2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Design of Linear Control Laws for Minimum Uniform Quantization Tracking Error

Mircea Şuşcă, *Member, IEEE*, Vlad Mihaly, *Member, IEEE*,
Simona Daiana Sim and Petru Dobra, *Member, IEEE*

***Abstract*— The uniform quantization effects present in the implementation of numeric regulators introduce undesired tracking errors, even though their continuous-time counterparts can ensure ideal steady-state response. Without perturbing the transient response, the state realization of the regulator can be scaled to reduce the influence of the steady-state artifacts. A main theoretical contribution is thus proposed, based on two complementary aspects. The starting point is given by an analytical bound of the quantization error. On one hand, this guaranteeable bound is minimized by the existence of an optimally scaled similarity matrix for the Jordan form of the closed-loop state matrix. On the other hand, a balancing scheme for the numeric regulator further reduces the quantization effects for a predefined hardware configuration. Mathematical guarantees to enforce said properties are then presented, developing sufficient conditions. Finally, the proposed method is illustrated on a case study which demonstrates the non-conservative nature of the optimized bound in comparison to the default value from the characterization theorem.**

I. Introduction

Numeric implementation of regulator models involves several key steps which require an analysis regarding sampling, discretization, quantization of system coefficients and quantization of involved signals. The previously-mentioned phenomena can affect both the transient and steady-state responses of the desired continuous-time regulator dynamics.

The effects of a fixed sampling rate selection on the transient response have been studied in a unified manner for single and multi-loop linear control systems in [1], while a time-variable sampling strategy for networked PID-based control systems is studied in [2]. The discretization method selection effects on the transient performance of resonant controllers are described in [3], while for robust control systems, a joint optimization problem for the regulator sampling rate and coefficients quantization step in order to maintain robust stability and performance is proposed in [4].

There are several types of quantizer circuits, such as fixed-point, floating-point, logarithmic, delta-sigma, to name a few. Fixed-point quantization is widely used as it is fast, can be implemented with low energy consumption, or it can be harvested to increase the parallel processing capabilities of high-performance graphics processing units. Apart from embedded systems, a newly-established use-case for fixed-point quantization, static or with dynamic scaling, is in deep neural network training and deployment. Using the framework from [5], the main advantage is that the desired network performance is maintained with minimal degradation, without supervised retraining on the labeled data. The fundamental difference encountered in control is given by the feedback connection which, due to quantized subsystems, leads to highly-nonlinear behaviour, such as limit cycles [6].

Quantization effects are studied in digital signal processing and control applications such as in the monograph [7], providing a probabilistic approach, and for embedded robust control system design in [8], where the authors present the phenomena which occur in practice, but exemplified through ad hoc studies for individual use cases. In [9] the stability and stabilization problems for input and output quantized feedback discrete systems have been addressed, while the authors of [10] propose a solution for stabilizing discrete-time linear systems considering a logarithmic quantizer for both input and output channels. Steady-state deviations with tight bounds in the context of DC-DC converters are modelled in [11], followed by means to combat limit cycles in the particular case of boost converters in [12], and an analytical bound provided for linear control systems in [13].

Recent applications involving quantization analysis can be found for the practical implementation of control barrier functions in [14] or robust model predictive control to maintain system stability during fixed-point encoding in [15]. By modelling output quantization and saturation, the paper [16] uses model reference control to provide convergence guarantees on the output tracking error without relying on polynomial coprimeness or initial condition assumptions. Similarly, using finite-and-quantized output feedback, the authors of [17] provide a pole-placement-based control law which guarantees that the tracking error converges to an arbitrarily-small residual set. The distributed consensus problem involving both uniform and logarithmic quantizers has been solved in a unified manner for the control of networked general linear systems in [18].

This work is an extension of [13] by moving forward to a design point of view. The purpose is to compute the discrete regulator balancing least affected by quantization errors, starting from a satisfactory design with a given input-output behaviour. As such, without degrading the desired transient response, a generalized tracking error bound is computed depending on the regulator and process state-space models, along with input and output converter specifications, and internal computation hardware capabilities, assuming fixed-point uniform numeric encoding leading to static (memo-

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.

All authors are part of the Department of Automation, Technical University of Cluj-Napoca, Romania (emails: {mircea.susca, vlad.mihaly, petru.dobra}@aut.utcluj.ro, sim.simonadaiana@gmail.com).

ryless) quantizers. The contributions of the paper are to: (i) propose three design problems to compute the least conservative guaranteeable quantization error, (ii) provide useful tools to characterize the resulting design problems, through Lemmas 1, 2, and (iii) illustrate the tightness of the optimized bound value on a numeric example.

Beyond the present introduction, Section II provides a short background on quantized closed-loop systems, Section III proposes a set of design problems to minimize the steady-state quantization effects and their numeric solutions, while Section IV illustrates the design problem on a numeric case study, closing with some concluding remarks.

Notations: Denote by $\rho(A)$ the spectral radius of a square matrix $A \in \mathbb{C}^{n\times n}$. A transformation matrix P denotes the similarity matrix used to bring A to its Jordan canonical form, i.e. $A = P{\cdot}J_A{\cdot}P^{-1}$. $\|\cdot\|$ will implicitly denote the ∞-norm. 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 square matrix with diagonal entries from a vector $x \in \mathbb{R}^n$ is written $D_x = \text{diag}\,(x_1, \ldots, x_n)$. The symbol $\overset{T}{\sim}$ denotes matrix and system similarity transformations through a matrix T. The general linear group of degree n is written $\text{GL}_n(\mathbb{C})$.

