Shift and Delta Operator Realizations for Digital Controllers with Finite-Word-Length Considerations*

J. Wu [1], S. Chen [2†], G. Li [3], R.H. Istepanian [4] and J. Chu [1]

[1] National Laboratory of Industrial Control Technology
Institute of Industrial Process Control
Zhejiang University, Hangzhou, 310027, P. R. China

[2] Department of Electronics and Computer Science
University of Southampton, Highfield
Southampton SO17 1BJ, U.K.

[3] School of of Electrical and Electronic Engineering
Nanyang Technological University, Singapore

[4] Department of Electrical and Computer Engineering
Ryerson Polytechnic University
Toronto, Ontario, Canada M5B 2K3

Abstract

This paper addresses implementation issues of digital controllers with finite word length (FWL) considerations. Both the shift and delta operator parameterizations of a general controller structure are considered. A unified formulation is adopted to derive a computationally tractable stability related measure that describes FWL closed-loop stability characteristics of different controller realizations. Within a given operator parameterization, the optimal FWL controller realization, which maximizes the proposed stability related measure, is the solution of a nonlinear optimization problem. Relationship between the z-operator and δ-operator controller parameterizations is analyzed, and it is shown that the δ parameterization has better FWL closed-loop stability margin than the z domain approach under a mild condition. A design example is included to verify the theoretical analysis and to illustrate the proposed optimization procedure.

*This work was partly supported by the National Natural Science Foundation of China under Grant 69504010 and Cao Guangbiao Foundation of Zhejiang University, China.

†Contact author. Tel/Fax: +44 (0)23 8059 6660/4508; Email: sqc@ecs.soton.ac.uk

1 Introduction

Modern controllers are typically implemented digitally, and it is well-known that a designed stable control system may achieve a lower than predicted performance or even become unstable when the control law is implemented with a finite-precision device due to the FWL effects. For many industrial and mass-market consumer applications, fixed-point implementations are more desired for the reasons of cost, simplicity, speed, memory space and power consumption. With a fixed-point processor, however, the detrimental FWL effects are markedly increased due to a reduced precision. The FWL effects on the closed-loop stability depend on the controller realization structure. This property can be utilized to "select" controller realization in order to improve the "robustness" of closed-loop stability under controller parameter perturbations. Currently, two approaches exist for determining the optimal controller realizations under different criteria, namely pole sensitivity measures [1]-[4] and complex stability radius measures [5],[6].

In the first approach, the pole sensitivity measures based on an l_2 norm [2] and an l_1 norm [3] are used to quantify the FWL effects on closed-loop stability. This approach leads to a nonlinear and non-smooth optimization problem in finding an optimal FWL controller realization. The need to solve for such a non-convex and non-smooth optimization problem had been seen as a disadvantage, as conventional optimization algorithms [7],[8], which are better known to the control community, may not guarantee to find a true optimal realization. However, the efficient global optimization techniques to tackle this kind of difficult optimization problems [9]-[14] are now widely available. More recently, Fialho and Georgiou [6] used the complex stability radius measure to formulate an optimal FWL controller realization problem that can be represented as a special H_∞ norm minimization problem and solved with the method of linear matrix inequality [15],[16]. In this second approach, the FWL perturbations are assumed to be complex-valued. Although this assumption is somewhat "artificial", the approach based on the complex stability radius measure has certain attractive features and requires further investigation.

Most studies on the FWL stability issues only consider the closed-loop systems with output feedback (OF) controllers. It is well known that there exists another class of controllers, namely observer-based (OB) controllers [17],[18]. Because state-space methods and observer theory are combined to form a direct multi-variable approach to linear control system synthesis [18], the design of OB controllers is more transparent and simpler than the design of OF controllers. Li and Gevers [19] have studied the sensitivity and the roundoff noise gain of the closed-loop system transfer function with an FWL implemented full-order OB controller. A recent study [20] has investigated the effects of FWL implementation on the closed-loop stability for full-order OB controllers. The first contribution of this paper is to develop a new framework of optimal FWL controller realizations for the generic digital controller structure that includes all the OF and OB controllers. A computationally tractable stability related measure is employed for the unified controller structure, using the well-tested pole sensitivity measure with the l_1 norm [3].

In most of the above-mentioned studies, digital controller structures are described and realized with the usual shift operator z. A discrete-time system can also be described and realized with a different operator, called the delta operator δ [21]. Two major advantages are known for the use of δ operator parameterization: a theoretically unified formulation of continuous-time and discrete-time systems; and better numerical properties in FWL implementations [1]. The benefits of using the δ operator as opposed to the shift operator in signal processing and control applications have been investigated [22]-[25]. In particular, a recent work has addressed the FWL closed-loop stability issues of OF controller structures using the δ operator formulation [26]. The second new contribution of this paper is to adopt a unified formulation to include both the z and δ operator parameterizations of the generic finite-precision controller structure and to analyze the underlying relationship between these two controller parameterizations.

The paper is organized as follows. Section 2 is devoted to establishing necessary notations and definitions. A generalized operator is adopted to represent either z or δ operator. This enables a unified formulation in the discussion. The problem to be dealt with is also formulated in this section. A closed-loop stability related measure that is computationally tractable is introduced in Section 3. The optimal controller realization problem is defined in Section 4, which is to find a realization that maximizes the proposed measure. The detailed optimization framework for obtaining the optimal FWL controller realization is also presented in this section. Section 5 is devoted to analyzing the underlying relationship between the z and δ controller realizations and to establishing the condition for the δ realization to have better FWL stability margin. In Section 6, a numerical design example is used to verify the theoretical results and to demonstrate the effectiveness of the proposed optimization strategy. The paper concludes in Section 7.

2 Notations, definitions and problem formulation

Let $\mathcal{R}$ denote the field of real numbers and $\mathcal{C}$ the field of complex numbers. For a complex-valued matrix $\mathbf{U} \in \mathcal{C}^{p\times q}$ with elements u_{ij}, we define the following matrix norm:

$$\|\mathbf{U}\|_S \triangleq \sum_{i=1}^{p}\sum_{j=1}^{q} |u_{ij}| \tag{1}$$

Let $\mathrm{Vec}(\cdot)$ be the column stacking operator such that $\mathrm{Vec}(\mathbf{U})$ is a qp-dimensional vector. As usual, $\mathbf{U}^T$ is the transposed matrix of $\mathbf{U}$, $\mathbf{U}^H$ is the Hermitian adjoint matrix of $\mathbf{U}$, and $\mathbf{U}^*$ is conjugate to $\mathbf{U}$. For a squared real-valued matrix $\mathbf{M} \in \mathcal{R}^{p\times p}$, let $\{\lambda_i(\mathbf{M}),\ 1 \le i \le p\}$ denote its eigenvalues. For diagonalisable $\mathbf{M}$, let $\mathbf{x}_i(\mathbf{M})$ be the right eigenvector corresponding to $\lambda_i(\mathbf{M})$, that is

$$\mathbf{M}\,\mathbf{x}_i(\mathbf{M}) = \lambda_i(\mathbf{M})\,\mathbf{x}_i(\mathbf{M}) \tag{2}$$

Since $\mathbf{M}$ is diagonalisable, the matrix

$$\mathbf{M}_x \triangleq [\,\mathbf{x}_1(\mathbf{M}) \quad \cdots \quad \mathbf{x}_p(\mathbf{M})\,] \tag{3}$$

is invertible. Define:

$$\mathbf{M}_y = [\,\mathbf{y}_1(\mathbf{M}) \quad \cdots \quad \mathbf{y}_p(\mathbf{M})\,] \triangleq \mathbf{M}_x^{-H} \tag{4}$$

$\mathbf{y}_i(\mathbf{M})$ is called the reciprocal left eigenvector corresponding to $\mathbf{x}_i(\mathbf{M})$ for the reason shown in the following lemma.

Lemma 1 $\mathbf{y}_i^H(\mathbf{M})\,\mathbf{M} = \lambda_i(\mathbf{M})\,\mathbf{y}_i^H(\mathbf{M})$, $\forall i$.

