log 2}$ (or $r_{bits/step} \geq \alpha$ if base 2 is used for the logarithm).

Proof: According to (5.27) the decrease in average entropy when new information is arrived is equal to the entropy of $\vec{p^k}$:

$$\begin{aligned}\mathrm{E}_{\vec{q}_k}\left(\mathrm{H}\left(f_{X_k|\vec{q}_k=\vec{d}_k}\right)\right) &= \sum_{i=1}^{n} p_i \times -\int_{Q_i} \frac{f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}(x)}{p_i} \log\left(\frac{f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}(x)}{p_i}\right) \mathrm{d}\, x \\ &= \mathrm{H}\left(f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) - \mathrm{H}\left(\vec{p^k}\right).\end{aligned}$$

The entropy of $\vec{p^k}$ is maximized for the uniform distribution, $p_i^k = \frac{1}{N}\ \forall i \in \{1,\ldots,N\}$, in which case $\mathrm{H}\left(\vec{p^k}\right) = \log N$. From Lemma 5.3.1 we know we must have $\log N \geq \mathrm{H}\left(\vec{p^k}\right) \geq \alpha$. As $N = 2^{r_{bits/step}}$ we conclude that $r_{bits/step} \geq \frac{\alpha}{\log 2}$. ■

For linear systems, $x_{k+1} = ax_k + u$, it is straightforward to show that

$$\mathrm{H}\left(f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) = \mathrm{H}\left(f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) + \log a.$$

Thus to have the state estimate converge in mean square we must have $r_{bits/step} \geq \log_2 a$.

5.3.3 Evaluating the Change in Entropy As a Sufficient Condition

In this section we seek to find how the necessary condition (5.31) can be refined in order to make it also sufficient for having the state estimate converge. We had hoped to show that the necessary condition for having the state estimate converge in mean square, (5.31), is also sufficient as it is. However, we are yet to show that. Nevertheless, to our knowledge even what we do show here has not been shown before. The difficulty is that in general $\mathrm{H}(f_X) \to -\infty$ does not imply $\mathrm{Cov}(X) \to 0$. Consider for example the pdf

$$f_{X_\Delta}(x) = \begin{cases} \frac{1}{4\Delta} & |1-|x|| \leq \Delta \\ 0 & \text{otherwise} \end{cases} \tag{5.34}$$

with $\Delta < 1$. Its entropy, $\mathrm{H}(f_{X_\Delta}) = \log(4\Delta)$, diverges to $-\infty$ as $\Delta \to 0$. Its covariance, however, converges to 1 and not to 0.

Note that in the above example, as $\Delta \to 0$ the pdf becomes increasingly more "jagged." If the entropy diverges to $-\infty$, but at the same time the pdf also becomes increasingly "smooth," then we are able to show that $\mathrm{Cov}(X) \to 0$. We use the notion of *total variation* to measure the smoothness of the function. We

slightly alter the standard definition of the total variation of a function and define it as

$$\mathrm{TV}(f) \doteq \sup_{m\in\mathbb{N}} \sup_{\substack{x_0,\ldots,x_m \in S(f) \\ x_0 < x_1 < \ldots < x_m}} \sum_{i=1}^{m} |f(x_i) - f(x_{i-1})| \tag{5.35}$$

where

$$S(f) \doteq \left]\ \inf\{x \,|f(x) > 0\} \quad , \quad \sup\{x \,|f(x) > 0\}\ \right[.$$

The use of the reversed brackets is to emphasize that $S(\cdot)$ is an open set. The standard definition of total variation is recovered if $S(f)$ in (5.35) is replaced with the domain of f. With the assumption that the set of points where f_X is not differentiable is at most countable, (5.35) is known to be equivalent to

$$\mathrm{TV}(f) \doteq \int_{S(f)\setminus D_0(f)} \left| \frac{\partial f(z)}{\partial z}_{|z=x} \right| \mathrm{d}\, x + \sum_{x\in D_0(f)\cap S(f)} \Delta f(x) \tag{5.36}$$

where

$$\Delta f(x) \doteq \left| \lim_{z\nearrow x} f(z) - \lim_{z\searrow x} f(z) \right|$$

and $D_0(f)$ is the set of points where f is not differentiable. Note that by intersecting $D_0(f)$ with $S(\cdot)$ we exclude the initial (final) jump of $f(x)$ from (to) zero, if the function is zero before (after) this jump.

Proposition 5.3.1. *Define*

$$\mu(f_X) \doteq b^{-\mathrm{H}(f_X)} - \mathrm{TV}(f_X).$$

If $\mu(f_X) > 0$ *then*

$$\mathrm{Cov}(X) \leq \frac{1}{12\mu(f_X)^2}.$$

Recall that b is the base over which the logarithm is defined.