II. Brief Theoretical Background

The one-degree-of-freedom (1DOF) linear time-invariant (LTI) numeric control structure with standard signal notations is found in Figure 1. Assume that the continuous regulator $K(s)$ ensures asymptotic stability for the process $G(s)$. Based on the structure of the numeric regulator $K(z)$ as in Figure 2 and that the digital-to-analog converter of the regulator implies the zero-order hold discretization method for the plant model, denote state-space representations as:

$$\left(K^{(0)}(z)\right): \begin{cases} x_c[k{+}1] & = A_1 x_c[k] + B_1 e[k]; \\ u[k] & = C_1 x_c[k] + D_1 e[k], \end{cases} \tag{1}$$

$$(G(z)): \begin{cases} x[k{+}1] & = A_2 x[k] + B_2 u[k]; \\ y[k] & = C_2 x[k] + D_2 u[k], \end{cases} \tag{2}$$

with dimensions $e, y \in \mathbb{R}^{n_y}$, $u \in \mathbb{R}^{n_u}$, $x \in \mathbb{R}^n$, $x_c \in \mathbb{R}^{n_c}$.

To account for the quantization effects of the regulator hardware, consider the definitions of the two classical fixed-point encoding quantizers $Q \in \{Q^t_\delta, Q^r_\delta\}$, namely midtread (rounding): $Q^t_\delta(x) = \delta \left\lfloor \frac{x}{\delta} + \frac{1}{2} \right\rfloor$, along with midriser (truncation): $Q^r_\delta(x) = \delta \left(\left\lfloor \frac{x}{\delta} \right\rfloor + \frac{1}{2} \right)$, for a quantization step $\delta > 0$ and arbitrary $x \in \mathbb{R}$. Both quantizer functions can be written in a unified manner as $Q_\delta(x) = x + \varphi(x,\delta)\cdot\delta$, $\varphi(x,\delta) \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, where $\varphi^t(x,\delta) = \frac{1}{2} - \left\{\frac{x}{\delta} + \frac{1}{2}\right\} \in \left[-\frac{1}{2}, \frac{1}{2}\right)$, for midtread, and $\varphi^r(x,\delta) = \frac{1}{2} - \left\{\frac{x}{\delta}\right\} \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ for midriser.

As such, irrespective of using rounding or truncation for the input, state, output, the quantized regulator $K(z)$, adapted from the ideal case of (1), can be rewritten maintaining its

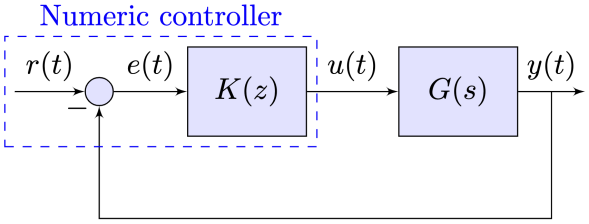

Fig. 1. One-degree-of-freedom control structure having a continuous-time plant and numeric regulator.

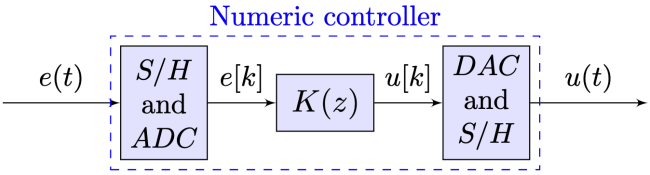

Fig. 2. Numeric regulator with interfacing devices: sample and hold with analog-to-digital converter (ADC), along with the digital-to-analog converter (DAC) followed by a sample and hold circuit.

LTI form as:

$$(K(z)): \begin{cases} \overline{e}[k] & = Q_{\delta_e}(e[k]) = e[k]{+}\eta_e[k]; \\ x_c[k{+}1] & = A_1 x_c[k]{+}B_1\,(e[k]{+}\eta_e[k]) {+}\eta_{x,1}[k]; \\ \overline{u}[k] & = C_1 x_c[k]{+}D_1\,(e[k]{+}\eta_e[k]) {+}\eta_{x,2}[k]; \\ u[k] & = Q_{\delta_u}(\overline{u}[k]) = \overline{u}[k]{+}\eta_u[k], \end{cases} \tag{3}$$

with additional disturbance inputs of corresponding sizes: η_e for the ADC with step δ_e; η_u for the DAC with step δ_u; $\eta_{x,1}$ and $\eta_{x,2}$ for the state and output computation errors with steps δ_x. By definition, $\|\eta_e[k]\| \leq \frac{\delta_e}{2}$, $\|\eta_u[k]\| \leq \frac{\delta_u}{2}$, $\|\eta_{x,1}[k], \eta_{x,2}[k]\| \leq \frac{\delta_x}{2}$. Consider the ideal discretized plant (2). Then, the open-loop system of order $n_\ell = n_c + n$, with input e, output y, extended state vector $x_s = \begin{pmatrix} x_c^\top & x^\top \end{pmatrix}^\top$ becomes:

$$(L(z)): \left(\begin{array}{cc|cccc:c} A_1 & O & B_1 & I & O & O & B_1 \\ B_2C_1 & A_2 & B_2D_1 & O & B_2 & B_2 & B_2D_1 \\ \hline D_2C_1 & C_2 & D_2D_1 & O & D_2 & D_2 & D_2D_1 \end{array}\right). \tag{4}$$

