Controller Redesign to Minimize Uniform Quantization Errors in Uncertain Linear Systems with Fixed Hardware Constraints[⋆]

Mircea Șușcă[a,*], Vlad Mihaly[a], Zsófia Lendek[a], Irinel-Constantin Morărescu[b,a], Petru Dobra[a]

[a] *Technical University of Cluj-Napoca, Str. Memorandumului, Nr. 28, Cluj-Napoca, 400114, Romania*
[b] *Université de Lorraine, CNRS, CRAN, F-54000, Nancy, France*

Abstract

We consider the classical emulation paradigm in which a controller is already designed for a linear time-invariant plant. Motivated by implementation constraints in real applications, we analyze the effects of ubiquitous low-cost quantizers on the closed-loop dynamics. Consequently, we address the robust control problem of an uncertain discrete-time linear process using a regulator affected by the effects of uniform quantization performed by the input-output converters and arithmetical unit. In this setup with fixed hardware resolutions, the regulator's state-space realization is balanced to minimize the process' state quantization error while simultaneously maintaining its desired transient response. To characterize the quantization error, we provide an ultimate bound for its worst-case scenario using the input-to-state stability framework. The minimization is performed using off-the-shelf tools, with a characterization of the resulting problem. Finally, a comparative numeric case study showing the tightness of the computed bound is discussed.

Keywords: quantized control, linear uncertain systems, input-to-state stability, state-space balancing, asymptotic gain.

1. Introduction

Digital implementation of continuous-time regulators requires the sampling, discretization, and quantization of system coefficients and of the involved signals. These operations generally affect both the transient and steady-state responses of the closed-loop system. There are several types of quantizer circuits, such as uniform, logarithmic, delta-sigma, spherical polar coordinate, generalizing to the implementation of arbitrary countable sets (Gray and Neuhoff, 1998; Wang, 2021; Fu, 2024). The feedback connection of quantized subsystems can lead to highly-nonlinear behaviour, with effects such as static steady-state errors, limit cycles, or chaotic behaviour, see (Delchamps, 1990; Franklin et al., 2006). This engineering problem has lead to two complementary research directions. The first focuses on adequate hardware to ensure closed-loop asymptotic stability, with notable design tools proposed by (Brockett and Liberzon, 2000; Liberzon, 2003, 2006; Fu and Xie, 2005, 2009; Liu et al., 2015; Zhou et al., 2010; Ferrante et al., 2015; Xia et al., 2020; Wang, 2021; Wang et al., 2022; Xu et al., 2022) etc. The second considers simple Lyapunov stability, with deviations as small as possible from the ideal asymptotic case.

Our focus is on the second paradigm, specifically on minimizing the quantization effects on the process' state signals through a controller redesign, when the regulator is implemented on fixed (non-adaptive) hardware which induces uniform quantization errors. There are many applications, particularly in the automotive and aerospace industries (Dajsuren and van der Brand, 2019; Llorente, 2020), which impose strict hardware constraints due to which asymptotic stability cannot be ensured. Tightening hardware constraints in automotive systems is accompanied by

[⋆] This work was been financially supported by project DECIDE, no. 57/14.11.2022, contract number 760069, funded under the PNRR I8 scheme by the Romanian Ministry of Research, Innovation, and Digitisation.

[*] Corresponding author.

Email addresses: Mircea.Susca@aut.utcluj.ro (Mircea Șușcă), Vlad.Mihaly@aut.utcluj.ro (Vlad Mihaly), Zsofia.Lendek@aut.utcluj.ro (Zsófia Lendek), Constantin.Morarescu@univ-lorraine.fr (Irinel-Constantin Morărescu), Petru.Dobra@aut.utcluj.ro (Petru Dobra)

stricter control performance expectations, forcing designers to exploit limited computational, memory, and sensing resources to their fullest to meet safety, functional, and cyber security requirements (ISO, 2018; UNECE, 2020*a*,*b*). More so, it is of interest to focus on the effects of uniform quantizers due to their pervasiveness in practice, as they are the de facto standard in industrial and embedded systems, as seen in the previous references alongside (Bolton, 2021) and (Petkov et al., 2018). An emerging use case is in training and deployment of neural networks (Wang et al., 2021). Specific target microcontrollers are often mandated once validated and certified; additional software must be integrated into these existing platforms rather than producing custom hardware for each application, so they must be exploited to their fullest extent to remain as cost-effective as possible. Our objective is therefore to maximize the performance attainable from a given hardware platform.

Transient response degradation with respect to external disturbances has been studied by Kameneva and Nešić (2010), controller sampling rate and coefficients' quantization by Şuşcă et al. (2023*a*,*b*), and has been characterized using passivity indices (Xu et al., 2020) or mean-square stability (Cheng et al., 2019). We show that if the regulator state-space matrices are exactly represented, the quantization errors affect only the steady-state behaviour of the control system. Consequently, we consider the closed-loop steady-state performance degradation of the quantized system compared to its ideal counterpart. This aspect has been extensively examined in recent studies; however, relevant research gaps remain, as shown in what follows.

For steady-state effects, Franklin et al. (2006, Ch. 10) and Widrow and Kollár (2008, Ch. 17) use the Bertram bound to analyze the quantization error of the output signal, but the approach is mostly limited to open-loop analysis. De Souza et al. (2010) consider finite logarithmic quantizers to design invariant attractor sets, but the technique is applicable only for single-input single-output systems. Error bounds for linear control systems affected by uniform quantization, assuming a diagonalizable closed-loop state matrix, are given in (Şuşcă et al., 2022, 2024). In specific cases, such as in DC-to-DC converter control, tight steady-state bounds are achieved in (Peng et al., 2007), with a limit cycle rejection technique proposed by Abdullah et al. (2023). Another limiting practical assumption of all state signals being accessible is considered in the adaptive backstepping control from (Wang et al., 2022). For exact models, the impact of both uniform and logarithmic finite quantizers has been studied through achievable regions of attraction of the closed-loop trajectories (Campos et al., 2016), and systematic ultimate bounds on the system's states and outputs (Haimovich et al., 2007). Furthermore, the interaction between the controller and process is performed through zero-order hold circuits, but there exist specific use cases where other hold circuits are necessary, such as first- and second-order hold, sinc hold, delta-sigma, etc., (Proakis and Manolakis, 2006).

Robustness to uncertainties and disturbances has been considered in multiple works, but most are explicitly addressed through the use of adaptive hardware, see (Yoo and Park, 2021; Liu et al., 2015; Ferrante and Tarbouriech, 2024; Ferdinando et al., 2022). Practical stability of the quantized closed-loop system has been achieved by Hayakawa et al. (2009); Liberzon (2006); Ferdinando et al. (2024), but with the allowance of time-varying sampling intervals or non-uniform quantization in the input/output channels.

In the works above, the quantization effects are assumed to stem from the input/output converters or input/state only, but not from the full configuration arising in practice, where the internal computations of the regulator are also affected by quantization. All considered works assume no feedthrough matrix on the process model. Additionally, many consider state feedback control, which implies the assumption of a fully-measurable state, which is not feasible in many practical use cases with limited hardware capabilities. Most studies – an exception being (Ferdinando et al., 2024) – focus on quantifying the impact of the quantization effects, and not on the subsequent step of adapting the regulator to minimize such effects.

This is why, in this paper, we consider a wide class of discrete-time linear control systems with: (*i*) uncertainties and feedthrough in the process model; (*ii*) dynamic output feedback; (*iii*) the possibility to interface the process using zero-order hold, first order-hold, or other more sophisticated hold devices; (*iv*) quantization effects in the internal computations of the regulator caused by the arithmetic logic unit (ALU). This work is an extension of (Şuşcă et al., 2022, 2024) with the added contribution of: (*a*) formalizing the induced quantization errors using the input-to-state stability framework, (*b*) quantifying the effect of system uncertainties on the errors, (*c*) relaxing the assumptions on the computed error bounds, (*d*) providing a nontrivial extension from vector bounds to element-wise bounds, and (*e*) proposing a regulator balancing method which ensures tighter bounds without altering the desired transient response. In developing the results, we also provide an algebraic relationship between the Markov parameters of a stable LTI system and its $\mathcal{H}_\infty$-norm. We start from a *robust* (in the sense of (Skogestad and Postlethwaite, 2005)) discrete-time regulator with given input-output behaviour and fixed sampling rate. The main objective is to minimize the quantization errors on the process' state through controller redesign, considering static (memoryless) uniform quantization on the regulator's input-output converters and internal computations of its arithmetical unit. The main contributions are:

1. computation of analytic ultimate bounds on the (vector and element-wise) quantization errors of the states and outputs of the process;
2. decoupling the transient and steady-state response degradations of the closed-loop system with respect to the quantization of the regulator;
3. balancing of regulator state-space matrices to minimize the quantization error bounds without affecting the transient response.

Paper Structure: Section 2 reviews the necessary concepts on input-to-state stability of nonlinear systems and robust control of uncertain linear systems in discrete-time domain. Section 3 provides the problem statement and the required tools for quantized control systems. Section 4 develops the proposed analytic bound on the uniform quantization error, while Section 5 continues with its minimization through a controller balancing scheme. Section 6 illustrates the achievable tightness of the optimized bound on a numeric example and compares it to other results from the literature. Conclusions and further extensions are given in Section 7.

Notations: $\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$ are the sets of natural, real, and complex numbers. The subscript ≥ 0 (> 0) denotes nonnegative (positive) numbers. The integer part of $x \in \mathbb{R}^n$ is symbolized as $\lfloor x \rfloor$, with its fractional part written as $\{x\}$, applied element-wise. The general linear group of degree n is written $\mathrm{GL}_n(\mathbb{C})$. The symbol $\overset{S}{\sim}$ denotes a general change of coordinates through a matrix $S \in \mathrm{GL}_n(\mathbb{C})$; $\overset{P}{\sim}$ brings a square matrix $A \in \mathbb{C}^{n\times n}$ into its Jordan canonical form J_A, i.e., $A \overset{P}{\sim} J_A \Leftrightarrow A = P \cdot J_A \cdot P^{-1}$. Let $\mu(A) = \mu(J_A)$ denote the maximum Jordan block (cell) dimension of J_A. The spectral radius, eigenvalue set, and maximum singular value of A are denoted $\rho(A)$, $\Lambda(A)$ and $\overline{\sigma}(A)$. Its ith row is symbolized as $A(i, :)$. $\|\cdot\|$ denotes infinity-type norms. The Euclidean norm, i.e., spectral norm for matrices, is denoted $|\cdot|$, $|A| = \overline{\sigma}(A)$. We use the input-to-state stability notations from (Mironchenko, 2003), including the $\mathcal{K}$, $\mathcal{K}_\infty$, $\mathcal{L}$, $\mathcal{KL}$ sets of comparison functions, the space ℓ^n of sequences of dimension $n \in \mathbb{N}_{>0}$, and the space $\ell(\mathbb{R}^m, \mathbb{R}^p)$ of discrete proper linear operators from $\mathbb{R}^m$ to $\mathbb{R}^p$. A discrete-time multi-input multi-output (MIMO) linear time-invariant (LTI) system $G \in \ell(\mathbb{R}^m, \mathbb{R}^p)$ with fixed sampling rate $\tau > 0$ has a state-space representation (A, B, C, D) and equivalent input-output transfer matrix representation:

$$G(z) = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \overset{\text{def}}{=} D + C\,(zI - A)^{-1}\,B.$$

The $\mathcal{H}_\infty$-norm of a stable system G is defined as:

$$\|G\| \overset{\text{def}}{=} \sup_{\Omega\in[0,2\pi)} \left|G\left(e^{j\Omega}\right)\right| = \sup_{z\in\mathbb{C},|z|=1} \left|D + C\,(zI - A)^{-1}\,B\right|.$$

A list of the most important notations is provided in Table 1.

2. Mathematical Background

2.1. Input-to-State Stability

The following concepts are adapted from (Mironchenko, 2003, Ch. 2) and (Jiang and Wang, 2001).

Let Σ be a discrete-time system, with a sufficiently smooth map $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$, $k \in \mathbb{N}$:

$$\Sigma : x(k{+}1) = f\left(x(k), u(k)\right), \; x_0 = x(0) \in \mathbb{R}^n. \tag{1}$$

Definition 1. *System* (1) *is called forward complete if, for all $x_0 \in \mathbb{R}^{n_x}$ and all $u \in \ell^{n_u}$, the solution $\phi(k, x_0, u)$, corresponding to initial condition x_0 and bounded input u exists and is unique for $k \in \mathbb{N}$.*

Assume that $f(0, 0) = 0$, i.e., $x = 0$ is an equilibrium point of the 0-input system.