Proof: First we show that for a given finite value of the entropy, the maximal value the distribution function attains is bounded from below. Let $c = \sup_x f_X(x)$ which implies $f_X(x) \leq c\ \forall x$. Then,

$$\mathrm{H}(f_X) = -\int_{\mathcal{X}} f_X(x) \log(f_X(x))\, \mathrm{d}\, x \geq -\int_{\mathcal{X}} f_X(x) \log(c)\, \mathrm{d}\, x = -\log(c).$$

Therefore we must have $c \geq b^{-\mathrm{H}(f_X)}$.

From (5.35) it is now obvious that

$$f_X(x) \geq b^{-\mathrm{H}(f_X)} - \mathrm{TV}(f_X) = \mu(f_X), \qquad \forall x \in S(f_X).$$

This implies (see Proposition 5.3.4 at the end of this section) that $\mathrm{Cov}(X) \leq \frac{1}{\mu(f_X)}$. ■

We are now ready to state our main result:

Theorem 5.3.3. *Assume that the entropy between time steps increases by no more than* $\log \bar{a}$*:*

$$\mathrm{H}\left(f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) \leq \mathrm{H}\left(f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) + \log \bar{a}, \qquad \forall k \in \mathbb{N},\, \forall \vec{d}_{k-1} \in \{1,\ldots,N\}^{k-1}.$$

Assume the quantization is such that at time steps the average entropy is reduced by at least $\mathrm{H}_p > \log \bar{a}$*:*

$$\mathrm{E}_{d_k} \mathrm{H}\left(f_{X_k|\vec{q}_k=[\vec{d}_{k-1},d_k]}\right) \leq \mathrm{H}\left(f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) - \mathrm{H}_p, \qquad \forall k \in \mathbb{N},\, \forall \vec{d}_{k-1} \in \{1,\ldots,N\}^{k-1}.$$

Assume the map F *is such that* $\partial F/\partial x \geq \underline{c} > 1$ *and* $\left|\partial^2 F/\partial x^2\right| \leq \bar{c}$ $\forall x \in \mathcal{X}$ *and* $\forall u \in \mathcal{U}$*. Then the state estimate converges with probability* ρ *-* $\forall \sigma > 0$*:*

$$\lim_{k\to\infty} \mathrm{Prob}\left(\vec{q}_k \left| \mathrm{Cov}\left(\tilde{X}_k|\vec{q}_k\right) > \sigma^2\right.\right) \leq 1 - \rho \tag{5.37}$$

where

$$\rho = 1 - \left(\frac{\log \bar{a} - \mathrm{H}_p - \log \min_{i,k} p_i^k - \log \underline{c}}{-\left(\log \min_{i,k} p_i^k + \log \underline{c}\right)}\right).$$

A few observations. If we use equally probable quantization regions, $p_1^k = p_2^k = \ldots = p_N^k$, $\forall k$, then (5.37) becomes

$$\lim_{k\to\infty} \mathrm{Prob}\left(\vec{q}_k \left| \mathrm{Cov}\left(\tilde{X}_k|\vec{q}_k\right) > \sigma^2\right.\right) \leq \frac{\log \bar{a} - \log \underline{c}}{\log N - \log \underline{c}}$$

where the right-hand side (RHS) can be made arbitrarily small by taking a larger N. If we further assume F is a linear map, then $\underline{c} = \bar{a} = \bar{c}$ and (5.37) becomes

$$\lim_{k\to\infty} \mathrm{Prob}\left(\vec{q}_k \left| \mathrm{Cov}\left(\tilde{X}_k|\vec{q}_k\right) > \sigma^2\right.\right) = 0$$

when $N > \bar{a}$. In this case the state estimate converges in probability.

Before proving Theorem (5.3.3) we state two easy to prove propositions:

Proposition 5.3.2. *If for a random variable, X, we are given its expectation, $\mathrm{E}\,X$, and its minimum possible value, $\min X$, then the following holds:*

$$Pr\left(X \leq \eta\right) \leq \frac{\mathrm{E}\,X - \min X}{\eta - \min X}.$$

Proposition 5.3.3. *Let X and Y be two random variables, then $\forall \eta$ and $\forall \delta$:*

$$\mathrm{Prob}\left(X - Y < \eta\right) \leq \mathrm{Prob}\left(X \leq \eta + \delta\right) + \mathrm{Prob}\left(Y \geq \delta\right).$$

Proof of Theorem 5.3.3: From the assumptions in the Theorem we have

$$\mathrm{E}_{\vec{q}_k}\,\mathrm{H}\left(f_{X_k|\vec{q}_k}\right) \leq \mathrm{H}\left(f_{X_0}\right) + k\left(\log \bar{a} - \mathrm{H}_p\right).$$