The following auxiliary notations are now considered:

$$D_s{=}D_2D_1;\ \widehat{D}{=}(I + D_s)^{-1};\ \widehat{C}{=}\begin{pmatrix} D_2C_1 & C_2 \end{pmatrix};\ H{=}D_2;$$

$$\widehat{B}{=}\begin{pmatrix} B_1 \\ B_2D_1 \end{pmatrix};\ F{=}\begin{pmatrix} I \\ O \end{pmatrix};\ V{=}\begin{pmatrix} O \\ B_2 \end{pmatrix};\ M{=}-\widehat{B}\widehat{D};$$

$$\Phi{=}\begin{pmatrix} A_1 & O \\ B_2C_1 & A_2 \end{pmatrix} - \widehat{B}\widehat{D}\widehat{C} = \widehat{A}{-}\widehat{B}\widehat{D}\widehat{C};\ R{=}V{-}\widehat{B}\widehat{D}H,$$

where Φ thus becomes the closed-loop state matrix.

Recall Theorem 6 from [13] which computes a worst-case bound for the steady-state limit cycles arising in LTI processes driven by fixed-point LTI regulators, assuming a diagonalizable closed-loop state matrix Φ.

Theorem 1: For an open-loop system $L(z)$ consisting of a quantized stabilizing numeric controller $K(z)$, with steps $\delta_e, \delta_x, \delta_u > 0$, in series with an ideal discretized process $G(z)$, then the guaranteeable worst-case deviation of $y[k]$

from the ideal closed-loop steady-state measurement $y^{(0)}[k]$ achievable by $K^{(0)}(z)$ without quantization errors is:

$$\sup_{k>k_\varepsilon}\left\|y[k]-y^{(0)}[k]\right\|\leq \tag{5}$$
$$\left\|\widehat{D}\widehat{C}P\right\|\frac{1}{1-\rho(\Phi)}\left\|P^{-1}R\right\|\left(\frac{\delta_x}{2}+\frac{\delta_u}{2}\right)+$$
$$\left\|\widehat{D}\widehat{C}P\right\|\frac{1}{1-\rho(\Phi)}\left(\left\|P^{-1}F\right\|\frac{\delta_x}{2}+\left\|P^{-1}M\right\|\frac{\delta_u}{2}\right)+$$
$$\left\|\widehat{D}H\right\|\left(\frac{\delta_x}{2}+\frac{\delta_u}{2}\right)+\left\|\widehat{D}D_s\right\|\frac{\delta_e}{2}\equiv\varepsilon_G\left(K,P,F\right),$$

where the matrix P diagonalizes Φ, i.e. $\Phi=P\cdot J_\Phi\cdot P^{-1}$, $J_\phi=\text{diag}\left(\lambda_1,\ldots,\lambda_{n_\ell}\right)$, in the complex number field.

III. Proposed Design Problem

A. Formulation

There are several available degrees-of-freedom in the quantization error bound $\varepsilon_G\left(K,P,F\right)$ from (5), such as the selection of discretization method for regulator $K^{(0)}$, the possibility to apply a similarity transformation to its state-space representation, and a coordinate change P for the Jordan canonical form of Φ. The term $\varepsilon.\left(\cdot,\cdot,F\right)$ is separately emphasized, as it does not depend on the regulator matrices in the same sense as R and M, as will be further exploited in Lemma 2 and Remark 2.

In case of the discretization method, its selection is usually preferable for the provided transient response of the regulator (see e.g. [3], [4]) and, as such, will not be the main focus of this section. Therefore, we assume that K provides the desired transient response and we will focus on the steady-state performance only.

The remainder of the design necessity is to propose an adequate scaling of the regulator matrices (A_1,B_1,C_1,D_1) and find the similarity matrix P to guarantee a least conservative bound to (5).

We start from a fixed quantized regulator K from (3). Denote its similarity transformation through a matrix $T\in\text{GL}_{n_c}(\mathbb{C})$ as a new regulator K_T with invariant input-output response:

$$K=\left(\begin{array}{c|c}A_1&B_1\\\hline C_1&D_1\end{array}\right)\overset{T}{\sim}\left(\begin{array}{c|c}T^{-1}A_1T&T^{-1}B_1\\\hline C_1T&D_1\end{array}\right)\equiv K_T. \tag{6}$$

As such, the following optimization problem arises in terms of possible similarity matrices for the Jordan form of Φ, forming the set $\mathcal{P}$, and of possible controller coordinate transformations, forming the set $\mathcal{T}$.

Problem 1: Given a discrete-time controller K, the least conservative upper bound of the closed-loop quantization error is the solution of the following optimization problem:

$$\mathcal{Q}=\min_{T\in\mathcal{T}}\min_{P\in\mathcal{P}}\varepsilon_G(K_{q,T},P,F). \tag{7}$$

While the first set is given by $\mathcal{T}=\text{GL}_{n_c}(\mathbb{C})$, the second set should be properly characterized. For the diagonalizable case, the following result will be considered.

Lemma 1: For a diagonalizable matrix Φ, the set of similarity matrices $P\in\text{GL}_{n_\ell}(\mathbb{C})$ can be obtained starting from a given similarity matrix P_0 left multiplied by an arbitrary nonsingular diagonal matrix.