Notation	Meaning
$\Delta = (A_\Delta, B_\Delta, C_\Delta, D_\Delta)$	Arbitrary stable uncertainty model (4); $\Delta \equiv D_\Delta$ particularizes to a static uncertainty
$G^{(0)}, G = \mathrm{ULFT}(G^{(0)}, \Delta)$	Process with nominal dynamics (5) versus uncertain process (6) with $\Delta \in \boldsymbol{\Delta}$
$K^{(0)}, K$	Ideal and quantized regulator models, Eqns. (7), (10), based on (A_1, B_1, C_1, D_1)
$L^{(0)} = G^{(0)}K, L = GK$	Open-loop plant without and with uncertainty block Δ from (12)
$T^{(0)}, T = \mathrm{ULFT}(T^{(0)}, \Delta)$	Closed-loop system (13) based on $L^{(0)}$ and L and the feedback $e = r - y$
$\Upsilon = (\mathbf{\Phi}_\Delta, \mathbf{\Gamma}_{\Delta,\eta}, \mathbf{C}_\Delta, \mathbf{D}_{\Delta,\eta})$	Error dynamics (15) of the quantized versus ideal closed-loop systems $T - T\vert_{\eta\equiv 0}$
$\mathbf{\Phi}_\Delta = P_\Delta \cdot J_{\mathbf{\Phi}_\Delta} \cdot P_\Delta^{-1}$	Complex Jordan form of closed-loop state matrix with Jordan cells of dimensions $\{N_1, \dots, N_s\}$ and maximum cell size $\mu(\mathbf{\Phi}_\Delta) = \max\{N_i\}, i = \overline{1, s}, s \le n_\Xi$
$\mathcal{Q}_\Delta^\Xi(k), \mathrm{bound}(\overline{\mathcal{Q}_\Delta^\Xi})$	Quantization error of the state vector Ξ, (18) and its computed bound (19)
$\mathcal{Q}_\Delta^y(k), \mathrm{bound}(\overline{\mathcal{Q}_\Delta^y})$	Quantization error of the output vector y (24) and its computed bound (25)
$\mathrm{bound}(\overline{\mathcal{Q}_\Delta^{\Xi_i}}), \mathrm{bound}(\overline{\mathcal{Q}_\Delta^{y_i}})$	Quantization bounds used for element-wise states Ξ_i (32a) and outputs y_i (32b)
$\mathrm{bound}(\overline{\mathcal{Q}_\Delta^\Xi}(K, P_\Delta))$	Computed bound with emphasized dependency on K and P_Δ (33)
$K(S)$	Balanced regulator (34) using arbitrary similarity transformation $S \in \mathrm{GL}_{n_\xi}(\mathbb{R}) = \mathcal{S}$
$\mathcal{P}_\Delta(S)$	Set of all similarity transforms which lead $\mathbf{\Phi}_\Delta(S)$ to its Jordan form, as in (35)
$\mathcal{D}_\alpha$	Set of block diagonal matrices (37) based on which $\mathcal{P}_\Delta(S)$ can be spanned
$\tilde{S} = \mathrm{diag}\left(I_{n_x+n_\zeta}, S\right)$	Shorthand notation $\tilde{S} \in \mathrm{GL}_{n_\Xi}(\mathbb{R})$ (38) which allows simplified computations in (39)
$\mathcal{N}_\xi, \mathcal{N}_u, \overline{\mathcal{N}}_\xi, \overline{\mathcal{N}}_u$	$\mathcal{H}_\infty$-norms of the state, output signals of K, with maximum allowed values (40), (41)

Table 1: Main mathematical objects used throughout the paper

Definition 2. *System* (1) *is called input-to-state stable (ISS), if* (1) *is forward complete and there exist* $\beta \in \mathcal{KL}$ *and* $\gamma \in \mathcal{K}$ *such that for all* $x_0 \in \mathbb{R}^{n_x}$, $u \in \ell^{n_u}$ *bounded,* $k \in \mathbb{N}$, *the following inequality holds:*

$$|\phi(k, x_0, u)| \le \beta(|x_0|, k) + \gamma(\|u\|).$$

The function γ *is called the asymptotic gain of system* (1) *and describes the influence of the exogenous input on the system. The map* β *describes the transient behaviour of the system.*

Theorem 1. *If* $f(\cdot, 0)$ *from system* (1) *is Lipschitz continuous, global asymptotic stability at zero (0-GAS) is equivalent to the existence of* $\beta \in \mathcal{KL}$ *such as for all* $x_0 \in \mathbb{R}^{n_x}$ *and all* $k \in \mathbb{N}$ *it holds that:*

$$|\phi(k, x_0, 0)| \le \beta(|x_0|, k).$$

Any ISS system is 0-GAS. On the other hand, for any ISS system there exists $\gamma \in \mathcal{K}$ so that for any $x_0 \in \mathbb{R}^{n_x}$ and any bounded $u \in \ell^{n_u}$, the inequality:

$$\limsup_{k\to\infty} |\phi(k, x_0, u)| \le \gamma\left(\|u\|\right), \tag{2}$$

is satisfied. Condition (2) is called the *asymptotic gain* (AG) property. A system having the AG property will be called an AG system. Every trajectory of an AG system converges to the neighborhood of an origin with a radius $\gamma\left(\|u\|\right)$. Furthermore:

Theorem 2. *System* (1) *is ISS if and only if it is AG and its origin without inputs is stable in the sense of Lyapunov.*

2.2. Robust Control of Discrete-Time Linear Systems

The following robust control concepts are adapted from (Skogestad and Postlethwaite, 2005). In robust control, the *process uncertainties* can be classified in unstructured (to describe residual dynamics) and parametric (to model inaccurate component characteristics). Both types can be encompassed into a single block-diagonal structured matrix uncertainty from the set:

$$\boldsymbol{\Delta} = \left\{\Delta = \mathrm{diag}\left(\Delta_1^p I_{n_1}, \dots, \Delta_{\underline{F}}^p I_{n_{\underline{F}}}, \Delta_1^u, \dots, \Delta_{\overline{F}}^u\right), \Delta_i^p \in \mathbb{C},\ \Delta_j^u \in \mathbb{C}^{m_j \times m_j},\ 1 \le i \le \underline{F},\ 1 \le j \le \overline{F}\right\}, \tag{3}$$

where blocks $\Delta_i^p I_{n_i}$ are used for *parametric uncertainties* and blocks Δ_j^u are used for *unstructured uncertainties*. By convention, $|\Delta| < 1$, $\forall \Delta \in \mathbf{\Delta}$. Formulation (3) is natural in frequency domain, but in the time domain, the complex terms translate to fixed LTI models $\Delta(z)$. From now on, we will refer to uncertainties $\Delta \in \mathbf{\Delta}$ only through their equivalent time-domain dynamical formulations, defined next. The dynamics of an arbitrary stable uncertainty block $\Delta \in \ell(\mathbb{R}^{n_v}, \mathbb{R}^{n_d})$ of order $n_\zeta \in \mathbb{N}$, $\|\Delta\| < 1$, and initial condition $\zeta(0) \equiv 0$, is:

$$\Delta : \begin{cases} \zeta(k+1) &= A_\Delta \zeta(k) + B_\Delta v(k); \\ d(k) &= C_\Delta \zeta(k) + D_\Delta v(k). \end{cases} \tag{4}$$

The equivalence is achieved by the identity $\Delta(z) = D_\Delta + C_\Delta (zI - A_\Delta)^{-1} B_\Delta$. A static uncertainty model corresponds to a feedthrough matrix $\Delta \equiv D_\Delta$, $|D_\Delta| < 1$. In practice, particular examples of the model (4) can be obtained through Monte Carlo sampling if the worst-case frequency response of the uncertainty model is known. Methods to experimentally obtain such models from data can be found in (Hindi et al., 2002; Balas et al., 2009) for LTI systems, with an extension to descriptor systems developed in (Markiş et al., 2024).

Figure 1 presents the one degree-of-freedom closed-loop discrete-time control system composed of an uncertain process $G \in \ell(\mathbb{R}^{n_u}, \mathbb{R}^{n_y})$ and a regulator (*assumed ideal in this section*) $K^{(0)} \in \ell(\mathbb{R}^{n_y}, \mathbb{R}^{n_u})$. The ideal process, decoupled from the uncertain dynamics $\Delta \in \ell(\mathbb{R}^{n_v}, \mathbb{R}^{n_d})$, is denoted by $G^{(0)}$ and has the augmented interface $\begin{pmatrix} v^\top & y^\top \end{pmatrix}^\top = G^{(0)} \begin{pmatrix} d^\top & u^\top \end{pmatrix}^\top$, with a default state-space representation and initial condition $x_0 = x(0) \in \mathbb{R}^{n_x}$:

$$G^{(0)} \colon \begin{cases} x(k+1) &= A_2 x(k) + \ B_d\, d(k) + B_2\, u(k); \\ v(k) &= C_v x(k) + D_{vd} d(k) + D_{vu} u(k); \\ y(k) &= C_2 x(k) + D_{yd} d(k) + D_2\, u(k). \end{cases} \tag{5}$$

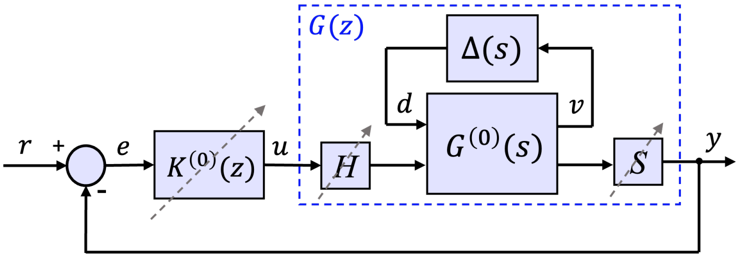

Figure 1: Closed-loop discrete-time system with process uncertainty Δ and quantized regulator K, i.e., the ideal regulator model $K^{(0)}$ implemented using non-ideal sample, hold, and arithmetic unit. This leads to quantization noise at the regulator's input (analog-to-digital converter, ADC), output (digital-to-analog converter, DAC), and internal computations (arithmetic logic unit, ALU).

The hardware setup from Figure 1 implies a zero-order hold discretization of the process. A causal adaptation can be determined, according to Franklin et al. (2006), to the first-order hold extrapolator also, which ultimately leads to different matrices in (5). The methodology remains adaptable to more specialized hold circuits. The signals $u \in \ell^{n_u}$ and $y \in \ell^{n_y}$ represent the command and measurement vectors, which are physically accessible, while $d \in \ell^{n_d}$ and $v \in \ell^{n_v}$ are the disturbance input and output vectors. The state vector is $x \in \ell^{n_x}$. For an arbitrary uncertainty model $\Delta \in \mathbf{\Delta}$ from (4), $d = \Delta v$, we compute the implicit form $y = Gu$ of the process using the upper linear fractional transformation (ULFT) connection $G = \mathrm{ULFT}\left(G^{(0)}, \Delta\right)$. Its expression, combining (5) with an arbitrary particular sample (4) from the uncertainty family, is:

$$G = \mathrm{ULFT}\left(G^{(0)}, \Delta\right) : \left(\begin{array}{c|c} A_{2,\Delta} & B_{2,\Delta} \\ \hline C_{2,\Delta} & D_{2,\Delta} \end{array}\right) \overset{\text{def}}{=} \left(\begin{array}{cc|c} A_2 + B_d \tilde{\mathbf{D}}^{-1} D_\Delta C_v & B_d \tilde{\mathbf{D}}^{-1} C_\Delta & B_2 + B_d \tilde{\mathbf{D}}^{-1} D_\Delta D_{vu} \\ B_\Delta \mathbf{D}^{-1} C_v & A_\Delta + B_\Delta \mathbf{D}^{-1} D_{vd} C_\Delta & B_\Delta \mathbf{D}^{-1} D_{vu} \\ \hline C_2 + D_{yd} D_\Delta \mathbf{D}^{-1} C_v & D_{yd} \tilde{\mathbf{D}}^{-1} C_\Delta & D_2 + D_{yd} D_\Delta \mathbf{D}^{-1} D_{vu} \end{array}\right), \tag{6}$$

with auxiliary notations $\mathbf{D} = I - D_{vd} D_\Delta$ and $\tilde{\mathbf{D}} = I - D_\Delta D_{vd}$. The ULFT connection is well-posed if both matrices $\mathbf{D}$ and $\tilde{\mathbf{D}}$ have full rank or, equivalently, the matrix $\begin{pmatrix} I & D_{vd} \\ D_\Delta & I \end{pmatrix}$ is invertible (Ionescu et al., 1999, Ch. 2). The state vector of the uncertain process is $\begin{pmatrix} x^\top & \zeta^\top \end{pmatrix}^\top \in \ell^{n_x+n_\zeta}$. The nominal process model can be recovered as $G^{(0)}\big|_{\Delta\equiv 0} \Leftrightarrow G|_{\Delta\equiv 0}$. This allows us to provide tailored quantization bound expressions in Section 4 to any system from the uncertainty family.

The ideal numeric regulator $K^{(0)}$ has an input-output mapping $u = K^{(0)}e$, using the error signal $e = r - y$ as input, and is connected to the process G in a feedback connection, with reference $r \in \mathbb{R}^{n_y}$. Its state-space representation is:

$$K^{(0)} : \begin{cases} \xi(k{+}1) & = A_1\xi(k) + B_1e(k); \\ u(k) & = C_1\xi(k) + D_1e(k), \end{cases} \tag{7}$$

with initial condition $\xi_0 = \xi(0) \in \mathbb{R}^{n_\xi}$, $e \in \ell^{n_y}$, $u \in \ell^{n_u}$, $\xi \in \ell^{n_\xi}$.

The problem of steady-state quantization has been shown by Haimovich et al. (2007) and Şuşcă et al. (2022) to be well-posed if the closed-loop system is stable. To ensure the validity of the study for the entire uncertainty family, we consider the robust stability (RS) property, i.e., regulator (7) ensures stability of all systems G against $\mathbf{\Delta}$. As such, we consider the assumption:

Assumption 1. *The already-designed regulator $K^{(0)}$ ensures robust stability of the uncertain system G for all $\Delta \in \mathbf{\Delta}$.*

We assume that a satisfactory controller (7) has already been designed and is given for the scope of the quantization analysis. This paradigm reflects common industrial practice, particularly in automotive and aerospace domains, where responsibilities are split across multiple teams: a design group develops the control law and certifies its functional properties, while separate implementation and integration teams adapt and deploy the controller on target hardware.

3. Uniform Quantized Linear Control Systems

Problem Statement: Let $K^{(0)}$ be an ideal nominal discrete-time regulator satisfying Assumption 1. Our objective is to perform a controller redesign to obtain the regulator $K(S)$ that minimizes the quantization errors induced by fixed hardware (see Figure 1) on the closed-loop system's states.

To solve the above problem we provide, in Section 4, a method to accurately characterize the errors induced by the quantization. Based on this characterization, we perform the controller redesign in Section 5.

In the following, we provide the model of the ideal regulator $K^{(0)}$ affected by the non-ideal hardware devices from Figure 1 inducing uniform quantization. The interface between the continuous-time process and discrete-time regulator is ensured through sample and hold circuits, like in (Chen and Francis, 1995; Haimovich et al., 2007) and (Ferdinando et al., 2024). Furthermore, we also consider that the internal computations from $K^{(0)}$ are affected by quantization noise, caused by a non-ideal arithmetic unit. This additional extension, normally not accounted for in the literature, is justified by the following motivating example.