We also have

$$\begin{aligned}
\max_{\vec{d}_k \in \{1,\ldots,N\}^k} \sup\left(f_{X_k|\vec{q}_k=\vec{d}_k}\right) &\leq \max_{\vec{d}_{k-1} \in \{1,\ldots,N\}^{k-1}} \max_{i \in \{1,\ldots,N\}} \frac{1}{p_i^k} \sup\left(f_{X_k|\vec{q}_{k-1}=\vec{d}_{k-1}}\right) \\
&\leq \max_{i \in \{1,\ldots,N\}} \frac{1}{p_i^k \underline{c}} \max_{\vec{d}_{k-1} \in \{1,\ldots,N\}^{k-1}} \sup\left(f_{X_{k-1}|\vec{q}_{k-1}=\vec{d}_{k-1}}\right),
\end{aligned}$$

which by recursion implies

$$\max_{\vec{d}_k \in \{1,\ldots,N\}^k} \sup\left(f_{X_k|\vec{q}_k=\vec{d}_k}\right) \leq \frac{1}{\left(\min_i p_i\right)^k \underline{c}^k} \sup\left(f_{X_0}\right).$$

Now, since

$$\mathrm{H}\left(f_X\right) = -\int_{\mathcal{X}} f_X\left(x\right) \log f_X\left(x\right) \mathrm{d}\,x \geq -\log \sup_x f_X\left(x\right) \int_{\mathcal{X}} f_X\left(x\right) \mathrm{d}\,x = -\log \sup_x f_X\left(x\right),$$

we can then write

$$\min_{\vec{q}_k} \mathrm{H}\left(f_{X_k|\vec{q}_k}\right) \geq -\log f_{X_0} + k \log \min_i p_i + k \log \underline{c}.$$

Now for the total variation, from (5.27) we have

$$\mathrm{E}_{\vec{q}_k}\,\mathrm{TV}\left(f_{X_k|\vec{q}_k}\right)=$$

$$\mathrm{E}_{\vec{q}_{k-1}}\sum_{i\in\{1,\ldots,N\}}p_i^k\times\left(\int_{\left(S\left(f_{X_k|\vec{q}_{k-1}}\right)\cap\left(Q_i^k\right)^\circ\right)\setminus D_0\left(f_{X_k|\vec{q}_{k-1}}\right)}\frac{1}{p_i^k}\left|\frac{\partial f_{X_k|\vec{q}_{k-1}}(z)}{\partial z}_{|z=x}\right|\mathrm{d}\,x+\frac{1}{p_i^k}\sum_{x\in D_0\left(f_{X_k|\vec{q}_{k-1}}\right)\cap S\left(f_{X_k|\vec{q}_{k-1}}\right)\cap\left(Q_i^k\right)^\circ}\Delta f_{X_k|\vec{q}_{k-1}}(x)\right)=$$

$$\mathrm{E}_{\vec{q}_{k-1}}\,\mathrm{TV}\left(f_{X_k|\vec{q}_{k-1}}\right)$$

where we used $\left(Q_i^k\right)^\circ$ to denote the open interior of Q_i^k. Note that had we used the original definition of total variation, the average total variation would have increased due to the possible new jumps in $f_{X_k|[\vec{q}_{k-1},i]}$, $i=1,\ldots,N$, on the boundaries of Q_i^k, which did not exist in $f_{X_k|\vec{q}_{k-1}}$. Because our modified definition of total variation only considers the interior of the support of each pdf, these possible new jumps are excluded. Similarly, from (5.26) we have

$$\begin{aligned}\mathrm{E}_{\vec{q}_{k-1}}\,\mathrm{TV}\left(f_{X_k|\vec{q}_{k-1}}\right)=&\,\mathrm{E}_{\vec{q}_{k-1}}\int_{F\left(S\left(f_{X_{k-1}|\vec{q}_{k-1}}\right),u_{k-1}\right)\setminus F\left(D_0\left(f_{X_{k-1}|\vec{q}_{k-1}}\right),u_{k-1}\right)}\left|\frac{\partial}{\partial x}\frac{f_{X_{k-1}|\vec{q}_{k-1}}\left(F^{-1}(x,u_{k-1})\right)}{\left|\partial F(z,u_{k-1})/\partial z_{|z=F^{-1}(x,u_{k-1})}\right|}\right|\mathrm{d}\,x+\\&\sum_{x\in F\left(D_0\left(f_{X_{k-1}|\vec{q}_{k-1}}\right)\cap S\left(f_{X_{k-1}|\vec{q}_{k-1}}\right),u_{k-1}\right)}\Delta\frac{f_{X_{k-1}|\vec{q}_{k-1}}\left(F^{-1}(x,u_{k-1})\right)}{\left|\partial F(z,u_{k-1})/\partial z_{|z=F^{-1}(x,u_{k-1})}\right|}\\\le&\,\mathrm{E}_{\vec{q}_{k-1}}\int_{S\left(f_{X_{k-1}|\vec{q}_{k-1}}\right)\setminus D_0\left(f_{X_{k-1}|\vec{q}_{k-1}}\right)}\left|\frac{f'_{X_{k-1}|\vec{q}_{k-1}}(x)}{F'(x,u_{k-1})}\right|+\left|\frac{f_{X_{k-1}|\vec{q}_{k-1}}(x)\,F''(x,u_{k-1})}{F'^2(x,u_{k-1})}\right|\mathrm{d}\,x+\\&\sum_{x\in D_0\left(f_{X_k|\vec{q}_{k-1}}\right)\cap S\left(f_{X_k|\vec{q}_{k-1}}\right)}\frac{\Delta f_{X_{k-1}|\vec{q}_{k-1}}(x)}{|F'(x,u_{k-1})|}\le\frac{1}{\underline{c}}E_{\vec{q}_{k-1}}TV\left(f_{X_{k-1}|\vec{q}_{k-1}}\right)+\frac{\bar{c}}{\underline{c}^2}.\end{aligned}\tag{5.38}$$