Proof: Let $\Lambda(\Phi)=\{\lambda_1,\lambda_2,\ldots,\lambda_{n_\ell}\}$ be the spectrum of the matrix Φ. Because the state matrix Φ is diagonalizable, the algebraic multiplicity for each eigenvalue is 1, which implies that for each eigenspace we have $\dim V_{\lambda_i}=1$. Therefore we have:

$$V_{\lambda_i}=\text{Span}\{p_i\},\quad i=\overline{1,n_\ell},$$

which can be used to construct the initial matrix P_0:

$$P_0=\left[\ p_1\mid p_2\mid\ \ldots\ \mid p_{n_\ell}\ \right].$$

Moreover, all bases of V_{λ_i} can be characterized as $\{\alpha\cdot p_i\}$, where $\alpha\in\mathbb{C}^\star$ is a non-zero complex number, and the conclusion follows. ■

Corollary 1: For the particular case of having the closed-loop state matrix $\Phi\in\mathbb{R}^{n_\ell\times n_\ell}$ with all eigenvalues distinct complex numbers, the similarity matrices can be characterized as a product between an arbitrary diagonal matrix D_α, $\alpha\in\mathbb{R}_+^{n_\ell}$, and a similarity matrix P_0.

According to Corollary 1, the diagonal matrices D_α are a good choice for considering the additional scaling factor set $\mathcal{P}$ for minimizing the upper bound $\varepsilon_G(K,P_0,F)$ from Theorem 1, $\alpha\in\mathbb{R}_+^{n_\ell}$ representing n_ℓ degrees-of-freedom for the optimization problem. We will further use the shorthand notation for the similarity matrix $P_\alpha=D_\alpha P_0$, $\alpha\in\mathbb{R}_+^{n_\ell}$, where $P_0=P_\alpha|_{\alpha=\mathbf{1}}$.

Considering the set $\mathcal{T}=\text{GL}_{n_c}(\mathbb{C})$ will present a major issue due to its lack of connectivity. Therefore, in a similar manner with the case of set $\mathcal{P}$, a diagonal transformation D_ξ, $\xi\in\mathbb{R}_+^{n_c}$ of the initial regulator K will be considered, resulting the set of regulators $K_{D_\xi}\equiv K_\xi$, its initial form being $K^{(0)}=K_\xi|_{\xi=\mathbf{1}}$.

To consider the worst-case steady-state guaranteeable bound (5) in the context of an optimization problem with vector variables, define the objective function:

$$\mathcal{J}:\mathbb{R}_+^{n_c}\times\mathbb{R}_+^{n_\ell}\to\mathbb{R}_+,\quad \mathcal{J}\left(\xi,\alpha\right)=\varepsilon_G\left(K_\xi,P_\alpha,F\right).$$

The hypothesis is that a change in K implies a change in the state matrix Φ which, in turn, leads to the search of a different similarity matrix P. This sequentiality assumption to select a regulator K, based upon which the matrix P will be further computed, leads to the following minimization problem derived from Problem 1.

Problem 2: Given a discrete quantized controller K and a similarity matrix P_0, a least conservative upper bound of the closed-loop quantization error is the solution of the following optimization problem:

$$\mathcal{Q}=\min_{\xi\in\mathbb{R}_+^{n_c}}\min_{\alpha\in\mathbb{R}_+^{n_\ell}}\mathcal{J}\left(\xi,\alpha\right)=\varepsilon_G\left(K_\xi,P_\alpha,F\right). \tag{8}$$

The main improvement of Problem 2 consists in formulating the optimization problem in terms of vector variables instead of nonsingular matrices while still maintaining low conservativeness: (i) the set $\mathcal{P}$ is properly characterized, and (ii) each state of the controller is individually scaled in the balancing scheme.

Next, we present a mechanism to properly characterize the term $\varepsilon_G\left(K_\xi, P_\alpha, F\right)$. First, starting from a diagonal matrix $D_\xi \in \mathbb{R}^{n_c \times n_c}$, an extended diagonal matrix $\overline{D}_\xi$ will also be defined as:

$$\overline{D}_\xi = \begin{pmatrix} D_\xi & O \\ O & I \end{pmatrix} \in \mathbb{R}^{n_\ell \times n_\ell}. \tag{9}$$

Lemma 2: The diagonal scalings applied to the controller K and its similarity matrix P_0 can be formulated as a joint diagonal scaling applied to P_0 and an auxiliary diagonal scaling of the non-symmetric term F:

$$\varepsilon_G\left(K_\xi, P_\alpha, F\right) = \varepsilon_G\left(K, \left(\overline{D}_\xi^2 D_\alpha\right) P_0, \overline{D}_\xi F\right). \tag{10}$$

Proof: The coordinate transformation D_ξ applied to K has the following effect on the closed-loop state matrix Φ:

$$\Phi \overset{D_\xi}{\sim} \begin{pmatrix} D_\xi^{-1} A_1 D_\xi & O \\ B_2 C_1 D_\xi & A_2 \end{pmatrix} - \begin{pmatrix} D_\xi^{-1} B_1 \\ B_2 D_1 \end{pmatrix} \hat{D} \begin{pmatrix} D_2 C_1 D_\xi & C_2 \end{pmatrix} = \\ = \overline{D}_\xi^{-1} \Phi \overline{D}_\xi,$$

while, considering the diagonal scaling D_α of P_0, we have:

$$\Phi = \left(\overline{D}_\xi D_\alpha P_0\right) J_\Phi \left(P_0^{-1} D_\alpha^{-1} \overline{D}_\xi^{-1}\right). \tag{11}$$

