Proceedings of the 45th IEEE Conference on Decision & Control
Manchester Grand Hyatt Hotel
San Diego, CA, USA, December 13-15, 2006

Optimal Controller Realisations with the Smallest Dynamic Range

Jun Wu, Sheng Chen, Gang Li and Jian Chu

Abstract— **An approach is proposed to design optimal finite word length (FWL) realisations of digital controllers implemented in fixed-point arithmetic. A minimax-based search procedure is first used to obtain an optimal controller realisation that optimises an FWL closed-loop stability measure. Since this FWL closed-loop stability measure is solely linked to the fractional part or precision of fixed-point format, the resulting realisation may not have the smallest dynamic range. A measure is derived to indicate the dynamic range of a realisation. Based on an orthogonal transformation of this dynamic range measure for the optimal precision controller realisation, a numerical optimisation method is developed to make the controller realisation having the smallest dynamic range without sacrificing FWL closed-loop stability robustness.**

I. Introduction

The detrimental finite word length (FWL) effect on a fixed-point implemented digital controller is particularly serious due to a reduced precision. Many works have focused on the FWL effect on closed-loop stability and have considered the design of optimal fixed-point digital controller realisations that optimise various FWL closed-loop stability measures [1]-[13]. We point out a limitation of these previous approaches in designing optimal fixed-point controller realisations. The FWL closed-loop stability measures used in these works are only linked to the fractional part of fixed-point representation. Optimising these measures, while minimising the bits required for the fractional part, may actually increase the integer part or dynamic range of fixed-point representation. Thus, the resulting "optimal" controller realisations are not necessarily true optimal ones in terms of the robustness to the FWL effects.

In a fixed-point implementation, the total available bits have to accommodate the dynamic range first to avoid overflow, and the remaining bits left are then used to implement the fractional part. Therefore, a better approach is to consider both a precision or FWL closed-loop stability measure and a dynamic range measure together. We adopt a "two-step" approach to tackle this multi-objective task. Firstly, we optimise the FWL closed-loop stability measure of [4] to obtain an optimal realisation using the efficient search algorithm of [13]. Secondly, we optimise a dynamic range measure for this optimal realisation. The value of the FWL closed-loop stability measure is invariant under orthogonal transformation of controller realisation [2]. We exploit this extra freedom of realisation to minimise the dynamic range of the controller realisation. The final realisation obtained in our two-step design has both the maximum FWL stability robustness and the smallest dynamic range, and thus it is a true optimal realisation for fixed-point implementation. The proposed approach is established in a unified framework for both the shift and delta operators to enable a comparison for the FWL closed-loop stability characteristics of the optimal controller realisations using these two operators.

J. Wu and J. Chu are with National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, P. R. China

S. Chen is with School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, U.K. sqc@ecs.soton.ac.uk

G. Li is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

S. Chen wish to thank the support of the United Kingdom Royal Academy of Engineering.

II. Notations and the problem formulation

Let $\mathbf{e}_i$ be the ith real coordinate vector, and for any $\mathbf{z} \in \mathcal{C}^n$ define $\Upsilon(\mathbf{z}) \triangleq [\Re(\mathbf{z})\ \Im(\mathbf{z})]$, where $\Re(\mathbf{z})$ and $\Im(\mathbf{z})$ denote the real and the imaginary parts of $\mathbf{z}$, respectively. For a complex-valued matrix $\mathbf{U} \in \mathcal{C}^{m\times n}$ with elements u_{ij}, we define the following matrix norms

$$\|\mathbf{U}\|_M \triangleq \max_{\substack{i\in\{1,\cdots,m\}\\ j\in\{1,\cdots,n\}}} |u_{ij}|, \quad \|\mathbf{U}\|_F \triangleq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|u_{ij}|^2}\ . \quad (1)$$