Therefore, by recursion,

$$\mathrm{E}_{\vec{q}_k}\,\mathrm{TV}\left(f_{X_k|\vec{q}_k}\right)\le\frac{1}{\underline{c}^k}\,\mathrm{TV}\left(f_{X_0}\right)+\left(\sum_{i=0}^{k-1}\frac{1}{\underline{c}^i}\right)\frac{\bar{c}}{\underline{c}^2}.$$

We also trivially have

$$\min_{\vec{q}_k}\mathrm{TV}\left(f_{X_k|\vec{q}_k}\right)\ge 0.$$

Now we analyze the random variable, $\mu\left(f_{X_k}\right)$. Using Propositions 5.3.3 and 5.3.2 we can get for every δ,

$$\begin{aligned}
&\operatorname{Prob}\left(\mu\left(f_{X_k}\right) \leq \eta\right) \leq \operatorname{Prob}\left(b^{-\operatorname{H}\left(f_{X_k}\right)} \leq \eta+\delta\right)+\operatorname{Prob}\left(\operatorname{TV}\left(f_{X_k}\right) \geq \delta\right) \\
&\quad=\operatorname{Prob}\left(\operatorname{H}\left(f_{X_k}\right) \geq-\log\left(\eta+\delta\right)\right)+\operatorname{Prob}\left(\operatorname{TV}\left(f_{X_k}\right) \geq \delta\right) \\
&\quad \leq \frac{\operatorname{E}\operatorname{H}\left(f_{X_k}\right)-\min \operatorname{H}\left(f_{X_k}\right)}{-\log\left(\eta+\delta\right)-\min \operatorname{H}\left(f_{X_k}\right)}+\frac{\operatorname{E}\operatorname{TV}\left(f_{X_k}\right)-\min \operatorname{TV}\left(f_{X_k}\right)}{\delta-\min \operatorname{TV}\left(f_{X_k}\right)} \\
&\quad \leq \frac{\left(\begin{array}{l} \operatorname{H}\left(f_{X_0}\right)+\log \sup f_X(x)+k\left(\log \bar{a}-\operatorname{H}\left(\vec{p}\right)-\log \min_i p_i-\log \underline{c}\right) \end{array}\right)}{\left(\begin{array}{l} -\log\left(\eta+\delta\right)+\log \sup f_X(x)-k\left(\log \min_i p_i+\log \underline{c}\right) \end{array}\right)}+\frac{\operatorname{TV}\left(f_{X_0}\right)}{\underline{c}^k \delta}+\left(\sum_{i=0}^{k-1} \frac{1}{\underline{c}^i}\right) \frac{\bar{c} \delta}{\underline{c}^2}.
\end{aligned}$$

In the limit as $k \rightarrow \infty$ and we get

$$\lim_{k \rightarrow \infty} \operatorname{Prob}\left(\mu\left(f_{X_k}\right) \leq \eta\right) \leq \frac{\log \bar{a}-\operatorname{H}\left(\vec{p}\right)-\log \min_i p_i-\log \underline{c}}{-\left(\log \min_i p_i+\log \underline{c}\right)}+\frac{1}{1-\underline{c}} \frac{\bar{c}}{\underline{c}^2 \delta}.$$

As this is true $\forall \delta$, we get that

$$\lim_{k \rightarrow \infty} \operatorname{Prob}\left(\mu\left(f_{X_k}\right) \leq \eta\right) \leq \frac{\log \bar{a}-\operatorname{H}\left(\vec{p}\right)-\log \min_i p_i-\log \underline{c}}{-\left(\log \min_i p_i+\log \underline{c}\right)}.$$

We complete the proof by using Proposition 5.3.1 which implies

$$\operatorname{Prob}\left(\vec{q}_k \left| \operatorname{Cov}\left(\tilde{X}_k | \vec{q}_k\right)>\sigma\right.\right) \leq \operatorname{Prob}\left(\mu\left(f_{X_k}\right) \leq \frac{1}{\sqrt{12 \sigma}}\right).$$

■

5.3.4 Discussion

In this work we sought to find sufficient conditions for convergence of an estimation error using arbitrary quantization. The sufficient conditions we derived only apply to scalar systems. They rely on the property of the total variation that, as it goes to zero, the function becomes constant. While extensions of the total variation to higher dimensions do exist, those that we are aware of do not posses this property. We also note that our sufficient conditions coincide with the necessary conditions only for linear systems with equiprobable quantization. We mention [7] as a potentially different approach for deriving sufficient conditions in more general settings.