Additionally, the scaling through D_ξ can be represented in a similar manner as a left multiplication by $\overline{D}_\xi^{-1}$ for the terms R and M, and a right multiplication by $\overline{D}_\xi$ for the term $\widehat{C}$ in (5). As such, considering the structure of $\varepsilon_G(K, P, F)$, one can easily obtain that:

$$\varepsilon_G\left(K_\xi, P_\alpha, F\right) = \varepsilon_G\left(K, \left(\overline{D}_\xi^2 D_\alpha\right) P_0, \overline{D}_\xi F\right), \tag{12}$$

which concludes the proof. ∎

The main problem of optimizing the controller coordinate transformations through $\xi \in \mathbb{R}_+^{n_c}$ consists in the possibility of obtaining arbitrarily small or large scaling matrices, i.e. $\|D_\xi\| \to \{0, \infty\}$, to minimize the upper bound of the steady-state quantization error. Let $\mathcal{N}_{x_c} = \|(A_1, B_1, I, O)\|$ 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 dynamics, respectively, assumed finite. However, considering Lemma 4 from [13], a maximum admissible regulator state signal norm $\overline{\mathcal{N}}_{x_c}$, also depending on the ADCs' and DACs' dynamic range should be imposed. Given that $\mathcal{N}_u$ is invariant to similarity transformations D_ξ, an upper bound $\overline{\mathcal{N}}_{x_c} = 2^{\lceil \log_2 \mathcal{N}_u \rceil}$ can be considered as a nonredundant starting point. Here, $\lceil \cdot \rceil$ signifies the ceiling function of a real number. This ensures the already necessary bound of the output signal and does not require any additional bits for state computations.

As such, enhancing Problem 2 according to Lemma 2 and the extra constraint on the state norm, a novel formulation of the tracking error (5) minimization problem occurs:

Problem 3: Given a discrete quantized controller K, with a maximum allowed $\mathcal{H}_\infty$-norm $\overline{\mathcal{N}}_{x_c}$ for its state signal, and a similarity matrix P_0, a least conservative upper bound of the closed-loop steady-state quantization error is the solution of the optimization problem:

$$\mathcal{Q} = \min_{(\xi,\alpha) \in \mathbb{R}_+^{n_c} \times \mathbb{R}_+^{n_\ell}} \mathcal{J}\left(\xi, \alpha\right) = \varepsilon_G\left(K, \left(\overline{D}_\xi^2 D_\alpha\right) P_0, \overline{D}_\xi F\right) \\ \text{s.t. } \left\| \left(D_\xi^{-1} A_1 D_\xi, D_\xi^{-1} B_1, I, O\right) \right\| < \overline{\mathcal{N}}_{x_c}. \tag{13}$$

As an overview, the domain of the objective function reduces from $\mathrm{GL}_{n_c}(\mathbb{R}) \times \mathrm{GL}_{n_\ell}(\mathbb{R}) < \mathbb{R}^{n_c^2} \times \mathbb{R}^{n_\ell^2}$ in the case of Problem 1 (which is non-connected) to $\mathbb{R}_+^{n_c} \times \mathbb{R}_+^{n_\ell}$ in Problem 2, and concluding with $\mathbb{R}_+^{n_c + n_\ell}$ with a single optimization variable (ξ, α) instead of a sequence of two minimization steps.

Remark 1: In a similar light to the pole-placement algorithm of [19], with the technical limitation that the multiplicity of the closed-loop poles cannot exceed the rank of the input matrix B_2, it may become desirable in this case to design the control system such that the closed-loop state matrix Φ is diagonalizable (which is not equivalent to forcing poles to have multiplicity one) in order to apply the framework of Theorem 1, which will lead to a simplified optimization problem as demonstrated through Problem 3.

B. Numeric Implementation Aspects

The resulting optimization Problems 1–3 are not convex by nature. With the following convention for the sign function:

$$\mathrm{sign}\left(x\right) = \begin{cases} -1, x < 0; \\ 1, x \geq 0, \end{cases} \tag{14}$$

the partial derivative of the matrix ∞-norm $\|\cdot\|$ is:

$$\frac{\partial \|X\|}{\partial x_{ij}} = \mathrm{sign}\left(x_{ij}\right) \delta_{kj}, \quad X \in \mathbb{R}^{n \times n} = \left[x_{ij}\right]_{i,j=\overline{1,n}}, \tag{15}$$

where δ_{kj} is the Kronecker delta function, with k being the row for which the maximum is achieved. As such, the functional $\mathcal{J}(\xi, \alpha)$ is differentiable, but its Jacobian is not continuous, being right-continuous only. As described in [20] and [21], such a minimization problem converges if the subgradient method is applied.

The experiments have been performed using the `fmincon` routine from MATLAB®, version R2022a, using several hyperparameter configurations and algorithms. From our findings, the sequential quadratic programming (`sqp`) algorithm works best, followed by the `interior-point` method, with its inherent advantage that it works for large-scale problems, but it halts at solutions with coarser tolerances and also tends to move away from the best found value in order to satisfy the first-order optimality conditions, and followed by the `active-set` algorithm which stalls prematurely for identical hyperparameters. The `trust-region-reflective` algorithm does not cover the constraint specified in (13).

The nonconvexity of the optimization problem can lead to a premature stopping of the algorithm into a local minimum point. A possible trick which can be used to reduce the possibility of early stopping is given in the next remark.