Let $\mathrm{Vec}(\cdot)$ be the column stacking operator such that $\mathrm{Vec}(\mathbf{U})$ is an mn-dimensional vector. For a real-valued square matrix $\mathbf{M} \in \mathcal{R}^{n\times n}$, let $\{\lambda_i(\mathbf{M}),\ 1 \le i \le n\}$ denote its eigenvalues, and let $\mathbf{x}_i(\mathbf{M})$ be the right eigenvector corresponding to $\lambda_i(\mathbf{M})$. Define $\mathbf{M}_x \triangleq [\mathbf{x}_1(\mathbf{M})\ \mathbf{x}_2(\mathbf{M})\cdots\mathbf{x}_n(\mathbf{M})]$ and $\mathbf{M}_y = [\mathbf{y}_1(\mathbf{M})\ \mathbf{y}_2(\mathbf{M})\cdots\mathbf{y}_n(\mathbf{M})] \triangleq \mathbf{M}_x^{-H}$, where $\mathbf{y}_i(\mathbf{M})$ is called the reciprocal left eigenvector related to $\mathbf{x}_i(\mathbf{M})$.

A discrete-time linear system can be described using either the shift operator z or delta operator δ. The latter is defined as $\delta \triangleq (z-1)/h$, where h is a positive real constant [14],[2]. In this paper, it is assumed that the value of h in the δ operator has an exact fixed-point representation (e.g. $h = 2^2$ or $h = 2^{-6}$) so that the source of FWL errors comes solely from a finite-precision implementation of the controller realisation. For the notational conciseness, we introduce a "generalised" operator ρ for the discrete-time system. The state-space description of the general discrete-time system using the operator ρ is

$$\begin{cases} \rho\mathbf{x}_g(k) = \mathbf{F}_{g,\rho}\mathbf{x}_g(k) + \mathbf{G}_{g,\rho}\mathbf{u}_{g,1}(k) + \mathbf{H}_{g,\rho}\mathbf{u}_{g,2}(k) \\ \mathbf{y}_g(k) = \mathbf{J}_{g,\rho}\mathbf{x}_g(k) + \mathbf{M}_{g,\rho}\mathbf{u}_{g,1}(k) \end{cases} \quad (2)$$

where all the matrices and vectors are real-valued with appropriate dimensions. Obviously, $\rho = z$ and $\rho = \delta$ give rise to the two equivalent representations of the same system,

with the following relationship

$$\begin{array}{l} \mathbf{F}_{g,\delta} = \frac{\mathbf{F}_{g,z}-\mathbf{I}}{h},\ \mathbf{G}_{g,\delta} = \frac{\mathbf{G}_{g,z}}{h},\ \mathbf{J}_{g,\delta} = \mathbf{J}_{g,z}, \\ \mathbf{M}_{g,\delta} = \mathbf{M}_{g,z},\ \mathbf{H}_{g,\delta} = \frac{\mathbf{H}_{g,z}}{h}, \end{array} \tag{3}$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension. With a proper index order, $\{\lambda_i(\mathbf{F}_{g,z})\}$ and $\{\lambda_i(\mathbf{F}_{g,\delta})\}$ can be one-to-one mapped with $\lambda_i(\mathbf{F}_{g,z}) = 1 + h\lambda_i(\mathbf{F}_{g,\delta})$, $\forall i$. Since the system $(\mathbf{F}_{g,z}, \mathbf{G}_{g,z}, \mathbf{J}_{g,z}, \mathbf{M}_{g,z}, \mathbf{H}_{g,z})$ is stable if and only if $|\lambda_i(\mathbf{F}_{g,z})| < 1$, $\forall i$, we have the stability condition for the same system described using δ operator.

Theorem 1: The discrete-time system $(\mathbf{F}_{g,\delta}, \mathbf{G}_{g,\delta}, \mathbf{J}_{g,\delta}, \mathbf{M}_{g,\delta}, \mathbf{H}_{g,\delta})$ is stable if and only if

$$\left|\lambda_i(\mathbf{F}_{g,\delta}) + \frac{1}{h}\right| < \frac{1}{h}, \quad \forall i. \tag{4}$$