Proof: Denote

$$\mathbf{\Sigma} \triangleq \begin{bmatrix} \lambda_1(\mathbf{M}) & & \\ & \ddots & \\ & & \lambda_p(\mathbf{M}) \end{bmatrix} \tag{5}$$

Clearly, $\mathbf{M}\mathbf{M}_x = \mathbf{M}_x\mathbf{\Sigma}$. It then follows from $\mathbf{M}_x\mathbf{M}_y^H = \mathbf{M}_y^H\mathbf{M}_x = \mathbf{I}$, the identity matrix, that $\mathbf{M}_y^H\mathbf{M} = \mathbf{\Sigma}\mathbf{M}_y^H$, which leads to lemma 1.

A discrete-time system can be described using either the usual z operator or the so-called δ operator. The latter is defined as [21]

$$\delta \triangleq \frac{z-1}{h} \tag{6}$$

where h is a positive real constant[1]. Let the state-space representation of a discrete-time system using z operator be

$$\begin{cases} z\mathbf{x}(k) = \mathbf{A}_z\mathbf{x}(k) + \mathbf{B}_z\mathbf{u}(k) \\ \mathbf{y}(k) = \mathbf{C}_z\mathbf{x}(k) + \mathbf{D}_z\mathbf{u}(k) \end{cases} \tag{7}$$

where all the matrices and vectors are real-valued and are assumed to have proper dimensions, and $z\mathbf{x}(k) = \mathbf{x}(k+1)$, as z is the forward shift operator. We can described the same discrete-time system by

$$\begin{cases} \delta\mathbf{x}(k) = \mathbf{A}_\delta\mathbf{x}(k) + \mathbf{B}_\delta\mathbf{u}(k) \\ \mathbf{y}(k) = \mathbf{C}_\delta\mathbf{x}(k) + \mathbf{D}_\delta\mathbf{u}(k) \end{cases} \tag{8}$$

using δ operator [27],[28], where

$$\mathbf{A}_\delta = \frac{\mathbf{A}_z - \mathbf{I}}{h},\ \mathbf{B}_\delta = \frac{\mathbf{B}_z}{h},\ \mathbf{C}_\delta = \mathbf{C}_z,\ \mathbf{D}_\delta = \mathbf{D}_z \tag{9}$$

with $\mathbf{I}$ denoting the identity matrix of appropriate dimension. Obviously, (7) and (8) are the two equivalent representations of the same system. The following lemma relates the eigenvalues and eigenvectors of $\mathbf{A}_z$ to those of $\mathbf{A}_\delta$.

Lemma 2 With a proper index order, $\{\lambda_i(\mathbf{A}_z)\}$ and $\{\lambda_i(\mathbf{A}_\delta)\}$ can be one-to-one mapped with

$$\lambda_i(\mathbf{A}_z) = 1 + h\lambda_i(\mathbf{A}_\delta),\ \forall i \tag{10}$$

Let $\lambda_i(\mathbf{A}_\delta)$ and $\lambda_i(\mathbf{A}_z)$ be related with (10). Then they have the same eigenvector set.

Proof: Let $\mathbf{x}_i(\mathbf{A}_z)$ be an eigenvector corresponding to $\lambda_i(\mathbf{A}_z)$. It follows from (9) that $\lambda_i(\mathbf{A}_z)\mathbf{x}_i(\mathbf{A}_z) = \mathbf{A}_z\mathbf{x}_i(\mathbf{A}_z) = h\,\mathbf{A}_\delta\mathbf{x}_i(\mathbf{A}_z) + \mathbf{x}_i(\mathbf{A}_z)$, which means that

$$\frac{\lambda_i(\mathbf{A}_z) - 1}{h}\mathbf{x}_i(\mathbf{A}_z) = \mathbf{A}_\delta\mathbf{x}_i(\mathbf{A}_z) \tag{11}$$

[1] In [21], h is limited to the sampling period. This constraint is removed in [24].

This, by definition, implies that $\frac{\lambda_i(\mathbf{A}_z)-1}{h}$ is an eigenvalue of $\mathbf{A}_\delta$, denoted as $\lambda_i(\mathbf{A}_\delta)$, and $\mathbf{x}_i(\mathbf{A}_z)$ is also an eigenvector of $\mathbf{A}_\delta$, corresponding to $\lambda_i(\mathbf{A}_\delta)$. Using the same procedure, one can show that if $\mathbf{x}_i(\mathbf{A}_\delta)$ is an eigenvector of $\lambda_i(\mathbf{A}_\delta)$, it is also an eigenvector related to an eigenvalue of $\mathbf{A}_z$ given by (10). This completes the proof.

It is well known that the discrete-time system $(\mathbf{A}_z, \mathbf{B}_z, \mathbf{C}_z, \mathbf{D}_z)$ is stable if and only if

$$|\lambda_i(\mathbf{A}_z)| < 1, \ \forall i \tag{12}$$

From lemma 2, we have the stability condition for the same system described using δ operator.

Lemma 3 The discrete-time system $(\mathbf{A}_\delta, \mathbf{B}_\delta, \mathbf{C}_\delta, \mathbf{D}_\delta)$ is stable if and only if

$$\left|\lambda_i(\mathbf{A}_\delta) + \frac{1}{h}\right| < \frac{1}{h}, \ \forall i \tag{13}$$

For the notational conciseness, we introduce a "generalized" operator ρ for the discrete-time systems. It is understood that $\rho = z$ or δ, depending on which operator is actually used. The two state-space representations (7) and (8) can then be unified as:

$$\begin{cases} \rho\,\mathbf{x}(k) = \mathbf{A}_\rho \mathbf{x}(k) + \mathbf{B}_\rho \mathbf{u}(k) \\ \mathbf{y}(k) = \mathbf{C}_\rho \mathbf{x}(k) + \mathbf{D}_\rho \mathbf{u}(k) \end{cases} \tag{14}$$

The use of this notation will avoid repeated derivations for the two operators in the following discussion.

Consider the discrete-time closed-loop control system depicted in Figure 1, where the linear time-invariant plant $\hat{P}$ has a state-space representation

$$\begin{cases} \rho\,\mathbf{x}(k) = \mathbf{A}_\rho \mathbf{x}(k) + \mathbf{B}_\rho \mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}_\rho \mathbf{x}(k) \end{cases} \tag{15}$$

which is assumed to be strictly proper, completely state controllable, and completely state observable, with $\mathbf{A}_\rho \in \mathcal{R}^{n\times n}$, $\mathbf{B}_\rho \in \mathcal{R}^{n\times p}$ and $\mathbf{C}_\rho \in \mathcal{R}^{q\times n}$; and the digital stabilizing controller $\hat{C}$ is described by the state-space representation:

$$\begin{cases} \rho\,\mathbf{v}(k) = \mathbf{F}_\rho \mathbf{v}(k) + \mathbf{G}_\rho \mathbf{y}(k) + \mathbf{H}_\rho \mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}_\rho \mathbf{v}(k) + \mathbf{M}_\rho \mathbf{y}(k) \end{cases} \tag{16}$$

where $\mathbf{F}_\rho \in \mathcal{R}^{m\times m}$, $\mathbf{G}_\rho \in \mathcal{R}^{m\times q}$, $\mathbf{J}_\rho \in \mathcal{R}^{p\times m}$, $\mathbf{M}_\rho \in \mathcal{R}^{p\times q}$ and $\mathbf{H}_\rho \in \mathcal{R}^{m\times p}$. The controller depicted in Figure 1 is generic and includes all the OF and OB controllers: $\hat{C}$ is an OF controller when $\mathbf{H}_\rho = \mathbf{0}$; a full-order OB controller when $\mathbf{F}_\rho = \mathbf{A}_\rho - \mathbf{G}_\rho\mathbf{C}_\rho$, $\mathbf{M}_\rho = \mathbf{0}$ and $\mathbf{H}_\rho = \mathbf{B}_\rho$; a reduced-order OB controller, otherwise [17],[18].