5.3.5 Technical Result

Proposition 5.3.4. *Let X be a random variable and f_X its corresponding pdf. Assume f_X is piecewise continuous. Given the constraint that $f_X(x) \geq \mu > 0 \ \forall x \in S(f_X)$ the maximum covariance of X is $\frac{1}{12\mu^2}$. It is attained when $f_X(x) = \mu \ \forall x \in S(f_X)$.*

Proof: Let f_X be a pdf which satisfies the constraint. Without loss of generality we can assume $\mathrm{E}\, X = 0$. If the set of points where $f_X > \mu$ is of measure zero, we can change f_X at these points to μ without changing the covariance. If the set of points where $f_X > \mu$ is of measure larger than zero, then by the assumption of piecewise continuity we can find an interval $I \subset S(f_X)$ such that $f_X(x) > d > \mu \ \forall x \in I$. Let $\beta = \sup S(f_X)$. Without loss of generality (due to symmetry), and by shrinking I if necessary, we can assume $I \subset [0, \beta - \tau)$ for some $\tau > 0$. Let l be the length of the interval I. Define f_{X_ϵ}, $\epsilon < d - \mu$ as follows:

$$f_{X_\epsilon}(x) = \begin{cases} f_X(x) - \epsilon & x \in I \\ \mu & x \in \left[\beta, \beta + \frac{\epsilon l}{\mu}\right] \\ f_X(x) & \text{otherwise} \end{cases}.$$

This new function is also a pdf which satisfies the constraint. Its covariance is

$$\begin{aligned} \mathrm{Cov}(X_\epsilon) &= \mathrm{E}\, X_\epsilon^2 - (\mathrm{E}\, X_\epsilon)^2 \\ &= \int_{S(f_X)} x^2 f_X(x)\,\mathrm{d}\,x - \int_I x^2 \epsilon\,\mathrm{d}\,x + \int_{\left[\beta, \beta + \frac{\epsilon l}{\mu}\right]} x^2 \mu\,\mathrm{d}\,x - \left(\mathrm{E}\, X - \int_I x\epsilon\,\mathrm{d}\,x + \int_{\left[\beta, \beta + \frac{\epsilon l}{\mu}\right]} x\mu\,\mathrm{d}\,x\right)^2 \\ &\geq \mathrm{Cov}(X) - (\beta - \tau)^2 \epsilon l + \beta^2 \epsilon l - \left(\beta + \frac{\epsilon l}{\mu}\right)^2 \epsilon^2 l^2. \end{aligned} \tag{5.39}$$

It is now easy to see that for sufficiently small ϵ we can make $\mathrm{Cov}(X_\epsilon) > \mathrm{Cov}(X)$. Thus the covariance is maximized only if $f_X(x) = \mu \ \forall x \in S(f_X)$ (except for a set of measure 0). For a uniform distribution over an interval of length $1/\mu$ the covariance becomes $\frac{1}{12\mu^2}$. ∎

Chapter 6

Conclusions

In this work we considered the byproducts that arise when using advanced sensing technology, namely quantized and faulty measurements. We developed and analyzed control tools that are adapted to these types of measurements. We proved the stability of these control tools to external disturbances, modeling uncertainties and delays, and we proved their robustness to faulty measurements. We demonstrated the applicability of these tools to state estimation from faulty GPS measurements and to automatic landing control using quantized vision-based measurements. The results we derived here are also applicable to remote sensing over limited bandwidth communication channels.

In the first chapter we showed how to achieve input-to-state stability with respect to external disturbances using measurements from a dynamic quantizer. We showed that our technique is applicable to output feedback, stable under modeling errors and delays, and can work with data rates arbitrarily close to the minimum data rate needed for unperturbed systems. We also showed that our technique can be extended to nonlinear systems. Ours is only the second work to consider disturbance attenuation for quantized systems in the sense of input-to-state stability, the first one to do so in the context of minimum data rates, and the only one to provide such comprehensive stability results for quantized systems.

In the second chapter we proved that the MSoD estimator, which was known to be *robust* with respect to corruption, is also *stable* with respect to noise. We developed new algorithms to quantify the robustness and stability properties of this estimator for deterministic matrices. Where an alternative algorithm already existed, we showed that in the settings of interest, our algorithm is more efficient. We also demonstrated the benefits of this estimator for estimation in a dynamical system.

In the last chapter we considered a vision-based control system and demonstrated the implications of using the feedback of such a system. We derived a linear-time varying model that is observable using quantities extracted from the image. We showed that due to the quantized nature of the measurements, a classic estimator will not converge fast enough to achieve the desired control requirements. We then developed a general estimator that is adapted to quantized measurements, proved its convergence, and showed its success in a vision-based landing control system.