Consider the discrete-time closed-loop control system, consisting of the linear time-invariant plant $\hat{P}$ and the generic digital stabilising controller $\hat{C}$. $\hat{P}$ is completely state controllable and observable with the state-space description

$$\begin{cases} \rho\,\mathbf{x}(k) = \mathbf{A}_\rho \mathbf{x}(k) + \mathbf{B}_\rho \mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}_\rho \mathbf{x}(k) \end{cases} \tag{5}$$

where $\mathbf{A}_\rho \in \mathcal{R}^{n\times n}$, $\mathbf{B}_\rho \in \mathcal{R}^{n\times p}$ and $\mathbf{C}_\rho \in \mathcal{R}^{q\times n}$; and $\hat{C}$ is described by the state-space description

$$\begin{cases} \rho\,\mathbf{v}(k) = \mathbf{F}_\rho \mathbf{v}(k) + \mathbf{G}_\rho \mathbf{y}(k) + \mathbf{H}_\rho \mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}_\rho \mathbf{v}(k) + \mathbf{M}_\rho \mathbf{y}(k) \end{cases} \tag{6}$$

with $\mathbf{F}_\rho \in \mathcal{R}^{m\times m}$, $\mathbf{G}_\rho \in \mathcal{R}^{m\times q}$, $\mathbf{J}_\rho \in \mathcal{R}^{p\times m}$, $\mathbf{M}_\rho \in \mathcal{R}^{p\times q}$ and $\mathbf{H}_\rho \in \mathcal{R}^{m\times p}$. $\hat{C}$ is an output feedback controller if $\mathbf{H}_\rho = \mathbf{0}$; a full-order observer-based controller if $\mathbf{F}_\rho = \mathbf{A}_\rho - \mathbf{G}_\rho\mathbf{C}_\rho$, $\mathbf{M}_\rho = \mathbf{0}$ and $\mathbf{H}_\rho = \mathbf{B}_\rho$; a reduced-order observer-based controller, otherwise [15],[16].

The state-space descriptions or realisations $(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho)$ of the controller $\hat{C}$ are not unique. Let $(\mathbf{F}_{\rho 0}, \mathbf{G}_{\rho 0}, \mathbf{J}_{\rho 0}, \mathbf{M}_{\rho 0}, \mathbf{H}_{\rho 0})$ be a realisation of $\hat{C}$ given by a standard controller design procedure. Then all the realisations of $\hat{C}$ form a realisation set

$$\begin{array}{rcl} \mathcal{S}_\rho & \triangleq & \{(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho) : \mathbf{F}_\rho = \mathbf{T}_\rho^{-1}\mathbf{F}_{\rho 0}\mathbf{T}_\rho, \\ & & \mathbf{G}_\rho = \mathbf{T}_\rho^{-1}\mathbf{G}_{\rho 0}, \mathbf{J}_\rho = \mathbf{J}_{\rho 0}\mathbf{T}_\rho, \\ & & \mathbf{M}_\rho = \mathbf{M}_{\rho 0}, \mathbf{H}_\rho = \mathbf{T}_\rho^{-1}\mathbf{H}_{\rho 0}\} \end{array} \tag{7}$$

where $\mathbf{T}_\rho \in \mathcal{R}^{m\times m}$ is any real-valued nonsingular matrix. Any two realisations in $\mathcal{S}_\rho$ are completely equivalent if they are implemented with infinite precision. Define

$$\mathbf{w}_\rho \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_\rho) \\ \mathrm{Vec}(\mathbf{G}_\rho) \\ \mathrm{Vec}(\mathbf{J}_\rho) \\ \mathrm{Vec}(\mathbf{M}_\rho) \\ \mathrm{Vec}(\mathbf{H}_\rho) \end{bmatrix}, \quad \mathbf{w}_{\rho 0} \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{G}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{J}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{M}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{H}_{\rho 0}) \end{bmatrix}. \tag{8}$$

The stability of the closed-loop control system depends on the eigenvalues of the transition matrix

$$\begin{array}{rcl} \overline{\mathbf{A}}(\mathbf{w}_\rho) & \triangleq & \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{G}_\rho\mathbf{C}_\rho + \mathbf{H}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho\mathbf{J}_\rho \end{bmatrix} \\ & \triangleq & \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_{\rho 0}) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho \end{bmatrix} \end{array} \tag{9}$$

where $\mathbf{0}$ denotes the zero matrix of appropriate dimension. Define the stability margin of $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))$ as

$$SM(\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))) \triangleq \begin{cases} 1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_z))|, & \text{if } \rho = z, \\ \frac{1}{h} - \left|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\delta)) + \frac{1}{h}\right|, & \text{if } \rho = \delta. \end{cases} \tag{10}$$

It is obvious that all the different controller realisations $\mathbf{w}_\rho \in \mathcal{S}_\rho$ have exactly the same set of the closed-loop eigenvalues if they are implemented with infinite precision.