It is a basic property of the linear system theory that the state-space realization $(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho)$ of the general controller $\hat{C}$ is not unique. Assume that a realization $(\mathbf{F}_{\rho 0}, \mathbf{G}_{\rho 0}, \mathbf{J}_{\rho 0}, \mathbf{M}_{\rho 0}, \mathbf{H}_{\rho 0})$ has been designed through a controllers design procedure for $\hat{C}$. All the realizations of $\hat{C}$ form a realization set:

$$\begin{aligned}\mathcal{S}_\rho \triangleq \{(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho) : \mathbf{F}_\rho = \mathbf{T}_\rho^{-1}\mathbf{F}_{\rho 0}\mathbf{T}, \mathbf{G}_\rho = \mathbf{T}_\rho^{-1}\mathbf{G}_{\rho 0},\\ \mathbf{J}_\rho = \mathbf{J}_{\rho 0}\mathbf{T}_\rho, \mathbf{M}_\rho = \mathbf{M}_{\rho 0}, \mathbf{H}_\rho = \mathbf{T}_\rho^{-1}\mathbf{H}_{\rho 0}\}\end{aligned} \tag{17}$$

where $\mathbf{T}_\rho \in \mathcal{R}^{m\times m}$ is any real-valued non-singular matrix, called a similarity transformation. Any two realizations in $\mathcal{S}_\rho$ are completely equivalent if they are implemented with infinite precision. Let

$$\mathbf{w}_\rho = \begin{bmatrix} w_{\rho 1} \\ w_{\rho 2} \\ \vdots \\ w_{\rho N} \end{bmatrix} \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_\rho) \\ \mathrm{Vec}(\mathbf{G}_\rho) \\ \mathrm{Vec}(\mathbf{J}_\rho) \\ \mathrm{Vec}(\mathbf{M}_\rho) \\ \mathrm{Vec}(\mathbf{H}_\rho) \end{bmatrix}, \quad \mathbf{w}_{\rho 0} \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{G}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{J}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{M}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{H}_{\rho 0}) \end{bmatrix} \tag{18}$$

where $N = (m+p)(m+q) + mp$. We also refer to $\mathbf{w}_\rho$ as a realization of $\hat{C}$. The stability of the closed-loop system in Figure 1 depends on the eigenvalues of the transition matrix

$$\begin{aligned}\bar{\mathbf{A}}(\mathbf{w}_\rho) &= \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{G}_\rho\mathbf{C}_\rho + \mathbf{H}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho\mathbf{J}_\rho \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_{\rho 0}\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_{\rho 0} \\ \mathbf{G}_{\rho 0}\mathbf{C}_\rho + \mathbf{H}_{\rho 0}\mathbf{M}_{\rho 0}\mathbf{C}_\rho & \mathbf{F}_{\rho 0} + \mathbf{H}_{\rho 0}\mathbf{J}_{\rho 0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho^{-1} \end{bmatrix} \bar{\mathbf{A}}(\mathbf{w}_{\rho 0}) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho \end{bmatrix}\end{aligned} \tag{19}$$

Let us define the "stability margin" of $\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))$ as

$$StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))) \triangleq \begin{cases} 1 - |\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))|, & \text{if } \rho = z \\ \frac{1}{h} - \left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta)) + \frac{1}{h}\right|, & \text{if } \rho = \delta \end{cases} \tag{20}$$

It follows, from the fact that the closed-loop system is designed to be stable,

$$StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))) = StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))) > 0, \ \forall i \in \{1, \cdots, m+n\} \tag{21}$$

which implies that all the different controller realizations $\mathbf{w}_\rho \in \mathcal{S}_\rho$ achieve exactly the same closed-loop poles if they are implemented with infinite precision.

In practice, however, a controller can only be implemented with finite precision. Different realizations will have different FWL characteristics. When $\mathbf{w}_\rho$ is implemented using a fixed-point processor, it is perturbed into $\mathbf{w}_\rho + \Delta\mathbf{w}_\rho$. Assume that the fixed-point processor uses B_f bits for the fractional part of a number. Define

$$\epsilon = 2^{-B_f} \tag{22}$$

Then, each element of $\Delta\mathbf{w}_\rho$ is bounded by $\pm\frac{\epsilon}{2}$, that is,

$$\mu(\Delta\mathbf{w}_\rho) \triangleq \max_{i\in\{1,\cdots,N\}} |\Delta w_{\rho i}| \leq \frac{\epsilon}{2} \tag{23}$$

With the perturbation $\Delta\mathbf{w}_\rho$, $\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))$ is moved to $\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho + \Delta\mathbf{w}_\rho))$. If an eigenvalue of $\bar{\mathbf{A}}(\mathbf{w}_\rho + \Delta\mathbf{w}_\rho)$ crosses over the stability boundary, the closed-loop system originally designed to be stable will become unstable. Intuitively, different controller realizations will have different degrees of robustness to this FWL effects. It is highly desired to be able to quantify how robustness a controller realization is in terms of its closed-loop stability under FWL implementation.

3 An FWL stability related measure

Roughly speaking, how easily the FWL error $\Delta\mathbf{w}_\rho$ can cause a stable control system to become unstable is determined by how close $\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))$ are to the stability boundary and how sensitive they are to the controller parameter perturbations. The first factor is determined by the stability margins of the eigenvalues, and the second factor is characterized by the derivatives of the eigenvalues with respect to the controller parameters. In this paper, we consider the following stability related measure [26]:

$$\mu_\rho(\mathbf{w}_\rho) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho)))}{\sum_{j=1}^{N}\left|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial w_{\rho j}}\right|} \tag{24}$$

Heuristically, the use of $\mu_\rho(\mathbf{w}_\rho)$ as a stability measure of $\mathbf{w}_\rho$ can be justified as follows. When the FWL error $\Delta\mathbf{w}_\rho$ is small, we have

$$\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho + \Delta\mathbf{w}_\rho)) \approx \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho)) + \sum_{j=1}^{N} \frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial w_{\rho j}}\Delta w_{\rho j},\ \forall i \in \{1,\ldots,m+n\} \tag{25}$$

It then follows that

$$-StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho + \Delta\mathbf{w}_\rho))) \le -StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))) + \sum_{j=1}^{N} \left| \frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial w_{\rho j}} \right| |\Delta w_{\rho j}|$$

$$\le -StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))) + \mu(\Delta\mathbf{w}_\rho) \sum_{j=1}^{N} \left| \frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial w_{\rho j}} \right|, \ \forall i \tag{26}$$

If $\mu(\Delta\mathbf{w}_\rho) < \mu_\rho(\mathbf{w}_\rho)$, from (24) and (26), we have

$$StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho + \Delta\mathbf{w}_\rho))) > 0 \tag{27}$$

This means that the closed-loop system remains stable under the FWL error $\Delta\mathbf{w}_\rho$. In other words, for a given realization $\mathbf{w}_\rho$, the closed-loop stability can tolerate those FWL perturbations $\Delta\mathbf{w}_\rho$, whose elements have magnitudes less than $\mu_\rho(\mathbf{w}_\rho)$. The larger $\mu_\rho(\mathbf{w}_\rho)$ is, the larger FWL errors the closed-loop system can tolerate.

The assumption that the controller coefficient perturbations are small is generally valid. For example, with a 10-bit accuracy for B_f, the FWL errors are bounded by 0.5%. The stability related measure $\mu_\rho(\mathbf{w}_\rho)$ is computationally tractable. To compute $\mu_\rho(\mathbf{w}_\rho)$, we need $\{\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial w_{\rho j}}\}$, which can be calculated with the following theorem.

Theorem 1 Let $\mathbf{A} = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2 \in \mathcal{R}^{m\times m}$ be diagonalisable where $\mathbf{X} \in \mathcal{R}^{l\times r}$, and $\mathbf{M}_0$, $\mathbf{M}_1$ and $\mathbf{M}_2$ are independent of $\mathbf{X}$ with proper dimensions. Let $\lambda_i(\mathbf{A})$ denote the ith eigenvalue of $\mathbf{A}$, and let $\mathbf{x}_i(\mathbf{A})$ and $\mathbf{y}_i(\mathbf{A})$ be the right and reciprocal left eigenvectors corresponding to $\lambda_i(\mathbf{A})$ respectively. Then

$$\frac{\partial \lambda_i(\mathbf{A})}{\partial \mathbf{X}} \triangleq \begin{bmatrix} \frac{\partial \lambda_i(\mathbf{A})}{\partial x_{11}} & \cdots & \frac{\partial \lambda_i(\mathbf{A})}{\partial x_{1r}} \\ \vdots & \cdots & \vdots \\ \frac{\partial \lambda_i(\mathbf{A})}{\partial x_{l1}} & \cdots & \frac{\partial \lambda_i(\mathbf{A})}{\partial x_{lr}} \end{bmatrix} = \mathbf{M}_1^T \mathbf{y}_i^*(\mathbf{A}) \mathbf{x}_i^T(\mathbf{A}) \mathbf{M}_2^T \tag{28}$$

The proof of this theorem can be found in [26].