6.1 Future Research

We showed that the dynamic quantization controller we developed creates a control system which is stable to external disturbances, modeling uncertainties, and delays, and can be used in nonlinear settings. However, we only considered systems that exhibit one or two of these properties. This was done mainly to simplify the derivation, and the next step would be to combine these results and show that our controller maintains stability in systems exhibiting all of these properties simultaneously. With respect to external disturbances, we explicitly derived the stability gain of the system. With respect to modeling uncertainties and delays, we only showed how the stability gains can be derived, but did not derive them explicitly. We also only showed that there exist strictly positive bounds on the modeling uncertainties and delays under which the stability is maintained, but, again, did not derive them explicitly. An addition to this work will be to derive the explicit formulas in these cases. Finally, our proof follows a worst-case analysis. It would be interesting to find the average response of the system given probabilistic characteristics of the disturbance.

The results for the MSoD estimator, dealing with faulty measurements, also followed a worst-case analysis. Here too, it would be beneficial to analyze what is the probability that a set of faulty measurements could arbitrarily corrupt the estimate. We also found out that by differentiating between the weight that each measurement is given, better performance can be achieved. However, an algorithm to find the best weighting has not been developed yet and is another direction for future research.

Given the preliminary stage of the tools we developed for a vision-based control system with unknown dynamics, this can lead to many new research directions. The first step would be to prove that our estimator not only converges but converges to the right value. The second step would be to prove that the stability of the closed-loop system is maintained when this estimator is being used. The third step would be to compute the optimal initial open-loop perturbation that will guarantee the fast convergence of the estimator. Once the first two steps, at least, are completed, several extensions should be pursued. One is a computationally faster solver for the nonlinear optimization problem that is the basis of our estimator, possibly using the non-smooth results we derived. Another extension is to improve the performance by integrating the tool we developed for dynamic quantization and using dynamic control of the camera's zoom and position. And yet another extension is to integrate the results for faulty measurements in order to deal with possible miss-detections by the computer vision algorithm.

References

[1] J. Baillieul, "Feedback designs in information-based control," in *Stochastic Theory and Control, LNCIS*, B. Pasik-Duncan, Ed. Berlin-Heidelberg, Germany: Springer-Verlag, 2002, pp. 35–57.

[2] M. J. Best and B. Ding, "On the continuity of the minimum in parametric quadratic programs," *J. Optimization Theory Application*, vol. 86, no. 1, pp. 245–250, 1995.

[3] O. Bourquardez and F. Chaumette, "Visual servoing of an airplane for alignment with respect to a runway," in *Proc. 2007 IEEE Int. Conf. on Robotics and Automation*, 2007, pp. 1330–1335.

[4] R. W. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Trans. Automatic Control*, vol. 45, no. 7, pp. 1279–1280, 2000.

[5] E. J. Candès and P. A. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Information Theory*, vol. 54, no. 7, pp. 2829–2840, 2008.

[6] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[7] T. P. Coleman, "A stochastic control viewpoint on 'posterior matching-style' communication schemes," in *Proc. 2009 IEEE Int. Symp. Information Theory*, 2009, pp. 1520–1524.

[8] J. Cortés, S. Martínez, and F. Bullo, "Spatially-distributed coverage optimization and control with limited-range interactions," *ESAIM: Control, Optimization Calculus Variation*, vol. 11, pp. 691–719, 2005.

[9] C. De Persis, "On stabilization of nonlinear systems under data rate constraints using output measurements," *Int. J. Robust Nonlinear Control*, vol. 16, no. 6, pp. 315–322, 2006.

[10] C. De Persis and F. Mazenc, "Stability of quantized time-delay nonlinear systems: A Lyapunov-Krasowskii-functional approach," in *Proc. 48th IEEE Conf. on Decision and Control*, 2009, pp. 4093–4098.

[11] D. F. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Trans. Automatic Control*, vol. 35, no. 8, pp. 916–924, 1990.

[12] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," Dept. of Statistics, Stanford Univ., Tech. Rep. 2005-4, 2005.

[13] D. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution," *Comm. Pure Applied Mathematics*, vol. 59, no. 7, pp. 907–934, March 2006.

[14] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *J. American Mathematical Society*, vol. 22, no. 1, pp. 1–53, 2009.

[15] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] Flightgear, an open-source flight simulator. [Online]. Available: http://www.flightgear.org

[17] R. A. Freeman and P. V. Kokotovic, "Global robustness of nonlinear systems to state neasurements disturbances," in *Proc. 32nd IEEE Conf. on Decision and Control*, 1993, pp. 1507–1512.

[18] R. A. Freeman and P. V. Kokotović, *Robust Nonlinear Control Design: State-Space and Lyapunov Techniques*. Boston, MA: Birkhäuser, 1996.

[19] E. Fridman and M. Dambrine, "Control under quantization, saturation and delay: An LMI approach," *Automatica*, vol. 10, no. 9, pp. 2258–2264, 2009.

[20] R. Z. Has'minskii, *Stochastic Stability of Differential Equations*. Alphen aan den Rijn, The Netherlands: Sijthoff & Noordhoff, 1980.

[21] T. Hayakawa, H. Ishii, and K. Tsumura, "Adaptive quantized control for nonlinear uncertain systems," *Systems Control Letters*, vol. 58, no. 9, pp. 625–632, 2009.