When $\mathbf{w}_\rho$ is implemented using a fixed-point processor of the bit length b, the b bits are assigned as follows: One bit is used for the sign, b_g bits are used for the integer part of the representation, and the remaining $b_f = b - b_g - 1$ bits are used to implement the fractional part of the representation. In order to avoid overflow in representing $\mathbf{w}_\rho$, b_g should be sufficiently large such that

$$\|\mathbf{w}_\rho\|_M \le 2^{b_g}. \tag{11}$$

$\|\mathbf{w}_\rho\|_M$ represents the dynamic range of $\mathbf{w}_\rho$ in fixed-point format. Even assuming no overflow, $\mathbf{w}_\rho$ is perturbed into $\mathbf{w}_\rho + \mathbf{\Delta}$ due to the finite b_f bits in the fractional part representation. It can easily be shown that each element of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(b_f+1)}$, that is , $\|\mathbf{\Delta}\|_M \le 2^{-(b_f+1)}$. With the perturbation $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho + \mathbf{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{w}_\rho + \mathbf{\Delta})$ crosses over the stability boundary, the closed-loop system, originally designed to be stable, becomes unstable. Under the condition of no overflow, the closed-loop stability depends only on the perturbation $\mathbf{\Delta}$, that is, the precision of the fractional part representation.

Because the total bit length b is divided between the dynamic range and precision of fixed-point format, the design of optimal FWL realisation is a multi-objective optimisation. Firstly, an optimal realisation should optimise some FWL closed-loop stability measure, and the value of such a stability measure only depends on the precision or fractional part of a realisation. Secondly, a desired realisation should also have the smallest dynamic range, since this will require the smallest number of b_g bits to avoid overflow and in turn leaves the most b_f bits to achieve the highest possible precision. We will adopt an effective two-step approach to tackle this multi-objective optimisation problem.

III. Optimising an FWL closed-loop stability measure

We use λ_i to replace $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))$ when doing so does not cause ambiguity, and we adopt the following FWL closed-loop stability measure as defined in [4]

$$f(\mathbf{w}_\rho) \triangleq \max_{i\in\{1,\cdots,m+n\}} \frac{\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F}{SM(\lambda_i)}. \tag{12}$$

The measure $f(\mathbf{w}_\rho)$ describes the robustness of closed-loop stability to the FWL perturbation $\mathbf{\Delta}$ for the realisation $\mathbf{w}_\rho$ under the condition of no overflow. With this measure, the optimal FWL realisation problem is given by

$$\nu \triangleq \min_{\mathbf{w}_\rho \in \mathcal{S}_\rho} f(\mathbf{w}_\rho). \tag{13}$$

Define

$$g(\mathbf{w}_\rho, i) \triangleq \frac{\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F}{SM(\lambda_i)}. \quad (14)$$

Obviously, the optimisation problem (13) can be viewed as

$$\nu = \min_{\mathbf{w}_\rho \in \mathcal{S}_\rho} \max_{i \in \{1,\cdots,m+n\}} g(\mathbf{w}_\rho, i). \quad (15)$$

The following results [17],[18] on saddle points play an important role in obtaining global optimal solutions of minimax-formulation problems.

Definition 1: $(\mathbf{w}'_\rho, i') \in \mathcal{S}_\rho \times \{1, \cdots, m+n\}$ is said to be a *saddle point* of $g(\mathbf{w}_\rho, i)$ if $\forall i \in \{1, \cdots, m+n\}$

$$g(\mathbf{w}'_\rho, i) \le g(\mathbf{w}'_\rho, i') \le g(\mathbf{w}_\rho, i'), \quad \forall \mathbf{w}_\rho \in \mathcal{S}_\rho. \quad (16)$$

We have the well-known Minimax Theorem in game theory.