Example 1. *Consider a perturbed version of controller* (7)*, where the signals are affected by the uniform quantization of the arithmetic unit (rounding and truncation), causing bounded uncorrelated additive noise vectors η_{ξ_1}, η_{ξ_2} of adequate sizes, see (Mullis and Roberts, 1976):*

$$K : \begin{cases} \tilde{\xi}(k{+}1) & = A_1\tilde{\xi}(k) + B_1e(k) + \eta_{\xi_1}(k); \\ \tilde{u}(k) & = C_1\tilde{\xi}(k) + D_1e(k) + \eta_{\xi_2}(k). \end{cases}$$

Each state can be written as $\tilde{\xi}(k) = \xi(k) + \eta_{\xi_1}(k{-}1)$. This leads to the perturbed command signal:

$$\tilde{u}(k) = C_1\left(\xi(k) + \eta_{\xi_1}(k{-}1)\right) + D_1e(k) + \eta_{\xi_2}(k).$$

The relative error of the command signal, assuming a non-zero forced equilibrium, i.e., $\lim_{k\to\infty} u(k) \neq 0$, is:

$$\frac{\|\tilde{u}(k) - u(k)\|}{\|u(k)\|} = \frac{\|C_1\eta_{\xi_1}(k{-}1) + \eta_{\xi_2}(k)\|}{\|C_1\xi(k) + D_1e(k)\|}. \tag{8}$$

The analysis of (8) *shows two possible quantization-sourced problems. First, the numerator can have non-negligible absolute errors in high-gain dynamic regulators (when $\|C_1\| \gg 1$). Second, arbitrarily large relative errors can appear in practical applications, e.g., computer numerical control machining, with variable arbitrary trajectories and steady-states. An equilibrium point implies $e(k) \to 0$, which means that state equilibria $\overline{\xi} \in \ker(C_1)$ cancels the numerator, leading to numeric singularities (Spong et al., 2020).* ■

To account for the quantization effects of the regulator hardware, consider mappings $Q(\chi, x) : \mathbb{R}_{>0} \times \mathbb{R}^n \to \mathbb{R}^n$. The argument χ is the resolution of the quantizer, with element-wise application on each component of the vector argument x. The two commonly-used functions are the *midtread* (*rounding*) and *midriser* (*truncation*), defined (Widrow and Kollár, 2008) as:

$$Q(\chi, x) = \chi \left\lfloor \frac{x}{\chi} + \frac{1}{2} \right\rfloor, \quad Q(\chi, x) = \chi \left(\left\lfloor \frac{x}{\chi} \right\rfloor + \frac{1}{2} \right). \tag{9}$$

We propose a rewriting of the quantization functions (9) in a unified additive manner as $Q(\chi, x) = x + \varphi(\chi, x) \odot \chi$, $\chi > 0$, $x \in \mathbb{R}^n$, with a considered uncertain but bounded term $\varphi(\chi, x) \in \left[-\frac{1}{2}, \frac{1}{2}\right]^n$, where $\varphi(\chi, x) = \frac{1}{2} - \left\{\frac{x}{\chi} + \frac{1}{2}\right\} \in \left[-\frac{1}{2}, \frac{1}{2}\right)^n$ for midtread, and $\varphi(\chi, x) = \frac{1}{2} - \left\{\frac{x}{\chi}\right\} \in \left(-\frac{1}{2}, \frac{1}{2}\right]^n$ for midriser, respectively. The symbol $\odot$ denotes element-wise multiplication. In practice, the truncation is usually translated by a half step, but it can be adapted back to its original form with average slope equal to one.

The hardware interfacing the process with the regulator from Figure 1 leads to quantization effects perturbing the input, state, and output signals of the ideal regulator $K^{(0)}$ from (7), as illustrated in Figure 2. Irrespective of the quantization function, the quantized regulator $K \in \ell\left(\mathbb{R}^{2n_y + n_\xi + 2n_u}, \mathbb{R}^{n_u}\right)$ can be modelled as:

$$K : \begin{cases} \xi(k+1) & = A_1 \xi(k) + B_1 e(k) + B_1 \eta_e(k) + \eta_{\xi_1}(k); \\ u(k) & = C_1 \xi(k) + D_1 e(k) + D_1 \eta_e(k) + \eta_{\xi_2}(k) + \eta_u(k), \end{cases} \tag{10}$$

with bounded disturbance inputs due to quantization as: $\eta_e \in \mathbb{R}^{n_y}$, $\|\eta_e\| = \frac{\chi_e}{2}$ (ADC), $\eta_{\xi_1} \in \mathbb{R}^{n_\xi}$, $\|\eta_{\xi_1}\| = \frac{\chi_\xi}{2}$ (ALU involved in state computations), $\eta_{\xi_2} \in \mathbb{R}^{n_u}$, $\|\eta_{\xi_2}\| = \frac{\chi_\xi}{2}$, (ALU involved in output computations) $\eta_u \in \mathbb{R}^{n_u}$, $\|\eta_u\| = \frac{\chi_u}{2}$ (DAC), for hardware resolutions $\chi_e, \chi_\xi, \chi_u \in \mathbb{R}_{>0}$. The disturbances can be lumped into a single vector:

$$\eta = \begin{pmatrix} \eta_e^\top & \eta_{\xi_1}^\top & \eta_{\xi_2}^\top & \eta_u^\top \end{pmatrix}^\top \in \ell^{n_y + n_\xi + 2n_u}. \tag{11}$$

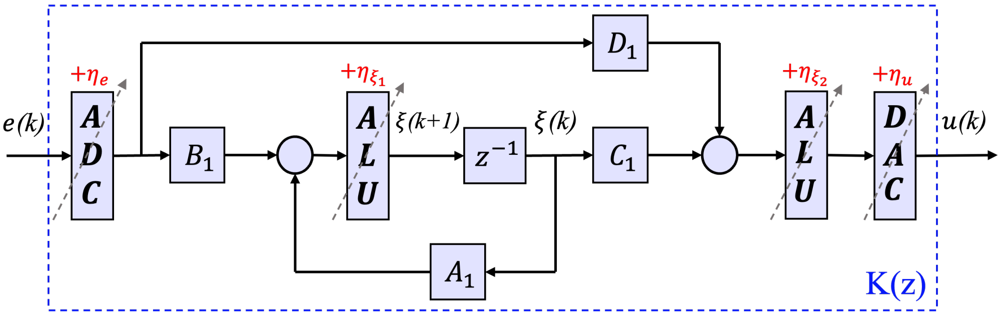

Figure 2: Ideal discrete-time controller (7) affected by quantization at its input (ADC), output (DAC), and internal computations (ALU).

We make the following assumption.

Assumption 2. *The reference r and matrices* (A_1, B_1, C_1, D_1) *from* (10) *are exactly represented.*

The reference signal must also be encoded in the same manner as the other signals involved in the controller's computations, irrespective of the fact that it may be provided by an external source. More so, coefficient encoding will not be further considered in this paper, as it is a separate problem which influences the transient response of the system, see (Şuşcă et al., 2023*a*). However, our analysis holds even for perturbed systems $(\tilde{A}_1, \tilde{B}_1, \tilde{C}_1, \tilde{D}_1) \approx (A_1, B_1, C_1, D_1)$ as long as RS is maintained for the entire uncertainty set $\mathbf{\Delta}$. The resulting non-ideal regulator has the extended input-output representation $u = K\begin{pmatrix}\eta^\top & e^\top\end{pmatrix}^\top$, with the ideal expression (7) recovered as $K^{(0)} = K|_{\eta \equiv 0}$.

4. Input-to-State Stability Characterization of System Quantization Errors

Using the setup from Section 3, we proceed to quantify the resulting quantization effects on the closed-loop system within the ISS framework. Using the quantized regulator K from (10), we obtain the open-loop system as the series connection $L^{(0)} = G^{(0)}K$, $\left(v^\top \quad y^\top\right)^\top = L^{(0)} \left(d^\top \quad \eta^\top \quad e^\top\right)^\top$, with an extended state Ξ of dimension $n_\Xi = n_x + n_\zeta + n_\xi$:

$$\Xi = \left(x^\top \quad \zeta^\top \quad \xi^\top\right)^\top .$$

The partial closed-loop with the uncertainty blocks $\Delta \in \mathbf{\Delta}$, $\|\Delta\| < 1$, is $L = \mathrm{ULFT}\left(L^{(0)}, \Delta\right)$, $d = \Delta v$, with the input-output representation $y = L\left(\eta^\top \quad e^\top\right)^\top$, $\quad L = GK$. Furthermore, the state-space representation of the open-loop plant can be expressed, based on (6), as:

$$L : \left(\begin{array}{cc|cccc:c} A_{2,\Delta} & B_{2,\Delta}C_1 & B_{2,\Delta}D_1 & O & B_{2,\Delta} & B_{2,\Delta} & B_{2,\Delta}D_1 \\ O & A_1 & B_1 & I & O & O & B_1 \\ \hline C_{2,\Delta} & D_{2,\Delta}C_1 & D_{2,\Delta}D_1 & O & D_{2,\Delta} & D_{2,\Delta} & D_{2,\Delta}D_1 \end{array}\right) \equiv \left(\begin{array}{c|c:c} A_{L,\Delta} & B_{L,\Delta,\eta} & B_{L,\Delta,e} \\ \hline C_{L,\Delta} & D_{L,\Delta,\eta} & D_{L,\Delta,e} \end{array}\right). \tag{12}$$

A feedback connection applied to L implies that the regulator input becomes $e = r - y$, which is a simplified form of the lower linear fractional transformation (LLFT) connection, see (Ionescu et al., 1999, Ch. 2). This leads to the closed-loop uncertain system model $T = \mathrm{ULFT}(T^{(0)}, \Delta)$, $y = T\left(\eta^\top \quad r^\top\right)^\top$. We make the following assumption.

Assumption 3. *Matrix* $E_\Delta \overset{\text{def}}{=} I + D_{L,\Delta,e} = I + D_{2,\Delta}D_1$ *is invertible.*

This assumption guarantees that the closed-loop system is well-posed. Thus, it is not restrictive in practice. Furthermore, regulators are usually designed such that the closed-loop system has *roll-off*, i.e., $D_{2,\Delta}D_1 = O$, or otherwise, the feedthrough gain is $\left|D_{2,\Delta}D_1\right| \ll 1$, to attenuate high-frequency noise. If the process or the regulator is strictly proper, i.e., $D_{2,\Delta} = O$ or $D_1 = O$, then $E_\Delta = I$. Due to this, in typical applications, the matrix E_Δ is diagonally dominant.

Then, the closed-loop uncertain model T has the state-space representation:

$$T : \begin{cases} \Xi(k{+}1) & = \mathbf{\Phi}_\Delta \Xi(k) + \mathbf{\Gamma}_\Delta \left(\eta^\top(k) \quad r^\top(k)\right)^\top ; \\ y(k) & = \mathbf{C}_\Delta \Xi(k) + \mathbf{D}_\Delta \left(\eta^\top(k) \quad r^\top(k)\right)^\top , \end{cases} \tag{13}$$

with initial condition $\Xi(0) = \left(\xi(0)^\top \quad x(0)^\top \quad \zeta(0)^\top\right)^\top \in \mathbb{R}^{n_\Xi}$. The closed-loop state and input matrices (well defined due to Assumption 3) are:

$$\mathbf{\Phi}_\Delta = A_{L,\Delta} - B_{L,\Delta,e}E_\Delta^{-1}C_{L,\Delta} = \begin{pmatrix} A_{2,\Delta} & B_{2,\Delta}C_1 \\ O & A_1 \end{pmatrix} - \begin{pmatrix} B_{2,\Delta}D_1 \\ B_1 \end{pmatrix} (I + D_{L,\Delta,e})^{-1} \left(C_{2,\Delta} \quad D_{2,\Delta}C_1\right);$$

$$\mathbf{\Gamma}_\Delta = \left(\begin{array}{c|c} B_{L,\Delta,\eta} & B_{L,\Delta,e} \end{array}\right) - B_{L,\Delta,e}\mathbf{D}_\Delta \equiv \left(\begin{array}{c|c} \mathbf{\Gamma}_{\Delta,\eta} & \mathbf{\Gamma}_{\Delta,r} \end{array}\right) = \left(\begin{array}{cccc|c} \mathbf{\Gamma}^e_{\Delta,\eta} & \mathbf{\Gamma}^{\xi_1}_{\Delta,\eta} & \mathbf{\Gamma}^{\xi_2}_{\Delta,\eta} & \mathbf{\Gamma}^u_{\Delta,\eta} & \mathbf{\Gamma}_{\Delta,r} \end{array}\right),$$

and the output and feedthrough matrices are:

$$\mathbf{C}_\Delta = E_\Delta^{-1}C_{L,\Delta} = \left(E_\Delta^{-1}C_{2,\Delta} \quad E_\Delta^{-1}D_{2,\Delta}C_1\right), \quad \mathbf{D}_\Delta = E_\Delta^{-1}\left(D_{L,\Delta,\eta} \quad D_{L,\Delta,e}\right) \equiv \left(\mathbf{D}_{\Delta,\eta} \quad \mathbf{D}_{\Delta,r} \right). \tag{14}$$

Denote the difference between the quantized closed-loop system's state $\Xi(k)$ and its ideal case counterpart as $\varepsilon(k) = \Xi(k) - \Xi(k)|_{\eta\equiv 0}$, and the difference between the corresponding outputs as $\delta(k) = y(k) - y(k)|_{\eta\equiv 0}$. The quantization error dynamics system Υ is then described by the state-space realization:

$$\Upsilon : \begin{cases} \varepsilon(k{+}1) & = \mathbf{\Phi}_\Delta \varepsilon(k) + \mathbf{\Gamma}_{\Delta,\eta}\eta(k); \\ \delta(k) & = \mathbf{C}_\Delta \varepsilon(k) + \mathbf{D}_{\Delta,\eta}\eta(k), \end{cases} \qquad \varepsilon_0 \in \mathbb{R}^{n_\Xi}. \tag{15}$$

Theorem 3. *Under Assumptions 1, 2, 3, the error dynamics* Υ *from* (15) *induced by uniform quantization, composed of system* G *from* (6) *and regulator* K *from* (10)*, is ISS.*

Proof 1. *Summing up the state equations of the error dynamics* (15) *for* $i = \overline{1, k{+}1}$ *scaled by the* $(k+1-i)$*th power of the state matrix* $\mathbf{\Phi}_\Delta$*, the* $(k{+}1)$*th sample depending on the initial state* $\varepsilon(0) = \varepsilon_0$ *is:*

$$\varepsilon(k{+}1) = \mathbf{\Phi}_\Delta^{k+1}\varepsilon(0) + \underbrace{\sum_{i=0}^{k} \mathbf{\Phi}_\Delta^i \mathbf{\Gamma}_{\Delta,\eta}\eta(k-i)}_{Q_\Delta^\Xi(k)}\,. \tag{16}$$