Remark 1: When $\mathbf{A}$ has no repeated poles, all the eigenvectors corresponding to $\lambda_i(\mathbf{A})$ can be characterized as $\mathbf{x}_i(\mathbf{A}) = \eta_i \mathbf{x}_{i0}(\mathbf{A})$, where η_i is a nonzero complex-valued constant and $\mathbf{x}_{i0}(\mathbf{A})$ is a given eigenvector of $\lambda_i(\mathbf{A})$. It is then easy to show that the corresponding reciprocal left eigenvector is $\mathbf{y}_i(\mathbf{A}) = \frac{1}{\eta_i^*}\mathbf{y}_{i0}(\mathbf{A})$ with $\mathbf{y}_{i0}(\mathbf{A})$ the reciprocal left eigenvector corresponding to $\mathbf{x}_{i0}(\mathbf{A})$. Therefore, $\mathbf{y}_i^*(\mathbf{A})\mathbf{x}_i^T(\mathbf{A}) = \mathbf{y}_{i0}^*(\mathbf{A})\mathbf{x}_{i0}^T(\mathbf{A})$, which means that though each eigenvalue has different eigenvectors, its sensitivity given by (28) is unique. In the sequel, the closed-loop system is assumed to have no repeated poles.

From (19), we know that

$$\bar{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{B}_\rho \mathbf{J}_\rho \\ \mathbf{G}_\rho \mathbf{C}_\rho + \mathbf{H}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{H}_\rho \mathbf{J}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{F}_\rho [\,\mathbf{0} \quad \mathbf{I}\,] \tag{29}$$

$$\bar{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{B}_\rho \mathbf{J}_\rho \\ \mathbf{H}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho \mathbf{J}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{G}_\rho [\,\mathbf{C}_\rho \quad \mathbf{0}\,] \tag{30}$$

$$\bar{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{0} \\ \mathbf{G}_\rho \mathbf{C}_\rho + \mathbf{H}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{F}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{B}_\rho \\ \mathbf{H}_\rho \end{bmatrix} \mathbf{J}_\rho [\,\mathbf{0} \quad \mathbf{I}\,] \tag{31}$$

$$\bar{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho & \mathbf{B}_\rho \mathbf{J}_\rho \\ \mathbf{G}_\rho \mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho \mathbf{J}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{B}_\rho \\ \mathbf{H}_\rho \end{bmatrix} \mathbf{M}_\rho [\,\mathbf{C}_\rho \quad \mathbf{0}\,] \tag{32}$$

$$\bar{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho \mathbf{M}_\rho \mathbf{C}_\rho & \mathbf{B}_\rho \mathbf{J}_\rho \\ \mathbf{G}_\rho \mathbf{C}_\rho & \mathbf{F}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{H}_\rho [\,\mathbf{M}_\rho \mathbf{C}_\rho \quad \mathbf{J}_\rho\,] \tag{33}$$

Applying theorem 1 gives rise to

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{F}_\rho} = [\,\mathbf{0} \quad \mathbf{I}\,]\, \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \tag{34}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{G}_\rho} = [\,\mathbf{0} \quad \mathbf{I}\,]\, \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \begin{bmatrix} \mathbf{C}_\rho^T \\ \mathbf{0} \end{bmatrix} \tag{35}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{J}_\rho} = [\,\mathbf{B}_\rho^T \quad \mathbf{H}_\rho^T\,]\, \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \tag{36}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{M}_\rho} = [\,\mathbf{B}_\rho^T \quad \mathbf{H}_\rho^T\,]\, \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \begin{bmatrix} \mathbf{C}_\rho^T \\ \mathbf{0} \end{bmatrix} \tag{37}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{H}_\rho} = [\,\mathbf{0} \quad \mathbf{I}\,]\, \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\rho)) \begin{bmatrix} \mathbf{C}_\rho^T \mathbf{M}_\rho^T \\ \mathbf{J}_\rho^T \end{bmatrix} \tag{38}$$

With these derivatives, $\mu_\rho(\mathbf{w}_\rho)$ can easily be computed using (24).

4 Optimal FWL controller realization

Since the stability related measure $\mu_\rho(\mathbf{w}_\rho)$ is a function of the controller realization $\mathbf{w}_\rho$, we can search for an "optimal" realization that maximizes $\mu_\rho(\mathbf{w}_\rho)$. Such a realization is optimal in the sense that it has a maximum closed-loop stability robustness to the FWL effects. Given an initial design $(\mathbf{F}_{\rho 0}, \mathbf{G}_{\rho 0}, \mathbf{J}_{\rho 0}, \mathbf{M}_{\rho 0}, \mathbf{H}_{\rho 0})$, any realization $(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho)$ can be characterized with (17). Thus, the optimal controller realization $\mathbf{w}_{\rho\,\mathrm{opt}}$ is the solution of the optimization problem:

$$\upsilon_\rho = \max_{\mathbf{w}_\rho \in \mathcal{S}_\rho} \mu_\rho(\mathbf{w}_\rho) \tag{39}$$

We now derive the detailed optimization procedure. $\forall i \in \{1, \cdots, m+n\}$, we partition the eigenvectors of $\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})$, $\mathbf{x}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))$ and $\mathbf{y}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))$, into:

$$\mathbf{x}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{x}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{x}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \quad \mathbf{y}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{y}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{y}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix} \tag{40}$$

where $\mathbf{x}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})), \mathbf{y}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^n$ and $\mathbf{x}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})), \mathbf{y}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^m$. It is easy to see from (19) that, $\forall i \in \{1, \cdots, m+n\}$,

$$\mathbf{x}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho})) = \begin{bmatrix} \mathbf{x}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{T}_\rho^{-1}\mathbf{x}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \quad \mathbf{y}_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho})) = \begin{bmatrix} \mathbf{y}_{i,1}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{T}_\rho^T\mathbf{y}_{i,2}(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix} \tag{41}$$

where $\mathbf{T}_\rho \in \mathcal{R}^{m\times m}$ and $\det(\mathbf{T}_\rho) \neq 0$. Applying (41) to (34)–(38) results in

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{F}_\rho} = \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{x}_{i,2}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T} \tag{42}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{G}_\rho} = \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{x}_{i,1}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^{T} \tag{43}$$

$$\begin{aligned}\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{J}_\rho} &= \left(\mathbf{B}_\rho^T \mathbf{y}_{i,1}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_\rho^T \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,2}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T} \\ &= \left(\mathbf{B}_\rho^T \mathbf{y}_{i,1}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,2}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T}\end{aligned} \tag{44}$$

$$\begin{aligned}\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{M}_\rho} &= \left(\mathbf{B}_\rho^T \mathbf{y}_{i,1}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_\rho^T \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,1}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^{T} \\ &= \left(\mathbf{B}_\rho^T \mathbf{y}_{i,1}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,1}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^{T}\end{aligned} \tag{45}$$

$$\begin{aligned}\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{H}_\rho} &= \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\left(\mathbf{x}_{i,1}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^T\mathbf{M}_\rho^T + \mathbf{x}_{i,2}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T}J_\rho^T\right) \\ &= \mathbf{T}_\rho^T \mathbf{y}_{i,2}^*(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\left(\mathbf{x}_{i,1}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^T\mathbf{M}_{\rho 0}^T + \mathbf{x}_{i,2}^T(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{J}_{\rho 0}^T\right)\end{aligned} \tag{46}$$

Define the following function of the similarity matrix $\mathbf{T}_\rho$:

$$f_\rho(\mathbf{T}_\rho) \stackrel{\triangle}{=} \mu_\rho(\mathbf{w}_\rho) = \min_{i\in\{1,\cdots,m+n\}} \frac{StMa(\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{\rho 0})))}{\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{F}_\rho}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{G}_\rho}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{J}_\rho}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{M}_\rho}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\rho))}{\partial \mathbf{H}_\rho}\|_S} \tag{47}$$