Theorem 2: If and only if there exists at least a saddle point $(\mathbf{w}'_\rho, i')$ of $g(\mathbf{w}_\rho, i)$, then

$$\begin{aligned} \min_{\mathbf{w}_\rho \in \mathcal{S}_\rho} \max_{i \in \{1,\cdots,m+n\}} g(\mathbf{w}_\rho, i) &= \max_{i \in \{1,\cdots,m+n\}} \min_{\mathbf{w}_\rho \in \mathcal{S}_\rho} g(\mathbf{w}_\rho, i) \\ &= g(\mathbf{w}'_\rho, i'). \end{aligned} \quad (17)$$

Theorem 3: Let

$$\eta_i \triangleq \min_{\mathbf{w}_\rho \in \mathcal{S}_\rho} g(\mathbf{w}_\rho, i) \quad \forall i \in \{1, \cdots, m+n\}, \quad (18)$$

$$i' \triangleq \arg \max_{i \in \{1,\cdots,m+n\}} \eta_i, \quad (19)$$

$$\mathcal{W} \triangleq \{\mathbf{w}_\rho : g(\mathbf{w}_\rho, i') = \eta_{i'}, \mathbf{w}_\rho \in \mathcal{S}_\rho\}. \quad (20)$$

Then $(\mathbf{w}'_\rho, i')$ is a saddle point of $g(\mathbf{w}_\rho, i)$ if and only if $\mathbf{w}'_\rho \in \mathcal{W}$ and

$$g(\mathbf{w}'_\rho, i) \le \eta_{i'}, \ \forall i \in \{1, \cdots, m+n\} \setminus \{i'\}. \quad (21)$$

A. Optimising single-pole FWL stability measure

To attain the single-pole measure η_i defined in (18) for the eigenvalue λ_i is equivalent to solve the minimisation problem of the single-pole sensitivity

$$\min_{\substack{\mathbf{T}_\rho \in \mathcal{R}^{m \times m} \\ \det \mathbf{T}_\rho \neq 0}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2. \quad (22)$$

$\forall i \in \{1, \cdots, m+n\}$, partition the eigenvectors of $\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})$

$$\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \quad (23)$$

$$\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \quad (24)$$

where $\mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^n$, $\mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^n$, $\mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^m$ and $\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^m$. Let

$$\begin{aligned} \alpha_i^2 &\triangleq \|\mathbf{C}_\rho \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2 + \\ &\quad \|\mathbf{M}_{\rho 0} \mathbf{C}_\rho \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \mathbf{J}_{\rho 0} \mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \quad (25) \\ \beta_i^2 &\triangleq \|\mathbf{B}_\rho^T \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \quad (26) \\ \tau_i^2 &\triangleq \|\mathbf{B}_\rho^T \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2 \\ &\quad \times \|\mathbf{C}_\rho \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \quad (27) \\ \mathbf{q}_i &\triangleq \mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})), \quad (28) \\ \mathbf{z}_i &\triangleq \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})). \quad (29) \end{aligned}$$

It can be shown that [13]

$$\begin{aligned} \left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2 &= \left\| \frac{\partial \lambda_i}{\partial \mathbf{F}_\rho} \right\|_F^2 + \left\| \frac{\partial \lambda_i}{\partial \mathbf{G}_\rho} \right\|_F^2 + \left\| \frac{\partial \lambda_i}{\partial \mathbf{J}_\rho} \right\|_F^2 + \\ &\quad \left\| \frac{\partial \lambda_i}{\partial \mathbf{M}_\rho} \right\|_F^2 + \left\| \frac{\partial \lambda_i}{\partial \mathbf{H}_\rho} \right\|_F^2 \\ &= \|\mathbf{T}_\rho^{-1} \mathbf{q}_i\|_F^2 \|\mathbf{T}_\rho^T \mathbf{z}_i\|_F^2 + \alpha_i^2 \|\mathbf{T}_\rho^T \mathbf{z}_i\|_F^2 + \\ &\quad \beta_i^2 \|\mathbf{T}_\rho^{-1} \mathbf{q}_i\|_F^2 + \tau_i^2. \end{aligned} \quad (30)$$

For the different cases of $\mathbf{q}_i$ and $\mathbf{z}_i$, the results on minimising $\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2$ and the related proofs are given in [13]. Based on these results, all the solutions to (18) can be specified. The following theorem lists the result for one case of $\mathbf{q}_i$ and $\mathbf{z}_i$.