By definition, the state trajectory of the difference equation (15) *coincides with* (16)*. As such,* $\phi(k, \varepsilon_0, \eta) \overset{\text{def}}{=} \varepsilon(k)$ *in Definition 2. Applying the* ∞*-norm on* (16)*, we obtain:*

$$\|\phi\,(k, \varepsilon_0, \eta)\| \le \underbrace{\left\|\mathbf{\Phi}_\Delta^k\right\| \cdot \|\varepsilon_0\|}_{\beta(\|\varepsilon_0\|,k)} + \underbrace{\sum_{\vartheta\in\{e,\xi_1,\xi_2,u\}} \left\|\sum_{i=0}^{k-1} \mathbf{\Phi}_\Delta^i \mathbf{\Gamma}_{\Delta,\eta}^\vartheta\right\| \cdot \|\eta_\vartheta(k{-}1{-}i)\|}_{\gamma(\|\eta\|)}, \tag{17}$$

which corresponds to Definition 2 of the ISS property with the Euclidean norm replaced by the ∞*-norm. The term* β *asymptotically decreases to zero, as shown next. Let* $\mathbf{\Phi}_\Delta = UX_{\Phi_\Delta}U^\star$ *be the Schur canonical form, where* U *is a unitary matrix and* $X_{\Phi_\Delta} \in \mathbb{C}^{n_\Xi \times n_\Xi}$ *is the upper triangular form. Then we have* $\left\|\mathbf{\Phi}_\Delta^k\right\| = \left\|X_{\Phi_\Delta}^k\right\|$*. We write* $X_{\Phi_\Delta} = D_X + F_X$*, where* D_X *and* F_X *are the main diagonal and upper triangular part of matrix* X_{Φ_Δ}*, respectively. Because* $F_X^{n_\Xi}$ *is the null matrix, we have:*

$$\left\|\mathbf{\Phi}_\Delta^k\right\| = \left\|X_{\Phi_\Delta}^k\right\| \le \sum_{i=0}^{n_\Xi-1} \binom{k}{i} \|D_X\|^{k-i} \|F_X\|^i = \sum_{i=0}^{n_\Xi-1} \binom{k}{i} \rho\,(\mathbf{\Phi}_\Delta)^{k-i} \|F_X\|^i\,,$$

where $\binom{k}{i}$ *is the notation for combinations. According to Assumption 1,* $\rho\,(\mathbf{\Phi}_\Delta) < 1$ *for all* $\Delta \in \mathbf{\Delta}$*. Then:*

$$\lim_{k\to\infty} \binom{k}{i} \rho\,(\mathbf{\Phi}_\Delta)^{k-i} = 0,$$

which, alongside (17)*, completes the proof.* ■

In conclusion, for $k \to \infty$, the quantization error is given only by the asymptotic gain $\gamma\,(\|\eta\|)$. Note that for $k = 0$, the initial process states $x(0)$ and $\zeta(0) = 0$ are invariant to the quantized and ideal regulators alike. If, furthermore, $\xi(0) = \xi(0)|_{\eta(0)\equiv 0}$, then $\varepsilon_0 = 0$ in (17) and the quantization error coincides with the asymptotic gain $\gamma\,(\|\eta\|)$ for any $k \in \mathbb{N}$, not just for the steady-state behaviour. Thanks to Assumption 2, the external reference signal vanishes from (15) and it does not influence the quantization error.

We further proceed to obtain an absolute bound on the asymptotic gain of the error dynamics. Define the ideal (non-conservative) upper bound of the partial sum:

$$\overline{Q_\Delta^\Xi} = \sup_{k\ge 0} \left\|Q_\Delta^\Xi(k)\right\|.$$

Denote the set of indices ϑ found in the outer sum of (17) as $\Theta \overset{\text{def}}{=} \{e, \xi_1, \xi_2, u\}$. We rewrite the input-dependent term from (16) as:

$$Q_\Delta^\Xi(k) = \sum_{\vartheta\in\Theta} \sum_{i=0}^{k-1} \mathbf{\Phi}_\Delta^i \mathbf{\Gamma}_{\Delta,\eta}^\vartheta \eta_\vartheta(k{-}1{-}i). \tag{18}$$

An absolute bound for the worst-case state quantization error can be computed based on the Jordan canonical form of the closed-loop state matrix $\mathbf{\Phi}_\Delta$, as folows:

Theorem 4 (State error vector bound). *Given* $\Delta \in \mathbf{\Delta}$*, the state* ε *in* (15) *is asymptotically bounded by:*

$$\overline{Q_\Delta^\Xi} \le \sum_{\vartheta\in\Theta} \left(\sum_{i=1}^{\mu(\mathbf{\Phi}_\Delta)} \frac{1}{(1-\rho\,(\mathbf{\Phi}_\Delta))^i} \right) \cdot \|P_\Delta\| \cdot \left\|P_\Delta^{-1}\mathbf{\Gamma}_{\Delta,\eta}^\vartheta\right\| \cdot \frac{\chi_\vartheta}{2} \overset{\text{def}}{=} \text{bound}\left(\overline{Q_\Delta^\Xi}\right), \tag{19}$$

with the (complex) Jordan canonical form $\mathbf{\Phi}_\Delta = P_\Delta \cdot J_{\mathbf{\Phi}_\Delta} \cdot P_\Delta^{-1}$ *and* $\mu(\mathbf{\Phi}_\Delta)$ *is the maximum Jordan cell dimension.*

The proof is found in Appendix A. Theorem 4 allows us to quantify the steady-state quantization error in terms of the infinity norm instead of the Euclidean norm, which would only provide a qualitative assessment in our case.

Remark 1. *It can be seen from* (19) *that* bound$\left(\overline{Q^{\Xi}_{\Delta}}\right)$ *depends on the regulator K and, through matrices $\mathbf{\Phi}_{\Delta}$, $\mathbf{\Gamma}_{\Delta,\eta}$, on the similarity transform P_{Δ}.*

Next, the quantization error of the uncertain model (6), $\Delta \not\equiv 0$, is shown to be separable into the sum of its nominal counterpart's error and a bounded term caused by the uncertainty. To state the next result, we introduce the order of growth notation $O(\cdot)$. A vector-valued function $\mathcal{F}(\mathbf{x}, \psi)$ is said to be $O(\psi)$ on a compact set $D_{\mathbf{x}} \times [0, \psi^{\star}]$ if there exist constants $c, \psi^{\star} > 0$ such that:

$$|\mathcal{F}(\mathbf{x}, \psi)| \leq c\psi, \ \forall \psi \in \left[0, \psi^{\star}\right], \ \forall \mathbf{x} \in D_{\mathbf{x}}.$$

Theorem 5 (Uncertain state error vector bound). *For an arbitrary $\Delta \in \mathbf{\Delta}$, the state ε in* (15)*, reduced to the process states x and regulator states ξ, is asymptotically bounded by the sum of the nominal quantization error and the size of the uncertainty as:*

$$\overline{Q^{\Xi}_{\Delta}} \leq \left.\overline{Q^{\Xi}_{\Delta}}\right|_{\Delta\equiv 0} + O\left(\|\Delta\|\right). \tag{20}$$

The proof is found in Appendix B. The above result shows that for sufficiently small uncertainties $\Delta \in \mathbf{\Delta}$, the quantization error of the uncertain system is well approximated by the nominal one's.

A bound on the worst-case quantization error of the process measurements can be similarly derived. Combining (15) with (16), the output dynamics can be rewritten as:

$$\delta(k) = \mathbf{C}_{\Delta}\varepsilon(k) + \mathbf{D}_{\Delta,\eta}\eta(k) = \mathbf{C}_{\Delta}\mathbf{\Phi}^{k}_{\Delta}\varepsilon(0) + \underbrace{\sum_{i=0}^{k-1} \mathbf{C}_{\Delta}\mathbf{\Phi}^{i}_{\Delta}\mathbf{\Gamma}_{\Delta,\eta}\eta(k-1-i) + \mathbf{D}_{\Delta,\eta}\eta(k)}_{Q^{y}_{\Delta}(k)}. \tag{21}$$

Applying the ∞-norm on expression (21), similarly to (17), we obtain the ISS-type formulation:

$$\|\delta(k)\| \leq \underbrace{\left\|\mathbf{C}_{\Delta}\mathbf{\Phi}^{k}_{\Delta}\right\| \cdot \|\varepsilon_0\|}_{\beta^{y}(\|\varepsilon_0\|, k)} + \underbrace{\sum_{\vartheta\in\Theta}\Big(\left\|\mathbf{D}^{\vartheta}_{\Delta,\eta}\right\| \cdot \|\eta_{\vartheta}(k)\| + \left\|\sum_{i=0}^{k-1} \mathbf{C}_{\Delta}\mathbf{\Phi}^{i}_{\Delta}\mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\right\| \cdot \|\eta_{\vartheta}(k-1-i)\|\Big)}_{\gamma^{y}(\|\eta\|)}. \tag{22}$$

For an uncertainty block $\Delta \in \mathbf{\Delta}$, define the ideal (non-conservative) upper bound of the above error as:

$$\overline{Q^{y}_{\Delta}} = \sup_{k\geq 0}\left\|Q^{y}_{\Delta}(k)\right\|. \tag{23}$$

We rewrite the input-dependent term from (21):

$$Q^{y}_{\Delta}(k) = \sum_{\vartheta\in\Theta}\left[\mathbf{D}^{\vartheta}_{\Delta,\eta}\eta_{\vartheta}(k) + \sum_{i=0}^{k-1} \mathbf{C}_{\Delta}\mathbf{\Phi}^{i}_{\Delta}\mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\eta_{\vartheta}(k-1-i)\right]. \tag{24}$$

An ultimate bound for the ideal worst-case output quantization error (23) can be computed as follows.

Corollary 1 (Output error vector bound). *Given $\Delta \in \mathbf{\Delta}$, the output δ in* (15) *is asymptotically bounded by:*

$$\overline{Q^{y}_{\Delta}} \leq \sum_{\vartheta\in\{e,\xi_1,\xi_2,u\}}\left(\left(\sum_{i=1}^{\mu(\mathbf{\Phi}_{\Delta})} \frac{1}{\left(1-\rho\left(\mathbf{\Phi}_{\Delta}\right)\right)^{i}}\right) \cdot \|\mathbf{C}_{\Delta}P_{\Delta}\| \cdot \left\|P^{-1}_{\Delta}\mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\right\| + \left\|\mathbf{D}^{\vartheta}_{\Delta,\eta}\right\|\right) \cdot \frac{\chi_{\vartheta}}{2} \overset{\text{def}}{=} \text{bound}\left(\overline{Q^{y}_{\Delta}}\right). \tag{25}$$

The framework of Corollary 1 can be further adapted to provide element-wise bounds for each state $\Xi_i \in \Xi$, $i = \overline{1, n_{\Xi}}$, and output $y_i \in y$, $i = \overline{1, n_y}$ from (15). This adaptation is not straightforward, as we need to maintain the feedback dynamics unaffected by the required matrix modifications to isolate individual errors. To isolate a worst-case bound for each component, we start from (21). Taking (14) into account, (21) becomes:

$$E_\Delta \delta(k) = C_{L,\Delta} \mathbf{\Phi}_\Delta^k \varepsilon(0) + D_{L,\Delta,\eta}\eta(k) + \sum_{i=0}^{k-1} C_{L,\Delta} \mathbf{\Phi}_\Delta^i \mathbf{\Gamma}_{\Delta,\eta}\eta(i). \tag{26}$$

To isolate output y_i, the pre-multiplications in (26) are replaced by:

$$E_\Delta^{y_i} = I(i,:) + D_{2,\Delta}^{y_i} D_1; \quad C_{\Delta,L}^{y_i} = \begin{pmatrix} C_{2,\Delta}^{y_i} & D_{2,\Delta}^{y_i} C_1 \end{pmatrix}; \quad D_{\Delta,L}^{y_i} = \left(\begin{array}{cccc|c} D_{2,\Delta}^{y_i} D_1 & O & D_{2,\Delta}^{y_i} & D_{2,\Delta}^{y_i} & D_{2,\Delta}^{y_i} D_1 \end{array} \right), \tag{27}$$

where, adapting from (5) and (6):

$$C_{2,\Delta}^{y_i} = \begin{pmatrix} C_2(i,:) + D_{yd}(i,:) D_\Delta \mathbf{D}^{-1} C_v & D_{yd}(i,:) \tilde{\mathbf{D}}^{-1} C_\Delta \end{pmatrix}; \quad D_{2,\Delta}^{y_i} = D_2(i,:) + D_{yd}(i,:) D_\Delta \mathbf{D}^{-1} D_{vu}, \tag{28}$$

and $\mathbf{D}$, $\tilde{\mathbf{D}}$ remain unchanged. The closed-loop matrices $\mathbf{\Phi}_\Delta$ and $\mathbf{\Gamma}_{\Delta,\eta}$ remain unaffected. Similarly, to highlight the states Ξ_i in (26), $i = \overline{1, n_\Xi}$, we replace in (28):

$$E_\Delta^{\Xi_i} \leftarrow I(i,:); \quad C_2(i,:) \leftarrow \begin{pmatrix} 0 & \dots & 0 & \underbrace{1}_{i} & 0 & \dots & 0 \end{pmatrix}; \quad D_2(i,:) \leftarrow 0; \quad D_{yd}(i,:) \leftarrow 0, \tag{29}$$

leading to the output matrices $C_{2,\Delta}^{\Xi_i}$, $D_{2,\Delta}^{\Xi_i}$ and $E_\Delta^{\Xi_i}$. The quantization errors $\delta^{\Xi_i}(k)$ and $\delta^{y_i}(k)$, respectively, have the dynamics:

$$E_\Delta^{\Xi_i} \delta^{\Xi_i}(k) = C_{L,\Delta}^{\Xi_i} \mathbf{\Phi}_\Delta^k \varepsilon(0) \underbrace{\sum_{j=0}^{k-1} C_{L,\Delta}^{\Xi_i} \mathbf{\Phi}_\Delta^j \mathbf{\Gamma}_{\Delta,\eta} \eta(k-1-j) + D_{L,\Delta,\eta}^{\Xi} \eta(k)}_{Q_\Delta^{\Xi_i}(k)}, \ i = \overline{1, n_\Xi}, \ k \geq 0; \tag{30}$$

$$E_\Delta^{y_i} \delta^{y_i}(k) = C_{L,\Delta}^{y_i} \mathbf{\Phi}_\Delta^k \varepsilon(0) \ + \underbrace{\sum_{j=0}^{k-1} C_{L,\Delta}^{y_i} \mathbf{\Phi}_\Delta^j \mathbf{\Gamma}_{\Delta,\eta} \eta(k-1-j) + D_{L,\Delta,\eta}^{y_i} \eta(k)}_{Q_\Delta^{y_i}(k)}, \ i = \overline{1, n_y}, \ k \geq 0. \tag{31}$$

Denote by $\text{bound}\left(\overline{Q_\Delta^{\Xi_i}}\right)$ and $\text{bound}\left(\overline{Q_\Delta^{y_i}}\right)$ the supremum of the terms indicated in (30) and (31), similarly to (19) and (25). We now proceed to provide ultimate element-wise quantization error bounds for the state and output signals of the closed-loop system T from (13).