Remark 2: Starting from a given similarity matrix P_0, each permutation matrix Π leads to a new similarity matrix

ΠP_0 which can locally generate a new subset of similarity matrices of the set $\mathcal{P}$ and, considering that:

$$\varepsilon_G\left(K_\xi, P_\alpha, F\right) = \varepsilon_G\left(K, \left(\overline{D}_\xi^2 D_\alpha \Pi\right) \cdot P_0, \overline{D}_\xi F\right), \quad (16)$$

a maximum prescribed number of function evaluations can be used to find the best possible local decreasing direction.

An alternative approach regarding the reformulation of Problem 1 in order to reduce the conservativeness would be to consider $\mathcal{T} = \mathrm{GL}_{n_c}(\mathbb{R})$. As such, the optimization problem (13) could be written as:

$$\begin{aligned} \mathcal{Q} = \min_{(T,\alpha)\in\mathbb{R}^{n_c\times n_c}\times\mathbb{R}_+^{n_\ell}} \mathcal{J}\left(T,\alpha\right) = \varepsilon_G\left(K, \left(\overline{T}^2 D_\alpha\right) P_0, \overline{T} F\right) \\ \text{s.t. } \left\|\left(T^{-1}A_1T, T^{-1}B_1, I, O\right)\right\| < \overline{\mathcal{N}}_{x_c} \\ \mathrm{rank}(T) = n_c, \end{aligned} \quad (17)$$

where $\overline{T} = \mathrm{diag}\left(T, I\right) \in \mathbb{R}^{n_\ell\times n_\ell}$. The main disadvantage of such a formulation is the non-differentiability of the rank-based equality constraint.

IV. Numeric Example

To illustrate the practical implications of the proposed results, we consider an academic example characterized by an underdamped system with a pair of complex poles and a left half-plane zero to be controlled, with the desire to compute the least conservative quantization error bound given an arbitrary hardware configuration. The continuous-time model is:

$$\left(G(s)\right) : \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) = \left(\begin{array}{cc|c} -0.2 & -0.5 & 1 \\ 0.5 & 0 & 0 \\ \hline 0.1 & 1 & 0 \end{array}\right), \quad (18)$$

with singularities $\hat{s}_{1,2} = -0.1 \pm 0.4899j$ and $\mathring{s}_1 = -5$. The process is further discretized using the zero-order hold method and a sampling rate $T = 0.1[s]$, leading to an ideal numeric representation $G(z) = (A_2, B_2, C_2, D_2)$, as specified in (2).

The regulator has been obtained through a $\mathcal{H}_\infty$ synthesis, tuned using loop-shaping for an overdamped response, with resulting state-space representation $K^{(0)}(z) = (A_1, B_1, C_1, D_1)$:

$$\left(\begin{array}{ccc|c} 0.9999017 & -0.000633 & 0.0004463 & 0.107559 \\ -0.000633 & -0.773041 & -0.162164 & 0.345551 \\ -0.000446 & 0.1621641 & 0.8841128 & 0.243469 \\ \hline 0.1075598 & 0.3455512 & -0.243469 & 0.531996 \end{array}\right). \quad (19)$$

The default $\mathcal{H}_\infty$ norms of the input-state and input-output dynamics of $K^{(0)}(z)$ are $\mathcal{N}_{x_c} = \|(A_1, B_1, I, O)\| = 1085.3$ and $\mathcal{N}_u = \|(A_1, B_1, C_1, D_1)\| = 117.7$, respectively. Keeping in mind that $\mathcal{N}_u$ is invariant to similarity transformations (6), it adds no benefit to constrain $\mathcal{N}_{x_c} < 2^7 = 128$. Considering $\overline{\mathcal{N}}_{x_c}$=512, then the minimum number of bits in the signal word length becomes $L_{x_c} = \log_2(512) + \max\{L_{ADC}, L_{DAC}\}$. Consider a standard configuration of $L_{ADC} = 12$, $L_{DAC} = 13$, in the supported range $e[k], u[k] \in [-5, 5]\,[V]$, leading to $L_{x_c} = 22$, without including the sign bit. As such, the working resolutions become $\delta_e = 2.441 \times 10^{-3}$, $\delta_x = 1.192 \times 10^{-6}$, $\delta_u = 1.22 \times 10^{-3}$. The default tracking error bound computed using (5) and the `eig` routine from MATLAB, without solving Problem 3, is $\mathcal{Q}_0 = 12.685{\times}10^{-3}$, with a corresponding similarity matrix:

$$P_0 = \begin{pmatrix} 0.0003 & 0.6448 & 0.6448 & -0.8051 & -0.0674 \\ 0.9949 & p_{22} & \overline{p}_{22} & -0.0100 & -0.0808 \\ -0.0987 & p_{32} & \overline{p}_{32} & 0.5404 & 0.9687 \\ -0.0210 & p_{42} & \overline{p}_{42} & 0.1184 & 0.2076 \\ 0.0001 & p_{52} & \overline{p}_{52} & -0.2135 & -0.0868 \end{pmatrix}, \quad (20)$$

with $p_{22} = -0.0224 + 0.0001j$, $p_{32} = 0.1037 + 0.6244j$, $p_{42} = 0.2890 + 0.0926j$, $p_{52} = 0.0355 - 0.2996j$.