Theorem 4: Given positive $\alpha_i, \beta_i \in \mathcal{R}$, $\mathbf{q}_i, \mathbf{z}_i \in \mathcal{C}^m$ and $\det((\Upsilon(\mathbf{z}_i))^T \Upsilon(\mathbf{q}_i)) > 0$, we have

$$\min_{\substack{\mathbf{T}_\rho \in \mathcal{R}^{m \times m} \\ \det \mathbf{T}_\rho \neq 0}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2 = (|\mathbf{z}_i^H \mathbf{q}_i| + \alpha_i \beta_i)^2 - \alpha_i^2 \beta_i^2 + \tau_i^2, \quad (31)$$

and $\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2$ achieves the minimum if and only if

$$\mathbf{T}_\rho = \mathbf{Q} \begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix} \mathbf{V} \quad (32)$$

where the orthogonal matrix $\mathbf{Q}$ can be obtained from the QR factorisation of $\Upsilon(\mathbf{z}_i)$

$$\Upsilon(\mathbf{z}_i) = \mathbf{Q} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \quad (33)$$

with nonzero $\gamma_{11}, \gamma_{22} \in \mathcal{R}$,

$$\begin{aligned} \mathbf{H} &\triangleq \frac{\beta_i}{\alpha_i} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} (\Upsilon(\mathbf{z}_i))^T \Upsilon(\mathbf{q}_i) \\ &\quad \times \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \end{aligned} \quad (34)$$

$$\mathbf{F} \triangleq \frac{\beta_i}{\alpha_i} \begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_m^T \end{bmatrix} \mathbf{Q}^T \Upsilon(\mathbf{q}_i) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \quad (35)$$

θ is the solution of

$$\begin{cases} \tan\theta = \frac{a_{21} - a_{12}}{a_{11} + a_{22}} \\ a_{11}\cos\theta - a_{12}\sin\theta > 0 \end{cases} \quad (36)$$

with

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \triangleq (\Upsilon(\mathbf{z}_i))^T \Upsilon(\mathbf{q}_i), \quad (37)$$

$\mathbf{\Omega} \in \mathcal{R}^{(m-2)\times(m-2)}$ is an arbitrary nonsingular matrix, and $\mathbf{V} \in \mathcal{R}^{m \times m}$ is an arbitrary orthogonal matrix.

B. Global optimal controller realisations

In Section III-A, the problem of attaining the single-pole FWL stability measure η_i is solved and hence the index i' is readily given from $\eta_{i'} = \max_{i\in\{1,\cdots,m+n\}} \eta_i$. Without the loss of generality, assume that $\lambda_{i'}$ is a complex-valued eigenvalue and $\det((\Upsilon(\mathbf{z}_{i'}))^T\Upsilon(\mathbf{q}_{i'})) > 0$. From Theorem 4, all the transformation matrices achieving $\eta_{i'}$ form the set

$$\mathcal{T} \triangleq \left\{ \mathbf{T}_\rho \left| \mathbf{T}_\rho = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix} \mathbf{V} \right. \right\}. \tag{38}$$

The realisation set $\mathcal{W}$ defined in (20) is described on the transformation set $\mathcal{T}$ as

$$\mathcal{W} = \left\{ \mathbf{w}_\rho : \mathbf{w}_\rho = \mathbf{w}_\rho(\mathbf{T}_\rho) = \begin{bmatrix} \mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{F}_{\rho 0}\mathbf{T}_\rho) \\ \mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{G}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{J}_{\rho 0}\mathbf{T}_\rho) \\ \mathrm{Vec}(\mathbf{M}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{H}_{\rho 0}) \end{bmatrix} \right\}. \tag{39}$$
$$\mathbf{T}_\rho \in \mathcal{T}$$