Then the problem (39) of finding an optimal controller realization $\mathbf{w}_{\rho\,\mathrm{opt}}$ is equivalent to obtaining an optimal similarity matrix that is the solution of the following nonlinear optimisation problem:

$$\mathbf{T}_{\rho\,\mathrm{opt}} = \arg \max_{\substack{\mathbf{T}_\rho\in\mathcal{R}^{m\times m} \\ \det(\mathbf{T}_\rho)\neq 0}} f_\rho(\mathbf{T}_\rho) \tag{48}$$

To find a $\mathbf{T}_{\rho\,\mathrm{opt}}$, we will adopt an iterative optimization procedure to generate a sequence $\{\mathbf{T}_{\rho\,0}, \mathbf{T}_{\rho\,1}, \cdots\}$, which converges to $\mathbf{T}_{\rho\,\mathrm{opt}}$. Define $\Omega \triangleq \{\mathbf{T}_\rho \in \mathcal{R}^{m\times m} : \ \det(\mathbf{T}_\rho) = 0\}$. As Ω is only a manifold in $\mathcal{R}^{m\times m}$, starting from a $\mathbf{T}_{\rho\,0} \notin \Omega$, it is rare for an iterative sequence $\{\mathbf{T}_{\rho\,i}\}$ to move into Ω. Thus, in the iterative procedure, the constraint $\det(\mathbf{T}_\rho) \neq 0$ can practically be ignored, leading to an "unconstrained" optimization problem:

$$\max_{\mathbf{T}_\rho \in \mathcal{R}^{m\times m}} f_\rho(\mathbf{T}_\rho) \tag{49}$$

The possible pitfall of violating the constraint can readily be avoided by monitoring the singular values of $\mathbf{T}_\rho$. If a singular value of $\mathbf{T}_\rho$ is too small, a small perturbation $\eta\mathbf{I}$ is added to $\mathbf{T}_\rho$ so that $\mathbf{T}_\rho + \eta\mathbf{I} \notin \Omega$. This small perturbation, which is rarely needed, will not affect the convergence of the iterative procedure. Because $f_\rho(\mathbf{T}_\rho)$ is non-smooth and non-convex, optimization must be based on a direct search without the aid of cost function derivatives. The conventional optimization methods for this kind of problem, such as Rosenbrock and Simplex algorithms [7],[8], generally can only find a local minimum. We will adopt an efficient global optimization strategy based on the ASA algorithm [12]–[14] to search for a true global optimum $\mathbf{T}_{\rho\,\mathrm{opt}}$. With $\mathbf{T}_{\rho\,\mathrm{opt}}$, we can readily obtain the optimal controller realization $\mathbf{w}_{\rho\,\mathrm{opt}}$. The detailed implementation of the ASA algorithm is given in [14].

5 Comparison between z and δ realizations

The z-operator controller realization $\mathbf{w}_z$ is completely equivalent to the δ-operator realization $\mathbf{w}_\delta$ under infinite-precision implementation. We analyze the underlying relationship between these two parameterizations of the controller structure and investigate their FWL implementation characteristics. We will assume that h in the δ operator has an exact FWL representation, e.g. $h = 2^2$, $h = 2^{-6}$. Thus, the source of FWL errors comes solely from the FWL implementation of $\mathbf{w}_\delta$. Define a map g_h from $\mathcal{S}_z$ to $\mathcal{S}_\delta$:

$$\mathbf{w}_\delta = g_h(\mathbf{w}_z) \Leftrightarrow \begin{cases} \mathbf{F}_\delta = \frac{\mathbf{F}_z - \mathbf{I}}{h} \\ \mathbf{G}_\delta = \frac{\mathbf{G}_z}{h} \\ \mathbf{J}_\delta = \mathbf{J}_z \\ \mathbf{M}_\delta = \mathbf{M}_z \\ \mathbf{H}_\delta = \frac{\mathbf{H}_z}{h} \end{cases} \tag{50}$$

We can see that g_h is a one-to-one map.

Lemma 4 $\mu_\delta(g_h(\mathbf{w}_z)) \geq \mu_z(\mathbf{w}_z)$ when $h < 1$; $\mu_\delta(g_h(\mathbf{w}_z)) = \mu_z(\mathbf{w}_z)$ when $h = 1$; $\mu_\delta(g_h(\mathbf{w}_z)) \leq \mu_z(\mathbf{w}_z)$ when $h > 1$.

Proof: For $\mathbf{w}_\delta = g_h(\mathbf{w}_z)$, it follows from Lemma 2 and Remark 1 that

$$\mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_\delta))\mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_\delta)) = \mathbf{y}_i^*(\bar{\mathbf{A}}(\mathbf{w}_z))\mathbf{x}_i^T(\bar{\mathbf{A}}(\mathbf{w}_z)) \tag{51}$$

Noting (50) and, for the plant, $\mathbf{A}_\delta = \frac{\mathbf{A}_z - \mathbf{I}}{h}$, $\mathbf{B}_\delta = \frac{\mathbf{B}_z}{h}$ and $\mathbf{C}_\delta = \mathbf{C}_z$, it then follows from (34)–(38) that

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{F}_\delta} = \frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{F}_z} \tag{52}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{G}_\delta} = \frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{G}_z} \tag{53}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{J}_\delta} = \frac{1}{h}\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{J}_z} \tag{54}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{M}_\delta} = \frac{1}{h}\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{M}_z} \tag{55}$$

$$\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{H}_\delta} = \frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{H}_z} \tag{56}$$

Hence, $\forall i \in \{1, \cdots, m+n\}$,

$$\begin{aligned}
&\frac{\frac{1}{h} - \left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta)) + \frac{1}{h}\right|}{\sum_{j=1}^N \left|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial w_{\delta j}}\right|} = \\
&\frac{\frac{1}{h} - \left|\frac{\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{h}\right|}{\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{F}_\delta}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{G}_\delta}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{J}_\delta}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{M}_\delta}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial \mathbf{H}_\delta}\|_S} = \\
&\frac{1 - \left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))\right|}{h\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{F}_z}\|_S + h\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{G}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{J}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{M}_z}\|_S + h\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{H}_z}\|_S}
\end{aligned} \tag{57}$$

On the other hand, $\forall i \in \{1, \cdots, m+n\}$,

$$\begin{aligned}
&\frac{1 - \left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))\right|}{\sum_{j=1}^N \left|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial w_{zj}}\right|} = \\
&\frac{1 - \left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))\right|}{\|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{F}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{G}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{J}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{M}_z}\|_S + \|\frac{\partial \lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial \mathbf{H}_z}\|_S}
\end{aligned} \tag{58}$$

When $h<1$, comparing (57) with (58) leads to

$$\frac{\frac{1}{h}-\left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))+\frac{1}{h}\right|}{\sum_{j=1}^{N}\left|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_\delta))}{\partial w_{\delta j}}\right|}\geq\frac{1-\left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))\right|}{\sum_{j=1}^{N}\left|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_z))}{\partial w_{zj}}\right|},\ \forall i\in\{1,\cdots,m+n\} \tag{59}$$

which means that $\mu_\delta(g_h(\mathbf{w}_z))\geq\mu_z(\mathbf{w}_z)$. The results for $h=1$ and $h>1$ can similarly be proved.

For $\upsilon_z=\mu_z(\mathbf{w}_{z\,\mathrm{opt}})$ and $\upsilon_\delta=\mu_\delta(\mathbf{w}_{\delta\,\mathrm{opt}})$, based on lemma 4, we have:

Corollary 1 $\upsilon_\delta\geq\upsilon_z$ when $h<1$; $\upsilon_\delta=\upsilon_z$ when $h=1$; $\upsilon_\delta\leq\upsilon_z$ when $h>1$.