Performing the optimization (13) with the observations from Section III-B, the least guaranteeable tracking error becomes $\mathcal{Q}^\star = 0.924{\times}10^{-3}$ for a solution $(\xi^\star, \alpha^\star) \in \mathbb{R}_+^{n_c+n_\ell}$:

$$\xi^{\star\top} = \begin{pmatrix} 12.1729 & 8.6178 & 7.6911 \end{pmatrix}; \quad (21a)$$

$$\alpha^{\star\top} = \begin{pmatrix} 7.5078 & 8.6092 & 7.7229 & 8.3333 & 3.3261 \end{pmatrix}, \quad (21b)$$

which provides an improvement factor of $\mathcal{Q}_0/\mathcal{Q}^\star = 13.99$ beyond the default value as deduced strictly by the theory developed in [13]. The admissible range for the state signal becomes $\mathcal{N}_{x_c}^\star = 89.15 < \overline{\mathcal{N}}_{x_c}$, denoting a feasible solution. Thus, the regulator representation which guarantees this bound is $K_{\xi^\star}$, with a corresponding coordinate matrix $P_{\alpha^\star}$, up to a permutation Π of its columns. Its corresponding scaled state realization $K_{\xi^\star}(z)$ becomes:

$$\left(\begin{array}{ccc|c} 0.9999017 & -0.0004485 & 0.0002819 & 0.008836 \\ -0.000895 & -0.773042 & -0.144725 & 0.040097 \\ -0.0007064 & 0.1817046 & 0.8841128 & 0.031656 \\ \hline 1.3093134 & 2.9779049 & -1.8725425 & 0.531996 \end{array}\right). \quad (22)$$

To further assess the tightness of $\mathcal{Q}^\star$, consider a series of experiments using $N = 500$ Monte Carlo simulations, with step reference signals varying in the range $r \in [0.5, 1.5]$, midriser quantizers for the ADC and DAC blocks, and midtread for the internal computations, respectively, along with a simulation time t_{sim} large enough for the system output to stabilize under the prescribed $\mathcal{Q}^\star$ deviation. The computed closed-loop settling time is $t_s = 15.6[s]$, considering the $\pm 2\%$ convention. Assuming a dominant closed-loop pole with real part $\mathrm{Re}\{\hat{s}_0\} = -\frac{4}{t_s}$, the necessary time frame for simulation such that the output signal's dominant oscillating mode is attenuated in the range of the quantization error is computed:

$$t_{\text{sim}} > \frac{t_s}{4} \ln\left(\frac{1}{\mathcal{Q}^\star}\right) = 27.32[s]. \quad (23)$$

Considering $t_{\text{sim}} = 120[s]$, the Monte Carlo simulations lead to a coverage $[0, 13.43]\,[\%]$ for $\mathcal{Q}_0$, with maximum at $r = 1.0128$ and $[0, 99.82]\,[\%]$ for $\mathcal{Q}^\star$, with maximum achieved for a reference $r = 1.2323$, respectively. Additionally, the reference signal values for which the maximum coverage is attained are not unique. The results described in this section can be summarized through the illustration of Figure 3. This

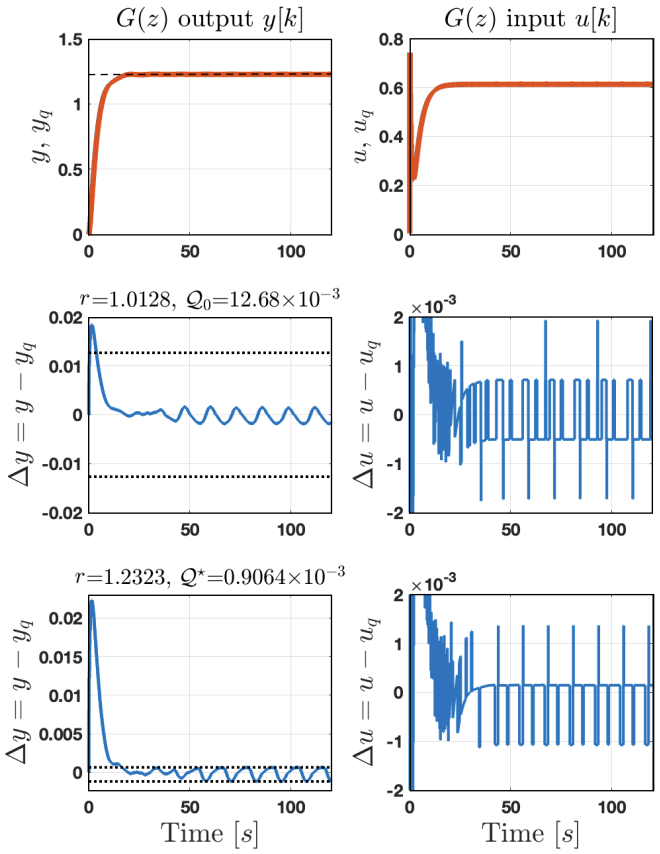

Fig. 3. Closed-loop simulations of the quantized control system, where the first column involves output signals $y, y_q, \Delta y = y - y_q$ and the second column involves input signals $u, u_q, \Delta u = u - u_q$. The first row shows the closed-loop step response for an arbitrary step reference, the second row illustrates the default application of Theorem 6 from [13] with a conservative bound $\mathcal{Q}_0$ and best achieved tightness of 13.43[%], while the third row shows the obtained improvements by solving Problem 3, with the bound $\mathcal{Q}^\star$ covered up to 99.82[%].

shows that the refined bound can become significantly lower than the default value and is also achievable in practice, so it cannot be further decreased. The command signal $u[k]$ is initially stochastic, but after a specific index $k > k_\varepsilon \in \mathbb{N}_+$, it converges either to a constant value or to a limit cycle trajectory.