Corollary 2 (Error element-wise bounds). *For an arbitrary* $\Delta \in \mathbf{\Delta}$*, each state* Ξ_i*,* $i = \overline{1, n_\Xi}$*, and output* δ_i*,* $i = \overline{1, n_y}$*, in* (15) *is asymptotically bounded by:*

$$\overline{Q_\Delta^{\Xi_i}} \leq \text{bound}\left(\overline{Q_\Delta^{\Xi_i}}\right), \ i = \overline{1, n_\Xi}; \tag{32a}$$

$$\overline{Q_\Delta^{y_i}} \leq \sum_{j=1}^{n_y} \left(\left| \tilde{e}_{ij} \right| \cdot \text{bound}\left(\overline{Q_\Delta^{y_i}}\right) \right), \ i = \overline{1, n_y}, \tag{32b}$$

where $\tilde{e}_{ij}$ *are the elements of the matrix inverse of* E_Δ*, i.e.,* $E_\Delta^{-1} = \left[\tilde{e}_{ij}\right]$, $1 \leq i, j \leq n_y$.

The proof is found in Appendix C. In practice, Corollary 2 shows, on the one hand, that the steady-state quantization error is directly influenced by the spectral radius of the closed-loop state matrix. On the other hand, it allows the computation of individual quantization errors based on an algebraic test on the regulator and process state-space matrices, alongside the hardware configuration parameters χ_e, χ_ξ, χ_u. For the best-case scenario of Assumption 3, i.e., $E_\Delta = I$, usually found in applications, (32b) is reduced to a decoupled version.

5. Controller Balancing for Asymptotic Gain Minimization

Emphasizing the dependency on the regulator K and the similarity matrix P_Δ, based on Remark 1, the state vector bounds from Section 4 can be written as:

$$\text{bound}\left(\overline{Q_\Delta^\Xi}\right) = \text{bound}\left(\overline{Q_\Delta^\Xi}(K, P_\Delta)\right). \tag{33}$$

This section proposes an adequate scaling of the regulator matrices (A_1, B_1, C_1, D_1) and a way to find the similarity matrix P_Δ, given K, to ensure optimized bounds of $\overline{Q_\Delta^\Xi}$.

There are several possible changes to the regulator to minimize the bounds from Theorem 4 and Corollaries 1 and 2. Among them, one can select the discretization method for $K^{(0)}$ or apply a similarity transformation to its state-space representation. The discretization method is usually selected to shape the transient response of the regulator, impacting the matrices mentioned in Assumption 2, see also (Yepes et al., 2010; Şuşcă et al., 2023*a*,*b*). The next invariance proposition is justified by relations (10), (13) and (15).

Proposition 1. *The quantization* (11) *of a process G from* (6) *controlled by regulator K in* (10) *does not affect the pole-zero structure of the closed-loop system, i.e., its transient response.*

The proof is straightforward due to the structure of the resulting closed-loop system matrices. As such, based on Assumptions 1, 2 and Proposition 1, the quantization effects from K affect only the steady-state response. Thus, to minimize the quantization error bounds, we consider the set of state-space representations of $K^{(0)}$ of minimal order n_ξ, which implicitly maintains the input-output response. Starting from a fixed ideal regulator $K^{(0)}$ from (7), through a similarity transformation $S \in \text{GL}_{n_\xi}(\mathbb{R})$ applied to K from (10), we obtain a new regulator $K(S)$:

$$K = \left(\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right) \overset{S}{\sim} \left(\begin{array}{c|c} S^{-1}A_1S & S^{-1}B_1 \\ \hline C_1S & D_1 \end{array}\right) \overset{\text{def}}{=} K(S). \tag{34}$$

According to (33), the state error bound is influenced by the decision variables $S \in \text{GL}_{n_\xi}(\mathbb{R}) \overset{\text{def}}{=} \mathcal{S}$, to balance the controller using (34) leading to the state matrix $\mathbf{\Phi}_\Delta(S)$, and $P_\Delta \in \mathcal{P}_\Delta(S)$,

$$\mathcal{P}_\Delta(S) \overset{\text{def}}{=} \left\{P_\Delta \in \text{GL}_{n_\Xi}(\mathbb{C}) : \mathbf{\Phi}_\Delta(S) = P_\Delta J_{\mathbf{\Phi}_\Delta(S)} P_\Delta^{-1}\right\}, \tag{35}$$

to obtain various coordinate changes to the Jordan form representation from the bound formula (33). The dimensionality of the decision variable is $n_\xi^2 + n_\Xi^2$. This leads to the following optimization problem:

$$\min_{S\in\mathcal{S}} \min_{P_\Delta \in \mathcal{P}_\Delta(S)} \text{bound}\left(\overline{Q_\Delta^\Xi}(K(S), P_\Delta)\right). \tag{36}$$

We next proceed to show that, due to the structure of the quantization error, the minimization (36) can be performed using a single decision variable (instead of a two step approach), with a reduced number of parameters. Suppose $\mathbf{\Phi}_\Delta(S) = P_\Delta^{(0)} \cdot J^{(0)}_{\mathbf{\Phi}_\Delta(S)} \cdot \left(P_\Delta^{(0)}\right)^{-1}$ is a fixed arbitrary Jordan form realization and that it is composed of $s \le n_\Xi$ Jordan blocks with geometric multiplicities N_i, $i = \overline{1, s}$, $\sum N_i = n_\Xi$. The following lemma gives a way to generate all similarity matrices from (35) starting from a single fixed example.

Lemma 1. *For any $S \in \text{GL}_{n_\xi}(\mathbb{R})$, $\mathcal{P}_\Delta(S)$ can be spanned by any similarity matrix $P_\Delta^{(0)}$ right multiplied by the block diagonal matrices D_α from the set:*

$$\mathcal{D}_\alpha = \left\{\text{diag}\left(\alpha_1 I_{N_1}, \alpha_2 I_{N_2}, \ldots, \alpha_s I_{N_s}\right) \,\Big|\, \alpha_{N_i} \in \mathbb{C} \setminus \{0\},\ i = \overline{1, s}\right\}, \tag{37}$$

for a given permutation of the Jordan structure $J^{(0)}_{\mathbf{\Phi}_\Delta(S)}$. All Jordan block permutations of $\mathbf{\Phi}_\Delta(S)$ can be recovered by a further right multiplication with permutation matrices that preserve its block structure.

The proof is found in Appendix D. The technical advantage of the previous lemma is that, instead of using matrices P_Δ from (35), we can iterate through a single decision vector $(\alpha_1, \ldots, \alpha_s)^\top \in \mathbb{R}^s_{>0}$ of reduced size, which allows the use of off-the-shelf optimization techniques.

Remark 2. *In the case of a diagonal matrix $J^{(0)}_{\mathbf{\Phi}_\Delta(S)}$, the scaling set $\mathcal{D}_\alpha$ from* (37) *is characterized by n_Ξ degrees-of-freedom $\alpha_i \in \mathbb{C} \setminus \{0\}$ as opposed to $s \leq n_\Xi$ in the general statement of Lemma 1. In practice, to achieve smaller quantization error bounds, it may be desirable to design a closed-loop state matrix $\mathbf{\Phi}_\Delta(S)$ with eigenspaces of dimension* 1 *(not equivalent to forcing poles to have multiplicity* 1*), leading to $\mu(\mathbf{\Phi}_\Delta) = 1$ in* (25)*.*

Thus, according to Lemma 1, the inner minimization for $P_\Delta \in \mathcal{P}_\Delta(S)$ in (36) reduces the search from a n_Ξ^2 dimensional manifold to the Cartesian product of an $s \leq n_\Xi$ dimensional complex space and $N \leq s!$ permutation matrices Π.

A change S in $K(S)$ implies a change in the state matrix $\mathbf{\Phi}_\Delta(S)$ which, in turn, leads to a different similarity matrix P_Δ. This sequentiality of selecting a regulator $K(S)$ based upon which the matrix P_Δ will be further computed can be bypassed by two independent decision variables, as follows. Denote an extension of $S \in \mathcal{S}$ up to the dimension n_Ξ of the closed-loop system order as:

$$\tilde{S} = \operatorname{diag}\left(I_{n_x+n_\zeta}, S\right) \in \mathrm{GL}_{n_\Xi}(\mathbb{R}). \tag{38}$$

Lemma 2. *For $\Delta \in \mathbf{\Delta}$, given $S \in \mathcal{S}$ such that $K \overset{S}{\sim} K(S)$, the asymptotic gain bound* (19) *can be expressed as:*

$$\operatorname{bound}\left(\overline{Q^{\Xi}_{\Delta}}(K(S), P_\Delta)\right) = \left(\sum_{i=1}^{\mu(\mathbf{\Phi}_\Delta)} \frac{1}{(1-\rho(\mathbf{\Phi}_\Delta))^i}\right) \cdot \left[\left\|\tilde{S}^{-1} P_\Delta\right\| \cdot \left(\left\|\tilde{S} P_\Delta^{-1} \mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}\right\| \cdot \frac{\delta_{\xi_1}}{2} + \sum_{\vartheta \in \{e,\xi_2,u\}} \left\|P_\Delta^{-1} \mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\right\| \cdot \frac{\chi_\vartheta}{2}\right)\right]\Bigg|_K . \tag{39}$$

The proof is found in Appendix E.

Remark 3. *Compared to the influence of S on the state quantization error* (18)*, which roughly translates to $Q^{\Xi}_{\Delta}(k)\big|_{K(S)} = \tilde{S}^{-1} \cdot Q^{\Xi}_{\Delta}(k)\big|_K + \mathrm{Residual}(S, K)$, in the case of the output error* (24) *it translates to $Q^{y}_{\Delta}(k)\big|_{K(S)} = Q^{y}_{\Delta}(k)\big|_K + \mathbf{C}_\Delta \cdot \mathrm{Residual}(S, K)$. The existence of the residual term (that depends on $\mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}$) shows that the quantization in the controller state computations cannot be easily compensated in the process's outputs, although it can be for the states. Furthermore, Lemma 2 shows that a change S in $K(S)$ due to the structure of the quantization errors does not induce a new similarity matrix P_Δ for the Jordan form of $\mathbf{\Phi}_\Delta$. Thus, $S \in \mathcal{S}$ and $P_\Delta \in \mathcal{P}_\Delta$ can be varied independently.*

A challenge stems from the possibility of arbitrarily low or large norms, i.e., $\|S\| \to 0$ or $\|S\| \to \infty$, which makes it practically impossible to encode the regulator signals in the considered hardware of the microprocessor. Let $\mathcal{N}_\xi = \|(A_1, B_1, I, O)\|$ and $\mathcal{N}_u = \|(A_1, B_1, C_1, D_1)\|$ denote the $\mathcal{H}_\infty$-norms of the regulator K state and output signals, respectively, assumed to be finite. Based on Lemma 4 from (Şuşcă et al., 2022), a maximum admissible regulator state norm $\overline{\mathcal{N}_\xi}$ should be imposed. Given that $\mathcal{N}_u$ is invariant to $S \in \mathrm{GL}_{n_\xi}(\mathbb{R})$ (due to Proposition 1), an additional design specification arises:

$$S \in \mathcal{S} \, \text{subject to} \left\|\left(S^{-1}A_1 S, S^{-1}B_1, I, O\right)\right\| \leq \overline{\mathcal{N}_\xi}. \tag{40}$$

Combining (36) with Lemma 1, (39) from Lemma 2, and constraint (40), the minimization (36) becomes:

$$\begin{aligned} \min_{\substack{S \in \mathcal{S} \\ D_\alpha \in \mathcal{D}_\alpha (37)}} \quad & \operatorname{bound}\left(\overline{Q^{\Xi}_{\Delta}}\left(K(S), P^{(0)}_\Delta D_\alpha \Pi\right)\right) \\ \text{subject to} \quad & \left\|\left(S^{-1}A_1 S, S^{-1}B_1, I, O\right)\right\| \leq \overline{\mathcal{N}_\xi}, \end{aligned} \tag{41}$$

for any permutation matrix Π corresponding to the Jordan block structure $J_{\mathbf{\Phi}_\Delta}$. The functional (41) is differentiable, having a right-continuous Jacobian due to the ∞-norms, and a solution can be obtained by subgradient methods, see (Clarke, 1990). The above results can also be adapted for $\overline{Q^{y}_{\Delta}}$, $\overline{Q^{\Xi_i}_{\Delta}}$ and $\overline{Q^{y_i}_{\Delta}}$. The importance of Lemmas 1 and 2 in practice is that, on the one hand, they reduce the dimensionality of the optimization problem (36), and on the other hand, they allow the joint optimization of both variables S and P_Δ. Furthermore, Theorem 3 ensures that the bounds are well-defined based on the ∞-norm, while Theorem 5 ensures the robustness and overall range of the computed bounds based on only a few samples from the uncertainty set. Once a software routine to solve (41) is implemented, it can also be used in a reverse manner, i.e., to design the minimal hardware configuration χ_e, χ_ξ, χ_u which ensures a quantization error less than a prescribed tolerance. The resulting scaling leads to the expression of the balanced regulator (34) to be implemented. The control engineer can select to minimize the bound on state error (Theorem 4), output error (Corollary 1), or on a subset of any individual state or output signal (Corollary 2).

6. Case Study and Relevant Comparison

We proceed to illustrate the proposed method on a benchmark eight-order MIMO LTI mass, spring, dashpot system, adapted from (Mihaly et al., 2022). The steps that are taken are to: obtain the analytical expression of the process family (6) using MATLAB®; design a robust continuous-time controller (7) which satisfies Assumption 1; discretize the regulator; define the quantization steps, with minimum number of bits to not require saturation, based on constraint (40); balance the regulator matrices for a suboptimal output vector bound (Corollary 4) by solving (41); present a comparative analysis with the computed element-wise bounds using Corollary 2 and the method from (Haimovich et al., 2007); illustrate the negligible effects of uncertainties on the computed bounds.