[22] J. P. Hespanha, A. Ortega, and L. Vasudevan, "Towards the control of linear systems with minimum bit-rate," in *Proc. 15th Int. Symp. Mathematical Theory Networks Systems*, 2002.

[23] H. Ishii and T. Başar, "Quantization in $\mathcal{H}^\infty$ parameter identification," *IEEE Trans. Automatic Control*, vol. 53, no. 9, pp. 2186–2192, 2008.

[24] H. Ishii, T. Başar, and R. Tempo, "Randomized algorithms for quadratic stability of quantized sampled-data systems," *Automatica*, vol. 40, no. 5, pp. 839–846, 2004.

[25] Z. P. Jiang, A. R. Teel, and L. Praly, "Small-gain theorem for ISS systems and applications," *Mathematics Control Signals Systems*, vol. 7, no. 2, pp. 95–120, 1994.

[26] Z. P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, 2001.

[27] R. Kalman, "Nonlinear aspects of sampled-data control systems," *Proc. Symp. Nonlinear Circuit Analysis*, pp. 273–313, 1956.

[28] T. Kameneva, "Robust stabilization of control systems with quantized feedback," Ph.D. dissertation, The University of Melbourne, 2008.

[29] T. Kameneva and D. Nešić, "On l_2 stabilization of linear systems with quantized control," *IEEE Trans. Automatic Control*, vol. 53, no. 1, pp. 399–405, 2008.

[30] M. Krstić and H. Deng, *Stabilization of Nonlinear Uncertain Systems*. London, UK: Springer-Verlag, 1998.

[31] D. Liberzon, "Hybrid feedback stabilization of systems with quantized signals," *Automatica*, vol. 39, no. 9, pp. 1543–1554, 2003.

[32] D. Liberzon, "On stabilization of linear systems with limited information," *IEEE Trans. Automatic Control*, vol. 48, no. 2, pp. 304–307, 2003.

[33] D. Liberzon, "Quantization, time delays, and nonlinear stabilization," *IEEE Trans. Automatic Control*, vol. 51, no. 7, pp. 1190–1195, 2006.

[34] D. Liberzon and J. P. Hespanha, "Stabilization of nonlinear systems with limited information feedback," *IEEE Trans. Automatic Control*, vol. 50, no. 6, pp. 910–915, 2005.

[35] D. Liberzon and D. Nešić, "Input-to-state stabilization of linear systems with quantized state measurements," *IEEE Trans. Automatic Control*, vol. 52, no. 5, pp. 767–781, 2007.

[36] H. P. Lopuhaä and P. J. Rousseeuw, "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices," *Ann. Statistics*, vol. 19, no. 1, pp. 229–248, 1991.

[37] N. C. Martins, "Finite gain l_p stabilization is impossible by bit-rate constrained feedback," in *Proc. 9th Int. Workshop on Hybrid Systems: Computation and Control, LNCS 3927*, J. Hespanha and A. Tiwari, Eds. Berlin-Heidelberg, Germany: Springer-Verlag, 2006, pp. 451–459.

[38] N. C. Martins, A. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Trans. Automatic Control*, vol. 51, no. 3, pp. 438–447, 2006.

[39] A. S. Matveev and A. V. Savkin, "Stabilization of stochastic linear plants via limited capacity stochastic communication channel," in *Proc. 45th IEEE Conf. on Decision and Control*, 2006, pp. 484–489.

[40] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[41] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[42] A. Miller, M. Shah, and D. Harper, "Landing a UAV on a runway using image registration," in *Proc. 2008 IEEE Int. Conf. Robotics Automation*, 2008, pp. 182–187.

[43] R. K. Miller, M. S. Mousa, and A. N. Michel, "Quantization and overflow effects in digital implementations of linear dynamic controllers," *IEEE Trans. Automatic Control*, vol. 33, no. 7, pp. 698–704, 1988.

[44] G. N. Nair and R. J. Evans, "Stabilization with data-rate-limited feedback:tightest attainable bounds," *Systems Control Letters*, vol. 41, no. 1, pp. 49–56, 2000.

[45] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM J. Control Optimization*, vol. 43, no. 2, pp. 413–436, 2004.

[46] G. N. Nair, R. J. Evans, I. M. Y. Mareels, and W. Moran, "Topological feedback entropy and nonlinear stabilization," *IEEE Trans. Automatic Control*, vol. 49, no. 9, pp. 1585–1597, 2004.

[47] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[48] D. Nešić, A. R. Teel, and E. D. Sontag, "Formulas relating $\mathcal{KL}$ stability estimates of discrete-time and sampled-data nonlinear systems," *Systems Control Letters*, vol. 38, no. 1, pp. 49–60, 1999.

[49] A. Okao, M. Ikeda, and R. Takahashi, "System identification for nano control: A finite wordlength problem," in *Proc. 2003 IEEE Conf. Control Applications*, 2003, pp. 49–53.

[50] I. R. Petersen and A. V. Savkin, "Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel," in *Proc. 40th IEEE Conf. on Decision and Control*, 2001, pp. 304–309.