V. Conclusions and Future Works

Obtained results can be summarized in the formulation of a low-dimensionality scaling problem to minimize the tracking error materialized through steady-state deviations or limit cycles and the guarantee of local convergence for its solution. The presented framework can be used as a design specification for the numeric controller to maintain the closed-loop poles far from the unit circle, as the term $\frac{1}{1-\rho(\Phi)}$ is the main degree-of-freedom of the fixed-point quantization error bound, given a specific hardware configuration. An adjacent design problem can be employed in rapid control prototyping to deduce the coarsest resolutions, equivalent to the least expensive hardware configuration, to guarantee the control system precision under a prescribed tolerance.

Research directions spanning from this work can be distinguished on three fronts: (i) generalizations to several classes of nonlinear systems, with focus on input-affine systems, frequently-arising in the field of robotics, (ii) different quantizer functions, such as logarithmic or floating-point and (iii) sensitivity analysis with mathematical guarantees for the application of output feedback linearization techniques through Lie derivatives or the Koopman operator.

References

[1] M. Șușcă, V. Mihaly, P. Dobra, Sampling Rate Selection for Multi-Loop Cascade Control Systems in an Optimal Manner, *IET Control Theory & Applications*, vol. 17(8), pp 1073–1087, May 2023.

[2] O.M.-Escrig, J.-A. Romero-Pérez, Regular quantisation with hysteresis: a new sampling strategy for event-based PID control systems, *IET Control Theory & Applications*, 14(15), pp. 2163–2175, Oct. 2020.

[3] A.G. Yepes, F.D. Freijedo, J. Doval-Gandoy, Ó. López, J. Malvar, P. Fernandez-Comesaña, Effects of Discretization Methods on the Performance of Resonant Controllers, *IEEE Transactions on Power Electronics*, vol. 25, no. 7, pp. 1692–1712, July 2010.

[4] M. Șușcă, V. Mihaly, P. Dobra, Maintaining Robust Stability and Performance through Sampling and Quantization, *IEEE American Control Conference (ACC)*, pp. 3852–3858, San Diego, CA, 2023.

[5] P. Wang, X. He, Q. Chen, A. Cheng, Q. Liu, J. Cheng, Unsupervised Network Quantization via Fixed-Point Factorization, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2706–2720, June 2021, doi: 10.1109/TNNLS.2020.3007749.

[6] M. Fu, L. Xie, The sector bound approach to quantized feedback control, *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1698–1711, Nov. 2005, doi: 10.1109/TAC.2005.858689.

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

[8] P.H. Petkov, T.N. Slavov, J.K. Kralev, *Design of Embedded Robust Control Systems Using MATLAB®/Simulink®*, The Institution of Engineering and Technology (IET), 2018.

[9] B. Picasso, A. Bicchi, On the Stabilization of Linear Systems Under Assigned I/O Quantization, *IEEE Transactions on Automatic Control*, vol. 52(10), pp. 1994–2000, 2007, doi:10.1109/TAC.2007.904283.

[10] D.F. Coutinho, M. Fu, C.E. de Souza, Input and Output Quantized Feedback Linear Systems, *IEEE Transactions on Automatic Control*, vol. 55(10), pp. 761-766, 2010, doi:10.1109/TAC.2010.2040497.

[11] H. Peng, A. Prodic, E. Alarcon, D. Maksimovic, Modeling of Quantization Effects in Digitally Controlled DC–DC Converters, *IEEE Trans. on Power Electronics*, vol. 22, no. 1, pp. 208–215, Jan. 2007.

[12] A. Abdullah, F. Musolino, P.S. Crovetti, Limit-Cycle Free, Digitally-Controlled Boost Converter Based on DDPWM, *IEEE Access*, vol. 11, pp. 9403–9414, 2023, doi: 10.1109/ACCESS.2023.3239883.

[13] M. Şuşcă, V. Mihaly, P. Dobra, Fixed-Point Uniform Quantization Analysis for Numerical Controllers, *IEEE 61st Conference on Decision and Control (CDC)*, Cancún, Mexico, pp. 3681–3686, 2022.

[14] J. Ma, W. Lan, X. Yu, Quantized feedback control of linear system with performance barrier, *International Journal of Robust and Nonlinear Control*, vol. 32, no. 12, pp. 7113–7131, August 2022.

[15] M.S. Darup, A. Redder, D. Quevedo, A fixed-point implementation of explicit MPC laws, *IEEE American Control Conference (ACC)*, Milwaukee, WI, USA, pp. 749–755, 2018.

[16] Y. Zhang, J.-F. Zhang, X.-K. Liu, Z.Liu, Quantized-output feedback model reference control of discrete-time linear systems, *Automatica*, 110027, vol. 137, March 2022, doi:10.1016/j.automatica.2021.110027.

[17] Y. Xu, Y. Zhang, J.-F. Zhang, A Pole Placement-Based Output Tracking Control Scheme by Finite-and-Quantized Output Feedback, *IEEE Control Systems Letters*, vol. 6, pp. 3200–3205, 2022.

[18] T. Xu, Z. Duan, Z. Sun, G. Chen, A unified control method for consensus with various quantizers, *Automatica*, vol. 136, 110090, February 2022, doi:10.1016/j.automatica.2021.110090.

[19] J. Kautsky, N.K. Nichols, P. Van Dooren, Robust pole assignment in linear state feedback, *Int. Journal of Control*, vol. 41, no. 5, 1985.

[20] A.S. Lewis, Nonsmooth optimization and robust control, *Annual Reviews in Control*, Volume 31, Issue 2, pp. 167–177, 2007.

[21] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1990.