Corollary1 shows that, if h is chosen to be smaller than 1, the optimal δ realization has better FWL stability characteristics than the optimal z realization; if h is chosen to be larger than 1, the optimal δ realization has worse FWL stability characteristics than the optimal z realization; if h is chosen to be equal to 1, the both optimal realizations have the same FWL stability robustness to the FWL effects. We notice that δ realizations are dependent of h while z realizations are independent of h. Thus, υ_δ is a function of h, which will be denoted as $\upsilon_\delta(h)$, while υ_z is not. Let us introduce the function:

$$f(h)=\min_{i\in\{1,\cdots,m+n\}}\frac{\kappa_i}{h\alpha_i+\beta_i} \tag{60}$$

where

$$\kappa_i=1-\left|\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))\right| \tag{61}$$

$$\alpha_i=\left\|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))}{\partial\mathbf{F}_z}\right\|_S+\left\|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))}{\partial\mathbf{G}_z}\right\|_S+\left\|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))}{\partial\mathbf{H}_z}\right\|_S \tag{62}$$

and

$$\beta_i=\left\|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))}{\partial\mathbf{J}_z}\right\|_S+\left\|\frac{\partial\lambda_i(\bar{\mathbf{A}}(\mathbf{w}_{z\,\mathrm{opt}}))}{\partial\mathbf{M}_z}\right\|_S \tag{63}$$

Theorem 2 $\upsilon_z=f(1)$ and $\upsilon_\delta(h)\geq f(h)$.

Proof: $\upsilon_z=f(1)$ can directly be obtained from the definitions of $\mu_z(\mathbf{w}_{z\,\mathrm{opt}})$ and $f(h)$. From the proof of lemma 4, it can easily be seen that $\mu_\delta(g_h(\mathbf{w}_{z\,\mathrm{opt}}))=f(h)$. Noting $\upsilon_\delta(h)=\max_{\mathbf{w}_\delta\in\mathcal{S}_\delta}\mu_\delta(\mathbf{w}_\delta)\geq\mu_\delta(g_h(\mathbf{w}_{z\,\mathrm{opt}}))$, we conclude that $\upsilon_\delta(h)\geq f(h)$.

Notice that $f(h)$ is defined in $(0, \infty)$ and $f(h)$ decreases as h increases. According to theorem 2, for $h \in (0,\ 1)$, the optimal δ realization has better FWL closed-loop stability performance than the optimal z realization and, furthermore, the smaller h is, the larger $\upsilon_\delta(h)$ is than υ_z. It is well-known that, when $h \to 0$, the δ-operator representation approaches the continuous-time representation. It is therefore expected that $f(h)$ and hence $\upsilon_\delta(h)$ will approach certain limit values as $h \to 0$.

6 A design example

We present a numerical example to illustrate the proposed optimization approach and verify the theoretical results given in the previous section. The plant model used is a modification of the plant studied in [2], which was a single-input single-output system. We have added one more output that is the first state in the original plant model. The state-space model of this modified plant, represented in the z operator, is given by

$$\mathbf{A}_z = \begin{bmatrix} 3.2439e-01 & -4.5451e+00 & -4.0535e+00 & -2.7003e-03 & 0 \\ 1.4518e-01 & 4.9477e-01 & -4.6945e-01 & -3.1274e-04 & 0 \\ 1.6814e-02 & 1.6491e-01 & 9.6681e-01 & -2.2114e-05 & 0 \\ 1.1889e-03 & 1.8209e-02 & 1.9829e-01 & 1.0000e+00 & 0 \\ 6.1301e-05 & 1.2609e-03 & 1.9930e-02 & 2.0000e-01 & 1 \end{bmatrix}$$

$$\mathbf{B}_z = \begin{bmatrix} 1.4518e-01 \\ 1.6814e-02 \\ 1.1889e-03 \\ 6.1301e-05 \\ 2.4979e-06 \end{bmatrix}$$

$$\mathbf{C}_z = \begin{bmatrix} 0 & 0 & 1.6188e+00 & -1.5750e-01 & -4.3943e+01 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The closed-loop poles as given in [2] were used in design, and the designed controller obtained using a standard design procedure [18] had a state-space form:

$$\mathbf{F}_{z0} = \begin{bmatrix} 0 & 1 \\ -9.3303e-01 & 1.9319e+00 \end{bmatrix}, \quad \mathbf{G}_{z0} = \begin{bmatrix} 4.1814e-02 & 2.7132e+02 \\ 3.9090e-02 & 1.0167e+03 \end{bmatrix}$$

$$\mathbf{J}_{z0} = \begin{bmatrix} 3.0000e-04 & 5.0000e-04 \end{bmatrix}, \quad \mathbf{M}_{z0} = \begin{bmatrix} 0 & 6.1250e-01 \end{bmatrix}, \quad \mathbf{H}_{z0} = \begin{bmatrix} 7.8047e+01 \\ 7.3849e+01 \end{bmatrix}$$

With this initial realization $\mathbf{w}_{z0}$, the corresponding transition matrix $\bar{\mathbf{A}}(\mathbf{w}_{z0})$ was formed using (19), from which the poles and the eigenvectors of the ideal closed-loop system were computed. The value of the stability related measure for $\mathbf{w}_{z0}$ is $\mu_z(\mathbf{w}_{z0}) = 4.0509e-07$.

Using the ASA algorithm to solve for the resulting optimization problem (48) gave rise to the following optimal similarity transformation matrix:

$$\mathbf{T}_{z\,\mathrm{opt}} = \begin{bmatrix} -1.7791e+01 & 3.5665e+00 \\ -1.6696e+01 & 3.5384e+00 \end{bmatrix}$$

The optimal z realization corresponding to $\mathbf{T}_{z\,\mathrm{opt}}$ was

$$\mathbf{F}_{z\,\mathrm{opt}} = \begin{bmatrix} 9.5253e-01 & -2.5578e-03 \\ 7.0338e-02 & 9.7934e-01 \end{bmatrix}, \quad \mathbf{G}_{z\,\mathrm{opt}} = \begin{bmatrix} -2.5073e-03 & 7.8274e+02 \\ -7.8313e-04 & 3.9806e+03 \end{bmatrix}$$

$$\mathbf{J}_{z\,\mathrm{opt}} = \begin{bmatrix} -1.3685e-02 & 2.8392e-03 \end{bmatrix}, \quad \mathbf{M}_{z\,\mathrm{opt}} = \begin{bmatrix} 0 & 6.1250e-01 \end{bmatrix}$$

$$\mathbf{H}_{z\,\mathrm{opt}} = \begin{bmatrix} -3.7504e+00 \\ 3.1750e+00 \end{bmatrix}$$

The optimal stability related measure was $\upsilon_z = \mu_z(\mathbf{w}_{z\,\mathrm{opt}}) = 3.8927e-06$. This represents an improvement, approximately a factor of ten, over the initial controller realization.

Similarly, we constructed and solved for the optimal δ realization problem for $h = 2^3 \sim 2^{-10}$. Figure 2 compares $\upsilon_\delta(h)$, the stability related measure for the optimal δ realization, with $f(h)$ and υ_z. It can be seen that the results of Figure 2 agree with the theoretical analysis of corollary 1 and theorem 2. As expected, for $h < 1$, the optimal δ realization has larger FWL closed-loop stability measure than the optimal z realization.

We also computed the unit impulse response of the closed-loop control system when the controllers were the infinite-precision implemented $\mathbf{w}_{z\,0}$ and various FWL implemented realizations with 10-bit accuracy for B_f, respectively. Note that any realization $\mathbf{w}_\rho \in \mathcal{S}_\rho$, implemented in infinite precision, will achieve the exact performance of the infinite-precision implemented $\mathbf{w}_{z0}$, which is the *designed* controller performance. For this reason, the infinite-precision implemented $\mathbf{w}_{z0}$ is referred to as the *ideal* controller realization $\mathbf{w}_{\mathrm{ideal}}$. Figures 3 to 6 compares the unit impulse response of the first plant output $y_1(k)$ for the ideal controller $\mathbf{w}_{\mathrm{ideal}}$ with those of various 10-bit implemented realizations. It can be seen that the closed-loop became unstable with a 10-bit implemented controller realization $\mathbf{w}_{z\,0}$. The results also clearly show the benefits of the proposed optimization process, as the closed-loop system remained stable with the 10-bit implemented $\mathbf{w}_{z\,\mathrm{opt}}$. Furthermore, the 10-bit implemented $\mathbf{w}_{\delta\,\mathrm{opt}}$ with $h = 2^{-1}$ was able to approximate closely the designed performance of the ideal infinite-precision controller. With h reduced to 2^{-7}, the 10-bit implemented $\mathbf{w}_{\delta\,\mathrm{opt}}$ achieved the designed controller performance.