The system is modelled through the Euler-Lagrange equation $\mathcal{M}\ddot{q}(t)+\mathcal{C}\dot{q}(t)+\mathcal{K}q(t) = \mathcal{T}u(t)$, where $q(t) = [q_i]_{i=\overline{1,4}}$ are the mechanical positions of the four masses. The matrices $\mathcal{M},\mathcal{C},\mathcal{K},\mathcal{T}$ are given by:

$$(\mathcal{M}|\mathcal{C}) = \left(\begin{array}{cccc|cccc} m_1 & 0 & 0 & 0 & c_1{+}c_2 & -c_2 & 0 & 0 \\ 0 & m_2 & 0 & 0 & -c_2 & c_2{+}c_3 & -c_3 & 0 \\ 0 & 0 & m_3 & 0 & 0 & -c_3 & c_3{+}c_4 & -c_4 \\ 0 & 0 & 0 & m_4 & 0 & 0 & -c_4 & c_4 \end{array}\right);\ (\mathcal{T}|\mathcal{K}) = \left(\begin{array}{cc|cccc} 0 & 0 & k_1{+}k_2 & -k_2 & 0 & 0 \\ 1 & 0 & -k_2 & k_2{+}k_3 & -k_3 & 0 \\ 0 & 0 & 0 & -k_3 & k_3{+}k_4 & -k_4 \\ 0 & 1 & 0 & 0 & -k_4 & k_4 \end{array}\right),$$

with parameters having nominal values $m_1 = m_2 = m_3 = m_4 = 1$ [kg], $k_1 = k_2 = k_3 = k_4 = 1$ [N/m], $c_1 = c_2 = c_3 = c_4 = 0.4$ [Ns/m] and tolerances $\mathrm{tol}_{m_1} = \pm 10[\%]$, $\mathrm{tol}_{m_2} = \mathrm{tol}_{m_3} = \mathrm{tol}_{m_4} = \pm 50[\%]$, $\mathrm{tol}_{c_1} = \mathrm{tol}_{c_2} = \mathrm{tol}_{c_3} = \mathrm{tol}_{c_4} = \mathrm{tol}_{k_1} = \mathrm{tol}_{k_2} = \mathrm{tol}_{k_3} = \mathrm{tol}_{k_4} = \pm 5[\%]$. This leads to static uncertainties in model (4), i.e., $\Delta \equiv D_\Delta$, normalized to $|D_\Delta| < 1$. Such models can be constructed using the `ultidyn` and `musynData` routines in MATLAB® (Balas et al., 2023). The state-space model G has state $x^\top = \begin{pmatrix} q^\top & \dot{q}^\top \end{pmatrix}$, input $u(t)$, and output $y^\top = \begin{pmatrix} x_2 & x_4 \end{pmatrix}$:

$$\left(\begin{array}{c} \dot{x}(t) \\ \hline y(t) \end{array}\right) = \left(\begin{array}{cc|c} O_{4\times 4} & I_{4\times 4} & O_{4\times 2} \\ -\mathcal{M}^{-1}\mathcal{K} & -\mathcal{M}^{-1}\mathcal{C} & \mathcal{M}^{-1}\mathcal{T} \\ \hline \mathcal{T}^\top & O_{2\times 4} & O_{2\times 2} \end{array}\right) \left(\begin{array}{c} x(t) \\ \hline u(t) \end{array}\right).$$

A closed-loop shaping procedure (Skogestad and Postlethwaite, 2005) was used to derive the continuous-time regulator $K^{(0)}(s)$. It was computed using the `musyn` routine from MATLAB®, version R2023b, with the weighting functions for the sensitivity $W_S(s) = \mathrm{diag}\left(\frac{0.4167s+0.15}{s+0.0015}, \frac{0.4167s+0.15}{s+0.0015}\right)$ and complementary sensitivity $W_T(s) = \mathrm{diag}\left(\frac{100(s+6)^2}{(s+92.95)^2}, \frac{100(s+6)^2}{(s+92.95)^2}\right)$, respectively. We further apply a balanced order-reduction algorithm with multiplicative error model on the resulting optimal controller. A zero-order hold discretization with a sampling period $T = 0.1$ [s] is applied to $G^{(0)}(s)$ and the form (5) is obtained using the `musynData` function. Through the Tustin discretization applied to $K^{(0)}(s)$, maintaining the previous sampling period, the following regulator is obtained[1]:

$$A_1 = \begin{pmatrix} 0.9999 & 0.0761 & -0.0495 & -0.0287 & -0.0764 & 0.0107 & 0.0166 & -0.0091 & 0.0640 & -0.5247 \\ -0.0004 & 0.6716 & 0.0176 & 0.1250 & 0.2698 & -0.1221 & -0.0427 & 0.1287 & 0.5797 & 0.8570 \\ -0.0015 & -0.0393 & -0.7132 & 0.1194 & -0.0710 & -0.1378 & -0.0416 & 0.1218 & -0.1231 & -0.4307 \\ 0.0001 & 0.3413 & -0.0293 & 0.2656 & 0.2556 & -0.0360 & -0.0525 & 0.0916 & -0.5154 & 0.0830 \\ 0.0010 & 0.1632 & -0.1870 & 0.4777 & -0.0220 & -0.2198 & 0.1721 & -0.0067 & 1.4104 & 1.3097 \\ 0.0006 & -0.0300 & -0.1184 & -0.1619 & -0.3565 & 0.3938 & -0.0215 & 0.3473 & 1.0975 & 1.3567 \\ 0.0002 & -0.0219 & -0.0279 & -0.1224 & 0.0007 & -0.1805 & 0.8873 & -0.1241 & 0.1434 & 1.3172 \\ -0.0004 & -0.0378 & 0.0716 & -0.0341 & 0.3010 & 0.2084 & 0.1611 & 0.9111 & -0.1920 & -0.3130 \\ 0.0001 & -0.0696 & -0.0201 & -0.2874 & 0.1754 & -0.2743 & -0.1554 & 0.1494 & -0.9181 & -0.3662 \\ -0.0000 & 0.0095 & 0.0064 & 0.0447 & -0.0147 & 0.0535 & 0.0233 & -0.0417 & 0.0952 & -0.7854 \end{pmatrix};$$

$$B_1^\top = \begin{pmatrix} 1.2515 & 3.2513 & -3.9534 & -2.3351 & 2.9414 & 4.5996 & -0.0712 & -0.9089 & 0.4562 & -0.0497 \\ -2.1297 & -2.2881 & -4.3578 & 3.9184 & 5.9039 & 2.8040 & -0.2377 & -0.2215 & -0.5109 & 0.1399 \end{pmatrix};$$

$$C_1 = \begin{pmatrix} 0.0097 & 0.0773 & -0.2420 & -3.2081 & 0.9643 & -2.6670 & -1.1576 & 1.5027 & -7.3888 & 2.8939 \\ -0.0020 & 0.4476 & -0.5423 & -3.1756 & 0.1145 & -2.8886 & -1.0948 & 1.5464 & -9.2361 & 6.9479 \end{pmatrix};$$

$$D_1 = \begin{pmatrix} 6.8053 & 6.3820 \\ 5.4449 & 12.1026 \end{pmatrix}. \tag{42}$$

[1] The source code and results with full precision can be found at the GitHub repository: https://github.com/mirceasusca/nahs-2025-quantization.

The scaling-invariant $\mathcal{H}_\infty$-norm of the controller is $\mathcal{N}_u = \|(A_1, B_1, C_1, D_1)\| = 213.25$, with dynamic range encoded on 8 bits. The range of the state signal is initially $\mathcal{N}_\xi = 21617.92$, which requires 7 extra bits to be correctly encoded. The quantizer resolutions (11) are $\chi_e = \frac{5}{2^{13}}, \chi_\xi = \frac{5}{2^{22}}, \chi_u = \frac{5}{2^{14}}$, calibrated to 5 [V] analog voltages, with truncation applied by the input-output devices and rounding applied to the internal computations.

For simplicity of implementation, we consider only positive diagonal scalings on S from (34), along with real-valued positive scalings D_α from (37) on P_Δ. This leads to $n_\Xi + n_\xi = 28$ $\mathbb{R}_{>0}$ decision variables. A good starting point for $\overline{\mathcal{N}_\xi}$ is 255, to maintain the same (already-required) dynamic range on the states. The minimization (41) has been performed using the `fmincon` routine on the nominal system, i.e., $\Delta \equiv 0$, with default hyperparameters, starting point $\mathtt{x}_0 = 1$, lower and upper bounds $\left[10^{-2}, 10^{2}\right]$ for S, and $\left[10^{-6}, 10^{8}\right]$ for D_α, respectively. A feasible suboptimal solution $x^\star$ which ensures the state bound is $\mathcal{N}_\xi^\star = 226.79$. The guaranteed state vector bound (39) decreases to 0.1961, compared to its starting value 4.9018, obtained by setting $S = I$, $D_\alpha = I$ and $P_\Delta^{(0)}$ returned by applying the `eig` routine to the nominal closed-loop state matrix $\mathbf{\Phi}_\Delta(I)|_{\Delta\equiv 0}$. This leads to the regulator $K(S^\star)$ by combining (34) with (42). $S^\star = \text{diag}(95.37, 98.43, 30.93, 98.43, 74.53, 96.07, 99.99, 70.52, 6.56, 13.42)$, with a corresponding $D_\alpha^\star = \text{diag}(4.74, 4.74,\ 0.38, 23.83, 13.77, 13.77, 38.42, 38.60, 1.70, 1.66, 1.2, 1.06, 10.09, 10.05, 19.33, 55.91, 35.90, 88.10)$ were obtained.

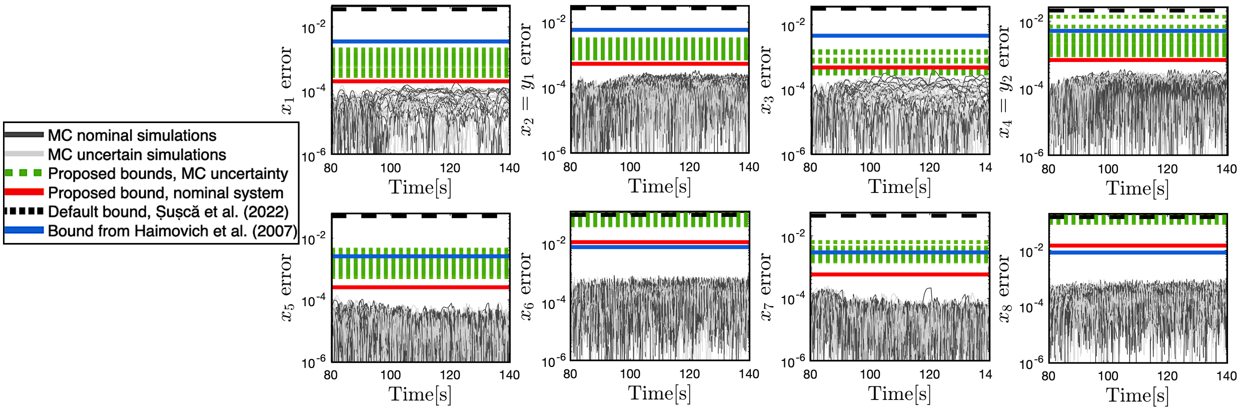

Figure 3: Element-wise state quantization error bounds bound$\left(Q_\Delta^{\Xi_i}\right)$ from (32a), for $i = \overline{1, 8}$, alongside Monte Carlo (MC) simulated closed-loop counterparts, as described in (*i*)-(*vi*) from Section 6. For this example, the experimental quantization errors in the nominal case (*v*) and with uncertainties (*vi*) are practically interchangeable, showing that the statement in Theorem 5, although providing a safety guarantee that the uncertainty does not arbitrarily break the quantization error, it is conservative. Furthermore, cases (*i*), (*iv*) show comparative results and the improvements obtained by using the proposed method.

Now, considering $K(S^\star)$ fixed, (32a) from Corollary 2 can be applied to compute the element-wise bounds for the process states, using only P_Δ as decision variable. This implicitly covers the output error bounds, as $y_1 = x_2$ and $y_2 = x_4$. Multiple experiments have been performed, with results gathered in Figure 3, where the y-axes are in logarithmic scale. For each state quantization error bound (32a), $i = \overline{1, 8}$, we consider: (*i*) an unoptimized *default* case with $P_\Delta = P_\Delta^{(0)}$, which is already an improvement compared to the full state vector bound from (Şuşcă et al., 2022) (dashed black lines); (*ii*) a set of 25 Monte Carlo suboptimal bounds for uncertain variations $\Delta \not\equiv 0$ of the process (green dotted lines); (*iii*) the best achieved bound for the nominal system $\Delta \equiv 0$ (red lines); (*iv*) the systematic ultimate bound computed with the method from (Haimovich et al., 2007) for $\Delta \equiv 0$ (blue lines). On top of that, we validate the computed bounds using two sets of 25 Monte Carlo closed-loop simulations for experimental quantization errors obtained using: (*v*) the nominal process (dark gray), and (*vi*) uncertain system configurations (light gray). The simulations have been performed for $t_{\text{sim}} = 150$ [s], which is beyond the settling time of the closed-loop system. The computed bounds of experiment (*iii*) are $10^{-4} \times (1.96, 5.44, 4.76, 6.99, 2.56, 115.19, 5.87, 147.32)$, while for (*iv*) they are $10^{-4} \times (34.83, 58.86, 45.16, 51.63, 25.69\ ,\ 77.86, 29.92, 83.72)$. It can be seen that the proposed method (*iii*) provides considerably tighter bounds on states x_1–x_5 and x_7, averaging to an improvement of an order of magnitude, while the bounds on x_6 and x_8 are approximately double in magnitude, compared to (*iv*).

The experiments in (*ii*) have been performed using a maximum of 3000 iterations in `fmincon`, compared to (*iii*), configured to run up to 40000 iterations, achieving tighter bounds. Figure 3 shows that the minimization of the bounds for states x_1–x_5 behave better than for x_6–x_8, as they tend, for various numbers of iterations and function tolerances, to reach the same value, and to less conservative bounds. Wider gaps can be seen between the green and red lines for x_6–x_8. An investigation of this discrepancy can be carried out as further research.

7. Conclusions and Future Works

This work analyses linear control systems perturbed by uniform quantization errors. It presents means to quantify the quantization effects on the system's states and outputs, and methods to minimize such effects through an adequate balancing of the regulator state-space realization. The guaranteed bounds are shown to scale proportionally with the norm of the uncertainty. The computed bounds can be easily adapted to allow different resolutions of each channel. Practical insights are given regarding the implementation and application of the method. In this work, we considered the regulator and hardware configuration as given, focusing on optimally leveraging the available resources to achieve the best possible performance. A natural extension involves the joint design of the least-cost hardware parameters χ_e, χ_ξ, χ_u with the objective of ensuring that the quantization error remains below a specified threshold.

A complementary research direction is to extend the proposed framework to the nonlinear system case. Assuming sufficiently smooth mappings, the recurrence relation that yields the state dynamics in (16) generalizes to an infinite composition of nonlinear functions. Such compositions could be analyzed by expanding them via Taylor series around the operating point, and applying the current method on the resulting linear approximation. It is also of significant interest to investigate the conditions under which diffeomorphisms, originally constructed in the continuous-time domain for system linearization, retain their effectiveness after numerical implementation.

References

Abdullah, A., Musolino, F. and Crovetti, P. (2023), 'Limit-cycle free, digitally-controlled boost converter based on ddpwm', *IEEE Access* **11**, 9403–9414.