[51] A. A. Proctor and E. N. Johnson, "Vision-only approach and landing," in *Proc. AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2005.

[52] D. Ralph and S. Dempe, "Directional derivatives of the solutions of a parametric nonlinear program," *Mathematical Programming*, vol. 70, pp. 159–172, 1995.

[53] S. M. Robinson, "Local structure of feasible sets in nonlinear programming, part iii: Stability and sensitivity," in *Nonlinear Analysis and Optimization*, ser. Mathematical Programming Studies, B. Cornet, V. H. Ngyuen, and J. P. Vial, Eds. Amsterdam, The Netherlands: North-Holland, 1987, vol. 30, pp. 45–66.

[54] R. Sailer and F. Wirth, "Stabilization of nonlinear systems with delayed data-rate-limited feedback," in *Proc. European Control Conf. 2009*, 2009, pp. 1734–1739.

[55] A. V. Savkin, "Analysis and synthesis of networked control systems: topological entropy, observability, robustness, and optimal control," *Automatica*, vol. 42, no. 1, pp. 51–62, 2006.

[56] M. S. Selig, R. Deters, and G. Dimock. (2002) Aircraft dynamics models for use with flightgear. [Online]. Available: http://www.ae.illinois.edu/m-selig/apasim/Aircraft-uiuc.html

[57] Y. Sharon and D. Liberzon, "Input-to-state stabilization with minimum number of quantization regions," in *Proc. 46th IEEE Conf. on Decision and Control*, 2007, pp. 20–25.

[58] Y. Sharon and D. Liberzon, "Input-to-state stabilization with quantized output feedback," in *Proc. 11th Int. Conf. on Hybrid Systems: Computation and Control, LNCS 4981*, M. Egerstedt and B. Mishra, Eds. Berlin-Heidelberg, Germany: Springer-Verlag, 2008, pp. 500–513.

[59] Y. Sharon and D. Liberzon, "Input to state stabilizing controller for systems with coarse quantization," *IEEE Trans. Automatic Control*, 2010, to appear. [Online]. Available: http://www.ysharon.info/papers/quantization.pdf

[60] Y. Sharon and D. Liberzon, "Stabilization of linear systems under coarse quantization and time delays," in *Proc. 2nd IFAC Workshop Distributed Estimation Control Networked Systems*, 2010, pp. 31–36.

[61] Y. Sharon, D. Liberzon, and Y. Ma, "Adaptive control using quantized measurements with application to vision-only landing control," in *Proc. 49th IEEE Conf. on Decision and Control*, 2010, to appear.

[62] Y. Sharon, J. Wright, and Y. Ma, "Minimum sum of distances estimator: robustness and stability," in *Proc. 2009 American Control Conf.*, 2009, pp. 524–530.

[63] E. D. Sontag, "Smooth stabilization implies coprime factorization," *IEEE Trans. Automatic Control*, vol. 34, no. 4, pp. 435–443, 1989.

[64] E. D. Sontag, *Mathematical Control Theory. Deterministic Finite-Dimensional Systems, 2nd edition.* New York, NY: Springer-Verlag, 1998.

[65] B. L. Stevens and F. L. Lewis, *Aircraft Control and Simulation, 2nd edition.* Hoboken, NJ: John Wiley & Sons, Inc., 2003.

[66] H. Sun, N. Hovakimyan, and T. Başar, "$\mathcal{L}_1$ adaptive controller for quantized systems," 2010, submitted.

[67] H. Sun, N. Hovakimyan, and T. Başar, "$\mathcal{L}_1$ adaptive controller for systems with input quantization," in *Proc. 2010 American Control Conf.*, 2010, pp. 253–258.

[68] H. Suzuki and T. Sugie, "System identification based on quantized i/o data corrupted with noises and its performance improvement," in *Proc. 45th IEEE Conf. on Decision and Control*, 2006, pp. 3684–3689.

[69] S. Tatikonda and S. Mitter, "Control over noisy channels," *IEEE Trans. Automatic Control*, vol. 49, no. 7, pp. 1196–1201, 2004.

[70] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Automatic Control*, vol. 49, no. 7, pp. 1056–1068, 2004.

[71] A. R. Teel, "Connections between Razumikhin-type theorems and the ISS nonlinear small gain theorem," *IEEE Trans. Automatic Control*, vol. 43, no. 7, pp. 960–964, 1998.

[72] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Royal Statistical Society Series B*, vol. 58, pp. 267–288, 1996.

[73] K. Tsumura, "Optimal quantization of signals for system identification," *IEEE Trans. Automatic Control*, vol. 54, no. 12, pp. 2909–2915, 2009.

[74] L. Vu and D. Liberzon, "Stabilizing uncertain systems with dynamic quantization," in *Proc. 47th IEEE Conf. on Decision and Control*, 2008, pp. 4681–4686.

[75] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints-ii: stabilization with limited information feedback," *IEEE Trans. Automatic Control*, vol. 44, no. 5, pp. 1049–1053, 1999.