7 Conclusions

We have studied the finite-precision implementation issues for digital controllers. A unified approach has been adopted to derive a tractable FWL closed-loop stability related measure for both the z and δ operator parameterizations of the general controller structure. An efficient optimization procedure has been developed for obtaining the optimal controller realization that maximizes the proposed measure. The underlying relationship connecting the z and δ realizations has been investigated. Because the FWL stability measure for δ controller realization is a function of the operator constant h, we can always obtain an optimal δ controller realization that has better closed-loop stability margin than the optimal z realization in FWL implementation. The theoretical results have been verified and the optimization procedure demonstrated using a numerical design example.

References

[1] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. London: Springer Verlag, 1993.

[2] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automatic Control*, Vol.43, pp.689–693, 1998.

[3] R.H. Istepanian, G. Li, J. Wu and J. Chu, "Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds," *IEE Proc. Control Theory and Applications*, Vol.145, No.5, pp.472–478, 1998.

[4] S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," *IEEE Trans. Automatic Control*, Vol.44, No.11, pp.2149–2153, 1999.

[5] I.J. Fialho and T.T. Georgiou, "On stability and performance of sampled data systems subject to word length constraint," *IEEE Trans. Automatic Control*, Vol.39, No.12, pp.2476–2481, 1994.

[6] I.J. Fialho and T.T. Georgiou, "Optimal finite worldlength digital controller realization," in *Proc. American Control Conf.* (San Diego, USA), June 2-4, 1999, pp.4326-4327.

[7] G.S.G. Beveridge and R.S. Schechter, *Optimization: Theory and Practice.* McGraw-Hill, 1970.

[8] L.C.W. Dixon, *Nonlinear Optimisation.* London: The English Universities Press Ltd, 1972.

[9] D.E. Goldberg, *Genetic Algorithms in Search, Optimisation and Machine Learning.* Addison Welsey, 1989.

[10] K.F. Man, K.S. Tang and S. Kwong, *Genetic Algorithms: Concepts and Design.* Springer-Verlag: London, 1998.

[11] C.M. Fonseca and P.J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms – Part I: A unified formulation," *IEEE Trans. Systems, Man, and Cybernetics Part A: Systems and Humans*, Vol.28, No.1, pp.26–37, 1998.

[12] L. Ingber, "Simulated annealing: practice versus theory," *Mathematical and Computer Modelling*, Vol.18, No.11, pp.29–57, 1993.

[13] L. Ingber, "Adaptive simulated annealing (ASA): lessons learned," *J. Control and Cybernetics*, Vol.25, No.1, pp.33-54, 1996.

[14] S. Chen and B.L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, Vol.79, No.11, pp.117-128, 1999.

[15] S. Boyd, L. EI Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory.* SIAM, 1994.

[16] R.E. Skelton, T. Iwasaki and K.M. Grigoriadid, *A Unified Algebraic Approach to Linear Control Design.* London: Taylor and Francis, 1998.

[17] T. Kailath, *Linear Systems.* Prentice-Hall, 1980.

[18] J. O'Reilly, *Observers for Linear Systems.* London: Academic Press, 1983.

[19] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits and Systems*, Vol.37, pp.1487–1498, 1990.

[20] J. Wu, S. Chen, G. Li and J. Chu, "Optimal finite-precision state-estimate feedback controller realization of discrete-time systems," *IEEE Trans. Automatic Control*, to appear, July 2000.

[21] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach.* Englewood Cliffs, NJ: Prentice Hall, 1990.

[22] R.M. Goodall and B.J. Donoghue, "Very high sample rate digital filters using the δ operator," *IEE Proc. G*, Vol.140, pp.199–206, 1993.

[23] Y. Sung and M. Kung, "Lower finite word-length effect on state space digital filter by δ operator realization," *Int. J. Electronics*, Vol.75, No.6, pp.1135–1141, 1993.

[24] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *IEEE Trans. Automatic Control*, Vol.38, No.5, pp.803–807, 1993.

[25] R.H. Istepanian, "Implementational issues for discrete PID algorithms using shift and delta operators parameterizations," in *Proc. 4th IFAC Workshop on Algorithms and Architectures for Real-Time Control* (Vilamoura, Portugal), 1997, pp.117–122.

[26] S. Chen, J. Wu, R.H. Istepanian, J. Chu and J.F. Whidborne, "Optimizing stability bounds of finite-precision controller structures for sampled-data systems in the delta operator domain," *IEE Proc. Control Theory and Applications*, Vol.146, No.6, pp.517–526, 1999.

[27] C.P. Neuman, "Transformations between delta and forward shift operator transfer function models," *IEEE Trans. System Man and Cybernetics*, Vol.23, pp.295–296, 1993.

[28] C.P. Neuman, "Properties of the delta operator model of dynamic physical systems," *IEEE Trans. System Man and Cybernetics*, Vol.23, pp.296–301, 1993.

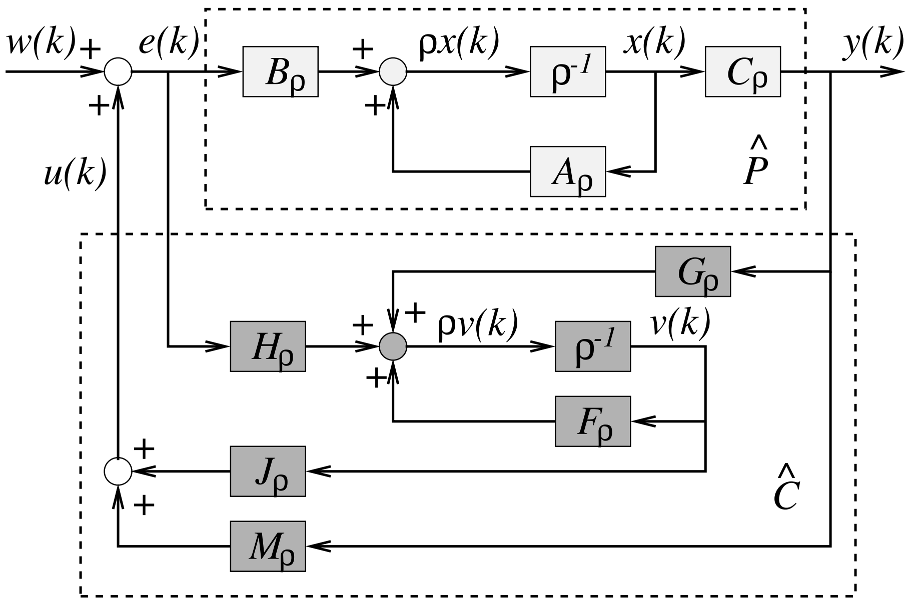

Figure 1: Discrete-time closed-loop system with a generic controller.

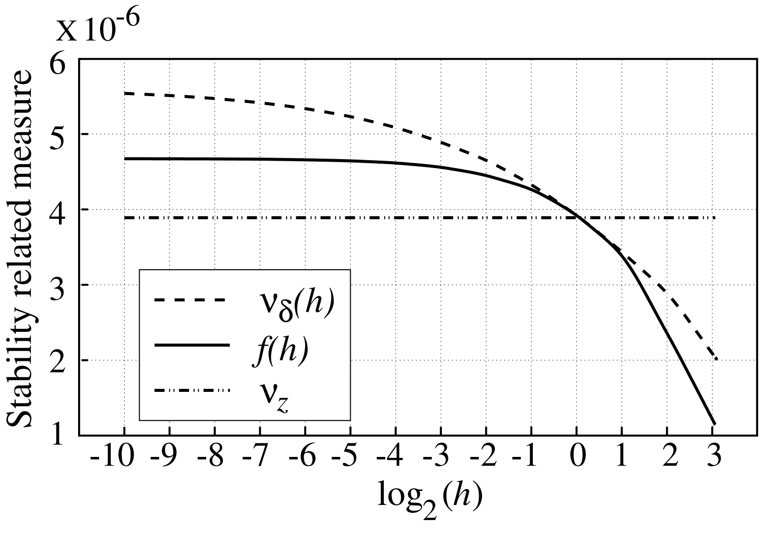

Figure 2: Comparison of the values of the stability related measure for the optimal δ realization $\mathbf{w}_{\delta\,\mathrm{opt}}$, the δ realization $\mathbf{w}_\delta = g_h(\mathbf{w}_{z\,\mathrm{opt}})$ and the optimal z realization $\mathbf{w}_{z\,\mathrm{opt}}$.