Balas, G., Chiang, R., Packard, A. and Safonov, M. (2023), *Robust Control Toolbox, User's Guide*, The MathWorks, Inc.

Balas, G., Packard, A. and Seiler, P. (2009), *Uncertain Model Set Calculation from Frequency Domain Data*, Hof, P. and Scherer, C. and Heuberger, P., eds, Model-Based Control, Springer, Boston, MA.

Bolton, W. (2021), *Instrumentation and Control Systems*, 3rd edn, Newnes, Elsevier.

Brockett, R. and Liberzon, D. (2000), 'Quantized feedback stabilization of linear systems', *IEEE Trans. on Automatic Control* **45**(7), 1279–1289.

Campos, G., da Silva, J., Tarbouriech, S. and Pereira, C. (2016), 'Stability of discrete-time control systems with uniform and logarithmic quantizers', *IFAC-PapersOnLine* **49**(30), 132–137.

Chen, T. and Francis, T. (1995), *Optimal Sampled-Data Control Systems*, Springer, Communications and Control Engineering (CCE).

Cheng, J., Park, J., Zhao, Z., Cao, J. and Qi, W. (2019), 'Static output feedback control of switched systems with quantization: A nonhomogeneous sojourn probability approach', *Int J Robust Nonlinear Control* **29**(17), 1–14.

Clarke, F. (1990), *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics (SIAM).

Dajsuren, Y. and van der Brand, M., eds (2019), *Automotive Systems and Software Engineering: State of the Art and Future Trends*, Springer Nature Switzerland.

Delchamps, D. (1990), 'Stabilizing a linear system with quantized state feedback', *IEEE Trans. on Automatic Control* **35**(8), 916–924.

Ferdinando, M., Castillo-Toledo, B., Gennaro, S. and Pepe, P. (2022), 'Robust quantized sampled–data stabilization for a class of lipschitz nonlinear systems with time–varying uncertainties', *IEEE Control Systems Letters* **6**, 1256–1261.

Ferdinando, M., Gennaro, S., Bianchi, D. and Pepe, P. (2024), 'On robust quantized sampled–data tracking control of nonlinear systems', *IEEE Trans. on Automatic Control* .

Ferrante, F., Gouaisbaut, F. and Tarbouriech, S. (2015), 'Stabilization of continuous-time linear systems subject to input quantization', *Automatica* **58**, 167–172.

Ferrante, F. and Tarbouriech, S. (2024), 'Sampled-data feedback control design in the presence of quantized actuators', *Nonlinear Analysis: Hybrid Systems, 101530* **54**.

Franklin, G., Powell, J. and Workman, M. (2006), *Digital Control of Dynamic Systems*, 3rd edn, Ellis-Kagle Press.

Fu, M. (2024), 'A tutorial on quantized feedback control', *IEEE/CAA Journal of Automatica Sinica* **11**(1), 5–17.

Fu, M. and Xie, L. (2005), 'The sector bound approach to quantized feedback control', *IEEE Trans. on Automatic Control* **50**(11), 1698–1711.

Fu, M. and Xie, L. (2009), 'Finite-level quantized feedback control for linear systems', *IEEE Trans. on Automatic Control* **54**(5), 1165–1170.

Gray, R. and Neuhoff, D. (1998), 'Quantization', *IEEE Trans. on Information Theory* **44**(6), 2325–2383.

Haimovich, H., Kofman, E. and Seron, M. (2007), 'Systematic ultimate bound computation for sampled-data systems with quantization', *Automatica* **43**(6), 1117–1123.

Hayakawa, T., Ishii, H. and Tsumura, K. (2009), 'Adaptive quantized control for linear uncertain discrete-time systems', *Automatica* **45**(3), 692–700.

Hindi, H., Seong, C.-Y. and Boyd, S. (2002), Computing optimal uncertainty models from frequency domain data, *in* 'Proceedings of the 41st IEEE Conference on Decision and Control', Vol. 3, Las Vegas, NV, USA, pp. 2898–2905.

Ionescu, V., Oara, C. and Weiss, M. (1999), *Generalized Riccati Theory and Robust Control: A Popov Function Approach*, John Wiley & Sons Ltd, Baffins Lane, Chichester, West Sussex P019 IUD, England.

ISO (2018), Road vehicles – Functional safety, Standard ISO 26262:2018, International Organization for Standardization, Geneva, Switzerland.

Jiang, Z. and Wang, Y. (2001), 'Input-to-state stability for discrete-time nonlinear systems', *Automatica* **37**(6), 857–869.

Kameneva, T. and Nešić, D. (2010), 'Robustness of nonlinear control systems with quantized feedback', *Nonlinear Analysis: Hybrid Systems* **4**(2), 306–318.

Liberzon, D. (2003), 'Hybrid feedback stabilization of systems with quantized signals', *Automatica* **39**(9), 1543–1554.

Liberzon, D. (2006), 'Quantization, time delays, and nonlinear stabilization', *IEEE Trans. on Automatic Control* **51**(7), 1190–1195.

Liu, K., Fridman, E. and Johansson, K. (2015), 'Dynamic quantization of uncertain linear networked control systems', *Automatica* **59**, 248–255.

Llorente, R. (2020), *Practical Control of Electric Machines: Model-Based Design and Simulation*, Springer.

Markiş, I., Mihaly, V., Şuşcă, M. and Dobra, P. (2024), Convex chebyshev approximation for descriptor systems for frequency domain data fitting, *in* '2024 28th International Conference on System Theory, Control and Computing (ICSTCC)', Sinaia, Romania, pp. 368–373.

Mihaly, V., Şuşcă, M., Dulf, E. and Dobra, P. (2022), Approximating the fractional-order element for the robust control framework, *in* 'IEEE American Control Conference (ACC)', Atlanta, GA, USA, pp. 1151–1157.

Mironchenko, A. (2003), *Input-to-State Stability Theory and Applications*, Communications and Control Engineering, Springer.

Mullis, C. and Roberts, R. (1976), 'Synthesis of minimum roundoff noise fixed point digital filters', *IEEE Trans. on Circuits and Systems* **23**(9), 551–562.

Peng, H., Prodic, A., Alarcon, E. and Maksimovic, D. (2007), 'Modeling of quantization effects in digitally controlled dc–dc converters', *IEEE Trans. on Power Electronics* **22**(1), 208–215.

Petkov, P., Slavov, T. and Kralev, J. (2018), *Design of Embedded Robust Control Systems Using MATLAB/Simulink*, IET Control, Robotics and Sensors, The Institution of Engineering and Technology.

Proakis, J. and Manolakis, D. (2006), *Digital Signal Processing*, 4th edn, Pearson, Prentice Hall.

Skogestad, S. and Postlethwaite, I. (2005), *Multivariable Feedback Control: Analysis and design*, 2nd edn, John Wiley & Sons.

de Souza, C., Coutinho, D. and Fu, M. (2010), 'Stability analysis of finite-level quantized discrete-time linear control systems', *European Journal of Control* **3**, 258–271.

Spong, M., Hutchinson, S. and Vidyasagar, M. (2020), *Robot Modeling and Control*, 2nd edn, John Wiley & Sons Ltd.

Şuşcă, M., Mihaly, V. and Dobra, P. (2022), Fixed-point uniform quantization analysis for numerical controllers, *in* '61st IEEE Conference on Decision and Control (CDC)', Cancún, Mexico, pp. 3681–3686.

Şuşcă, M., Mihaly, V. and Dobra, P. (2023*a*), Maintaining robust stability and performance through sampling and quantization, *in* '2023 IEEE American Control Conf. (ACC)', San Diego, CA, USA, pp. 3852–3858.

Şuşcă, M., Mihaly, V. and Dobra, P. (2023*b*), 'Sampling rate selection for multi-loop cascade control systems in an optimal manner', *IET Control Theory & Applications* **17**(8), 1073–1087.

Şuşcă, M., Mihaly, V., Sim, S. and Dobra, P. (2024), Design of linear control laws for minimum uniform quantization tracking error, *in* 'IEEE European Control Conference (ECC)', Stockholm, Sweden, pp. 3617–3622.

UNECE (2020*a*), UN Regulation No. 155: Uniform provisions concerning the approval of vehicles with regards to cyber security and a cyber security management system, Regulation, United Nations Economic Commission for Europe, Geneva, Switzerland.

UNECE (2020*b*), UN Regulation No. 156: Uniform provisions concerning the approval of vehicles with regards to software update management, Regulation, United Nations Economic Commission for Europe, Geneva, Switzerland.

Wang, J. (2021), 'Quantized feedback control based on spherical polar coordinate quantizer', *IEEE Trans. on Automatic Control* **66**(12), 6077–6084.

Wang, P., He, X., Chen, Q., Cheng, A., Liu, Q. and Cheng, J. (2021), 'Unsupervised network quantization via fixed-point factorization', *IEEE Trans. on Neural Networks and Learning Systems* **32**(6), 2706–2720.

Wang, W., Zhou, J., Wen, C. and Long, J. (2022), 'Adaptive backstepping control of uncertain nonlinear systems with input and state quantization', *IEEE Trans. on Automatic Control* **67**(12), 6754–6761.

Widrow, B. and Kollár, I. (2008), *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*, Cambridge University Press, Cambridge, UK.

Xia, M., Gahinet, P., Abroug, N., Buhr, C. and Laroche, E. (2020), 'Sector bounds in stability analysis and control design', *International Journal of Robust and Nonlinear Control* **30**, 7857 – 7882.

Xu, T., Duan, Z., Sun, Z. and Chen, G. (2022), 'A unified control method for consensus with various quantizers', *Automatica, 110090* **136**.

Xu, X., Ozay, N. and Gupta, V. (2020), 'Passivity-based analysis of sampled and quantized control implementations', *Automatica, 109064* **119**.

Yepes, A., Freijedo, F., Doval-Gandoy, J., Lopez, O., Malvar, J. and Fernandez-Comesana, P. (2010), 'Effects of discretization methods on the performance of resonant controllers', *IEEE Trans. on Power Electronics* **25**(7), 1692–1712.

Yoo, S. and Park, B. (2021), 'Quantized-states-based adaptive control against unknown slippage effects of uncertain mobile robots with input and state quantization', *Nonlinear Analysis: Hybrid Systems, 101077* **42**.

Zhou, B., Duan, G.-R. and Lam, J. (2010), 'On the absolute stability approach to quantized feedback control', *Automatica* **46**(2), 337–346.

Appendix A. Proof of Theorem 4

Denote:

$$\Sigma_k\left(\mathbf{\Phi}_\Delta, \mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}, \eta_\vartheta\right) = \sum_{i=0}^{k-1} \mathbf{\Phi}^i_\Delta \mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\eta_\vartheta(k-1-i),\ \vartheta \in \Theta.$$

An upper bound for $\overline{Q_{\Delta}^{\Xi}}$ is:

$$\overline{Q_{\Delta}^{\Xi}} \leq \sup_{k\geq 0} \sum_{\vartheta\in\Theta} \left\| \Sigma_k \left(\mathbf{\Phi}_{\Delta}, \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta}, \eta_{\vartheta} \right) \right\|. \tag{A.1}$$

Each partial sum from (A.1) can be rewritten as:

$$\left\| \Sigma_k \left(\mathbf{\Phi}_{\Delta}, \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta}, \eta_{\vartheta} \right) \right\| \leq \sum_{i=0}^{k-1} \left\| \mathbf{\Phi}_{\Delta}^{i} \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta} \eta_{\vartheta}(k-1-i) \right\| \leq \|P_{\Delta}\| \cdot \sum_{i=0}^{k-1} \left\| J_{\mathbf{\Phi}_{\Delta}}^{i} \right\| \cdot \left\| P_{\Delta}^{-1} \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta} \right\| \cdot \frac{\chi_{\vartheta}}{2}.$$

For each Jordan cell $J_{N_i}(\lambda_i)$ of dimension N_i, $i = \overline{1, s}$, corresponding to eigenvalue $\lambda_i \in \Lambda\left(\mathbf{\Phi}_{\Delta}\right)$ we have:

$$J_{N_i}^{m}(\lambda_i) = \begin{bmatrix} \lambda_i^m & \binom{m}{1}\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \dots \\ & \lambda_i^m & \binom{m}{1}\lambda^{m-1} & \dots \\ & & \ddots & \ddots \\ & & & \lambda_i^m \end{bmatrix},$$

which implies

$$\|J_{N_i}^{m}(\lambda_i)\| \leq \sum_{j=0}^{N_i-1} |\lambda_i|^{m-j} \binom{m}{j} \leq \sum_{j=0}^{N_i-1} \rho\left(\mathbf{\Phi}_{\Delta}\right)^{m-j} \binom{m}{j},$$

where $\binom{m}{j} = 0$, if $j > m$. As $k \to \infty$, we have:

$$\sum_{m=0}^{k-1} \|J_{N_i}^{m}(\lambda_i)\| \leq \sum_{m=0}^{k-1} \sum_{j=0}^{N_i-1} \rho\left(\mathbf{\Phi}_{\Delta}\right)^{m-j} \binom{m}{j} = \sum_{j=0}^{N_i-1} \sum_{m=0}^{k-1} \rho\left(\mathbf{\Phi}_{\Delta}\right)^{m-j} \binom{m}{j} \leq \sum_{j=1}^{N_i} \frac{1}{\left(1-\rho\left(\mathbf{\Phi}_{\mathbf{\Delta}}\right)\right)^{j}}.$$

Therefore, considering $\mu\left(\mathbf{\Phi}_{\Delta}\right) = \max\{N_1, N_2, \dots, N_s\}$, we have

$$\sum_{i=0}^{k-1} \left\| J_{\mathbf{\Phi}_{\Delta}}^{i} \right\| \leq \sum_{i=1}^{\mu(\mathbf{\Phi}_{\Delta})} \frac{1}{\left(1-\rho\left(\mathbf{\Phi}_{\mathbf{\Delta}}\right)\right)^{i}},$$

which leads to the upper bound:

$$\left\| \Sigma_k \left(\mathbf{\Phi}_{\Delta}, \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta}, \eta_{\vartheta} \right) \right\| \leq \left(\sum_{i=1}^{\mu(\mathbf{\Phi}_{\Delta})} \frac{1}{\left(1-\rho\left(\mathbf{\Phi}_{\Delta}\right)\right)^{i}} \right) \cdot \|P_{\Delta}\| \cdot \left\| P_{\Delta}^{-1} \mathbf{\Gamma}_{\Delta,\eta}^{\vartheta} \right\| \cdot \frac{\chi_{\vartheta}}{2}. \tag{A.2}$$

Combining (A.1) with (A.2), (19) follows. ■

Appendix B. Proof of Theorem 5

We begin with an auxiliary result needed to prove Theorem 5. The following lemma establishes a connection between the Markov parameters of a stable LTI system and its $\mathcal{H}_{\infty}$-norm.

Lemma 3. *If a discrete-time system (A, B, C, D) has the $\mathcal{H}_{\infty}$-norm ψ, then $|D| \leq \psi$ and for each $k > 0$:*

$$\sup_{z\in\mathbb{C},|z|=1} \left| D + CBz^{-1} + \ldots + CA^{k-1}Bz^{-k} \right| \leq \psi. \tag{B.1}$$

Proof. From the $\mathcal{H}_{\infty}$-norm definition, we have:

$$\sup_{z\in\mathbb{C},|z|=1} \left| D + C\left(zI - A\right)^{-1} B \right| = \psi. \tag{B.2}$$

Let $h_0 = D$ and $h_n = CA^{n-1}B$, $n \geq 1$, be the impulse response of the discrete-time system, i.e., its Markov parameters. Because the system is Schur stable, the spectral radius $\rho(A) < 1$, so A^n converges exponentially to the null matrix. Moreover:

$$|h_n| = |CA^{n-1}B| \leq |C||A|^{n-1}|B|,$$

so the sequence of norms $|h_n|$ decays exponentially. Therefore, the series $\sum\limits_{n=0}^{\infty} |h_n|$ is convergent. Additionally, according to (B.2), we have that $\left|\sum\limits_{n=0}^{\infty} h_n z^n\right| \leq \psi$, for each complex number z with $|z| = 1$. Therefore, the power series $\sum\limits_{n=0}^{\infty} h_n z^n$ is uniformly convergent on the unit circle.

We first prove that $|h_0| = |D| \leq \psi$. For $z \equiv \varepsilon_\ell \sqrt[n]{z}$, where $\varepsilon_\ell = e^{j\frac{2\pi\ell}{n}}$, $\ell \in \{0, 1, \ldots, n-1\}$ are the n^{th} roots of unity and z is an arbitrary complex number on the unit circle, we have:

$$\left|D + CB\varepsilon_\ell^{-1}\sqrt[n]{z}^{-1} + CAB\varepsilon_\ell^{-2}\sqrt[n]{z}^{-2} + \ldots\right| \leq \psi,$$

for $\ell \in \overline{1, n}$ and $|z| = 1$. Summing up the above n inequalities, and using the triangle inequality along with the following fact:

$$\sum_{\ell=0}^{n-1} \varepsilon_\ell^i = \sum_{\ell=0}^{n-1} e^{i \cdot j\frac{2\pi\ell}{n}} = \begin{cases} 0, \text{ if } \ i \not\equiv 0 \ (\text{mod } n); \\ n, \text{ if } \ i \equiv 0 \ (\text{mod } n), \end{cases}$$

it leads to:

$$\left|D + CA^{n-1}Bz^{-1} + CA^{2n-1}Bz^{-2} + \ldots\right| \leq \psi. \tag{B.3}$$

Because the spectral radius $\rho(A) < 1$, we have $\lim\limits_{n\to\infty} A^n = O$, so relation (B.3) implies $|D| \leq \psi$.

For $k \geq 1$ we proceed by induction for an adequate system which has the partial sum of the first k Markov parameters in its feedforward matrix D. We consider the response of the (A, B, C, D) system to the input:

$$u(n) = \delta(n) + v\delta(n+1) + \cdots + v^k\delta(n+k),$$

where $\delta(n)$ is the discrete impulse signal and v is a complex number with $|v| = 1$. Then, we will have the same property for a new system with Markov parameters $\tilde{h}_n = h_n + vh_{n+1} + \cdots + v^k h_{n+k}$, $n \geq 0$, and, according to the case $k = 0$, using the uniform convergence of $\sum\limits_{n=0}^{\infty} h_n z^n$, we have that each partial term satisfies:

$$\sup_{|v|=1} \left|h_0 + h_1 v + \cdots + h_k v^k\right| \leq \psi,$$

which covers all the remaining cases of statement (B.1). ■

Next, we proceed to analyze $Q_\Delta^\Xi(k)$ from (18) in terms of the nominal ($\Delta \equiv 0$) system matrices perturbed by residuals due to $\Delta \not\equiv 0$, $\|\Delta\| < 1$. Define the shift operator $\mathbb{S} : \ell^{n_\zeta} \to \ell^{n_\zeta}$, $\mathbb{S}x(n) = x(n-1)$. Then, for $k \in \mathbb{N}$ and initial condition $\zeta(0) = 0$, the output of the discrete-time uncertainty model (4) is:

$$d(k) = D_\Delta v(k) + \sum_{i=0}^{k-1} C_\Delta A_\Delta^{k-1-i} B_\Delta v(i) = \left(D_\Delta + \sum_{i=0}^{k-1} \mathbb{S}^{i+1} C_\Delta A_\Delta^{k-1-i} B_\Delta\right) v(k) \overset{\text{def}}{=} \Delta(k) \cdot v(k).$$

Then, according to Lemma 3, $|\Delta(k)| \leq \|\Delta\|$, for all $k \in \mathbb{N}$. This means that system G from (6), with only its original states $x \in \mathbb{R}^{n_x}$, can be seen as a linear time-varying system with matrices:

$$\begin{pmatrix} A_{2,\Delta}(k) & B_{2,\Delta}(k) \\ C_{2,\Delta}(k) & D_{2,\Delta}(k) \end{pmatrix} = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} + \begin{pmatrix} B_d \\ D_{yd} \end{pmatrix} (I - \Delta(k) D_{vd})^{-1} \ \Delta(k) \begin{pmatrix} C_v & D_{vu} \end{pmatrix} = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} + \text{Res}(\Delta(k)),$$

which perturbs the nominal matrices with a term that scales with the uncertainty norm, i.e., $|\mathrm{Res}(\Delta(k))| = \mathcal{O}(\|\Delta\|)$. Furthermore, it follows that matrices from the steady-state quantization error (18) are affected in the same manner: $\mathbf{\Phi}_\Delta = \mathbf{\Phi}_\Delta|_{\Delta\equiv 0} + \mathrm{Res}(\Delta(k))$ and $\mathbf{\Gamma}_{\Delta,\eta} = \mathbf{\Gamma}_{\Delta,\eta}\big|_{\Delta\equiv 0} + \mathrm{Res}(\Delta(k))$. Therefore:

$$\mathcal{Q}^{\Xi}_{\Delta}(k) = \sum_{i=0}^{k-1} \left[\mathbf{\Phi}_\Delta|_{\Delta\equiv 0} + \mathrm{Res}\,(\Delta(k))\right]^i \cdot \left[\mathbf{\Gamma}_{\Delta,\eta}\big|_{\Delta\equiv 0} + \mathrm{Res}\,(\Delta(k))\right] \eta(k-1-i) = \mathcal{Q}^{\Xi}_{\Delta}(k)\big|_{\Delta\equiv 0} + \sum_{i=0}^{k-1} \left[\sum \mathrm{Res}\,(\Delta(k))\right] \eta(k-1-i). \tag{B.4}$$

All resulting uncertain control systems are stable, due to Assumption 1, leading to convergent residual series in (B.4). By applying the supremum with respect to $k \in \mathbb{N}$ and the ∞-norm, (20) follows. ■

Appendix C. Proof of Corollary 2

In the case of element-wise state bounds (32a), the individual signal errors δ^{Ξ_i}, $i = \overline{1, n_\Xi}$ from (30) are decoupled, see (29), which implies that the bounds on $\left|\delta^{\Xi_i}(k)\right|$ for $k \to \infty$ coincide with bound$\left(\overline{\mathcal{Q}^{\Xi_i}_{\Delta}}\right)$, $i = \overline{1, n_\Xi}$.

On the other hand, for the element-wise outputs (32b), the individual signal errors δ^{y_i}, $i = \overline{1, n_y}$ are linked, based on (27) and (28), as:

$$\underbrace{\begin{pmatrix} E^{y_1}_{\Delta} \\ \vdots \\ E^{y_{n_y}}_{\Delta} \end{pmatrix}}_{E_\Delta} \begin{pmatrix} \delta^{y_1}(k) \\ \vdots \\ \delta^{y_{n_y}}(k) \end{pmatrix} = \begin{pmatrix} \mathcal{Q}^{y_1}_{\Delta}(k) \\ \vdots \\ \mathcal{Q}^{y_{n_y}}_{\Delta}(k) \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \delta^{y_1}(k) \\ \vdots \\ \delta^{y_{n_y}}(k) \end{pmatrix} = E^{-1}_{\Delta} \begin{pmatrix} \mathcal{Q}^{y_1}_{\Delta}(k) \\ \vdots \\ \mathcal{Q}^{y_{n_y}}_{\Delta}(k) \end{pmatrix}.$$

Therefore, each signal δ^{y_i} is bounded by:

$$\sup_{k\geq 0} |\delta^{y_i}(k)| = \sup_{k\geq 0} \left|\sum_{j=1}^{n_y} \tilde{e}_{ij} \cdot \mathcal{Q}^{y_j}_{\Delta}(k)\right| \leq \sum_{j=1}^{n_y} \left|\tilde{e}_{ij}\right| \cdot \sup_{k\geq 0} \left|\mathcal{Q}^{y_j}_{\Delta}(k)\right| \leq \sum_{j=1}^{n_y} \left(\left|\tilde{e}_{ij}\right| \cdot \mathrm{bound}\left(\overline{\mathcal{Q}^{y_i}_{\Delta}}\right)\right),$$

which concludes the proof. ■

Appendix D. Proof of Lemma 1

The similarity matrix $P^{(0)}_{\Delta}$ has the eigenvector structure:

$$P^{(0)}_{\Delta} = \left(\begin{array}{ccc|c|ccc} p^{(1)}_1 & \ldots & p^{(1)}_{N_1} & \ldots & p^{(s)}_1 & \ldots & p^{(s)}_{N_s} \end{array}\right) \in \mathbb{C}^{n_\Xi \times n_\Xi},$$

according to the Jordan block structure of $J^{(0)}_{\mathbf{\Phi}_\Delta(S)}$, as each eigenspace V_{λ_i} of $P^{(0)}_{\Delta}$ has the dimension $\dim V_{\lambda_i} = N_i$, leading to:

$$V_{\lambda_i} = \mathrm{Span}\left\{p^{(i)}_1, \ldots, p^{(i)}_{N_i}\right\},\ i = \overline{1, s},\ \lambda_i \in \Lambda\left(\mathbf{\Phi}_\Delta(S)\right).$$

To preserve the change of coordinates from $\mathbf{\Phi}_\Delta(S)$ to $J^{(0)}_{\mathbf{\Phi}_\Delta(S)}$, the basis of each subspace V_{λ_i} can at most be scaled by a non-zero value $\alpha_i \in \mathbb{C} \setminus \{0\}$. It follows that all Jordan form representations of $\mathbf{\Phi}_\Delta(S)$ can be written as:

$$\mathbf{\Phi}_\Delta(S) = \left(P^{(0)}_{\Delta} D_\alpha \Pi\right) \cdot \left(\Pi^{-1} J^{(0)}_{\mathbf{\Phi}_\Delta(S)} \Pi\right) \cdot \left(\Pi^{-1} D^{-1}_{\alpha} (P^{(0)}_{\Delta})^{-1}\right) = P_\Delta \cdot J_{\mathbf{\Phi}_\Delta}(S) \cdot P^{-1}_{\Delta},$$

where $D_\alpha = \mathrm{diag}(\alpha_1 I_{N_1}, \ldots, \alpha_s I_{N_s}) \in \mathcal{D}_\alpha$ as in (37) and Π are permutation matrices of order s, with units extended to size N_i, $i = \overline{1, s}$, corresponding to each Jordan block of $J^{(0)}_{\mathbf{\Phi}_\Delta(S)}$. ■

Appendix E. Proof of Lemma 2

An analysis of the bound (19), considering a change of coordinates to from K to $K(S)$ implies two different occurrences of matrix $\tilde{S}$ from (38). First, for the state matrix, there is the similarity transformation:

$$\mathbf{\Phi}_\Delta(S) = \begin{pmatrix} A_{2,\Delta} & B_{2,\Delta}C_1S \\ O & S^{-1}A_1S \end{pmatrix} - \begin{pmatrix} B_{2,\Delta}D_1 \\ S^{-1}B_1 \end{pmatrix} (I + D_{L,\Delta,e})^{-1} \begin{pmatrix} C_{2,\Delta} & D_{2,\Delta}C_1S \end{pmatrix} = \tilde{S}^{-1}\mathbf{\Phi}_\Delta(I)\tilde{S} \overset{\tilde{S}}{\sim} \mathbf{\Phi}_\Delta(I).$$

Furthermore, matrices $\mathbf{C}_\Delta$ and $\mathbf{\Gamma}_{\Delta,\eta}$ from (12) are transformed to:

$$\mathbf{C}_\Delta|_{K(S)} = \mathbf{C}_\Delta|_K \cdot \tilde{S}; \quad \mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}\Big|_{K(S)} = \mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}\Big|_K; \quad \mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\Big|_{K(S)} = \tilde{S}^{-1} \cdot \mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\Big|_K, \quad \vartheta \in \{e, \xi_2, u\}.$$

As the state quantization error (16) involves products of the form $\mathbf{\Phi}^i_\Delta \cdot \mathbf{\Gamma}_\Delta$, it follows that, for the scaled regulator $K(S)$, formula (18) reduces to a left-multiplication by $\tilde{S}^{-1}$. An exception is the invariant term $\mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}$, as it does not depend on the regulator matrices by $\tilde{S}^{-1}$ to cancel the effect of $\tilde{S}$ from $\mathbf{\Phi}^i_\Delta$:

$$Q^{\Xi}_\Delta(k) = \sum_{\vartheta\in\{e,\xi_2,u\}} \sum_{i=0}^{k-1} \tilde{S}^{-1}\mathbf{\Phi}^i_\Delta \cancel{\tilde{S}}\cancel{\tilde{S}^{-1}}\mathbf{\Gamma}^{\vartheta}_{\Delta,\eta}\eta_\vartheta(k-1-i) + \sum_{i=0}^{k-1} \tilde{S}^{-1}\mathbf{\Phi}^i_\Delta\tilde{S} \cdot \mathbf{\Gamma}^{\xi_1}_{\Delta,\eta}\eta_{\xi_1}(k-1-i). \tag{E.1}$$

By applying (19) to the set of matrices (E.1), the bound (39) for the scaled regulator results. ■