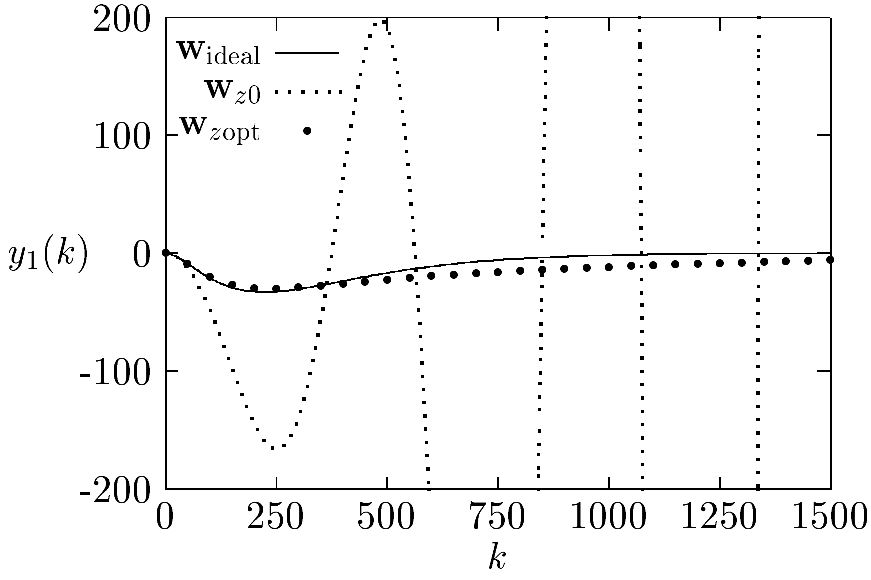

Figure 3: Comparison of unit impulse response for the ideal infinite-precision controller implementation $\mathbf{w}_{\text{ideal}}$ with those for the two 10-bit implemented controller realizations $\mathbf{w}_{z0}$ and $\mathbf{w}_{z\text{opt}}$.

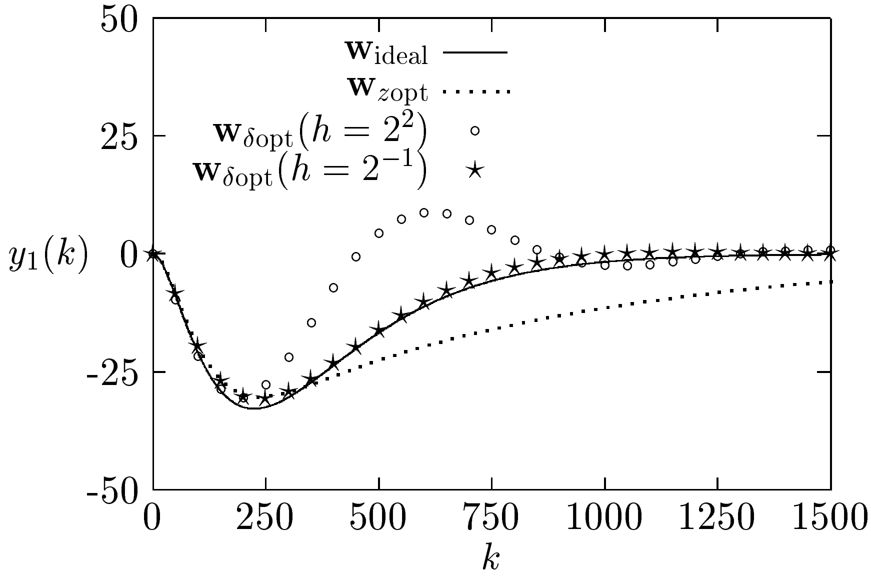

Figure 4: Comparison of unit impulse response for the ideal infinite-precision controller implementation $\mathbf{w}_{\text{ideal}}$ with those for the three 10-bit implemented controller realizations $\mathbf{w}_{z\text{opt}}$, $\mathbf{w}_{\delta\text{opt}}$ $(h = 2^2)$ and $\mathbf{w}_{\delta\text{opt}}$ $(h = 2^{-1})$.

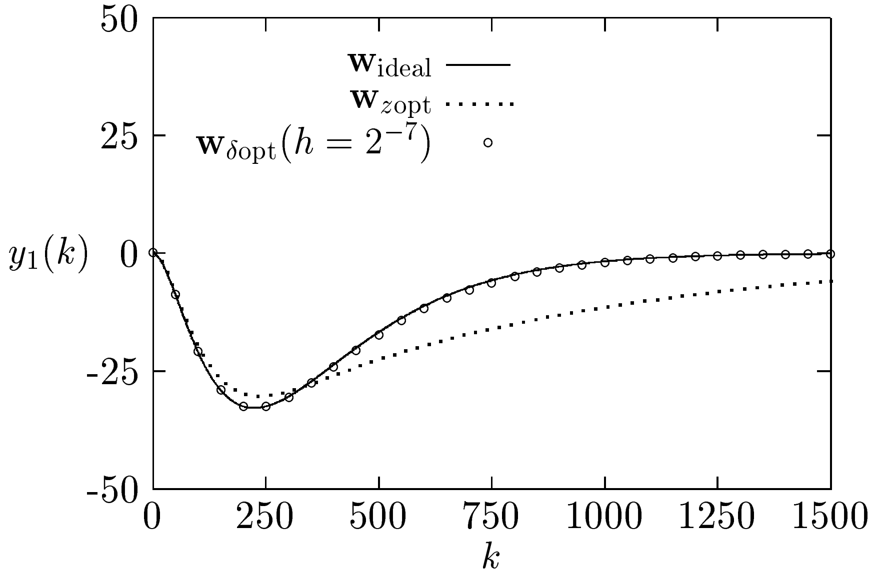

Figure 5: Comparison of unit impulse response for the ideal infinite-precision controller implementation $\mathbf{w}_{\text{ideal}}$ with those for the two 10-bit implemented controller realizations $\mathbf{w}_{z\text{opt}}$ and $\mathbf{w}_{\delta\text{opt}}$ $(h = 2^{-7})$.

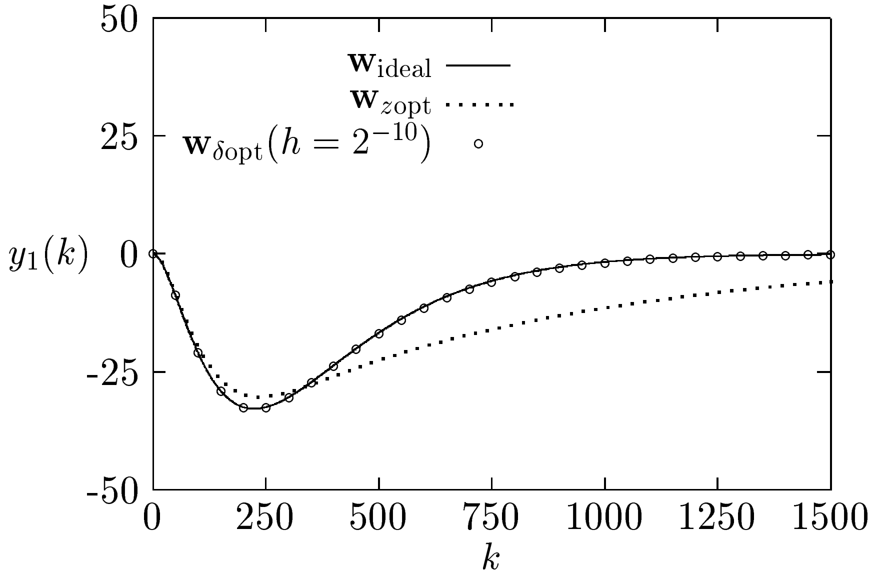

Figure 6: Comparison of unit impulse response for the ideal infinite-precision controller implementation $\mathbf{w}_{\text{ideal}}$ with those for the two 10-bit implemented controller realizations $\mathbf{w}_{z\text{opt}}$ and $\mathbf{w}_{\delta\text{opt}}$ $(h = 2^{-10})$.