It can readily be shown that $g(\mathbf{w}_\rho(\mathbf{T}_\rho), i) = g(\mathbf{w}_\rho(\mathbf{T}_\rho\mathbf{V}), i)$ for any orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$ and nonsingular $\mathbf{T}_\rho \in \mathcal{R}^{m\times m}$ [2]. This means that $\mathbf{V}$ plays no role in computing $g(\mathbf{w}_\rho, i)$ and hence we can simply set $\mathbf{V} = \mathbf{I}$. Therefore

$$\mathbf{T}_\rho = \mathbf{T}_\rho(\mathbf{\Omega}) = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix} \tag{40}$$

are explored for a nonsingular $\mathbf{\Omega}_{\rm opt} \in \mathcal{R}^{(m-2)\times(m-2)}$ such that $g(\mathbf{w}_\rho(\mathbf{T}_\rho(\mathbf{\Omega}_{\rm opt})), i) \le \eta_{i'}, \forall i$. The subgradient algorithm detailed in [13] can be used to seek $\mathbf{\Omega}_{\rm opt}$.

IV. Optimal Realisation with the Smallest Dynamic Range

In Section III, we construct a controller realisation $\mathbf{w}_{\rho\rm opt} = \mathbf{w}_\rho(\mathbf{T}_\rho(\mathbf{\Omega}_{\rm opt}))$ that achieves the minimum value of the FWL closed-loop stability measure (12). Since the FWL stability measure (12) is concerned with the FWL error $\mathbf{\Delta}$ that depends only on the fractional bit length b_f, an optimal realisation that minimises this precision measure is not guaranteed to have a small dynamic range. We now consider how to modify the optimal controller realisation obtained in Section III to achieve the smallest dynamic range under the condition that it remains to be a minimum solution of the optimisation problem (13). Because $\|\mathbf{w}_\rho\|_M$ indicates the dynamic range of $\mathbf{w}_\rho$, it is appropriate to use it as the dynamic range measure of a realisation, that is,

$$d(\mathbf{w}_\rho) \triangleq \|\mathbf{w}_\rho\|_M. \tag{41}$$

It is straightforward to prove the following theorem [13].

Theorem 5: For two realisations $\mathbf{w}_{\rho 1}$ and $\mathbf{w}_{\rho 2}$ (or equivalently $(\mathbf{F}_{\rho 1}, \mathbf{G}_{\rho 1}, \mathbf{J}_{\rho 1}, \mathbf{M}_{\rho 1}, \mathbf{H}_{\rho 1})$ and $(\mathbf{F}_{\rho 2}, \mathbf{G}_{\rho 2}, \mathbf{J}_{\rho 2}, \mathbf{M}_{\rho 2}, \mathbf{H}_{\rho 2})$), if there exists an orthogonal transformation $\mathbf{V} \in \mathcal{R}^{m\times m}$ such that

$$\begin{aligned} &\mathbf{F}_{\rho 2} = \mathbf{V}^{-1}\mathbf{F}_{\rho 1}\mathbf{V},\ \mathbf{G}_{\rho 2} = \mathbf{V}^{-1}\mathbf{G}_{\rho 1},\ \mathbf{J}_{\rho 2} = \mathbf{J}_{\rho 1}\mathbf{V}, \\ &\mathbf{M}_{\rho 2} = \mathbf{M}_{\rho 1},\ \mathbf{H}_{\rho 2} = \mathbf{V}^{-1}\mathbf{H}_{\rho 1}, \end{aligned} \tag{42}$$

then $f(\mathbf{w}_{\rho 1}) = f(\mathbf{w}_{\rho 2})$.

Given $\mathbf{w}_{\rho\rm opt}$ (that is, $(\mathbf{F}_{\rho\rm opt}, \mathbf{G}_{\rho\rm opt}, \mathbf{J}_{\rho\rm opt}, \mathbf{M}_{\rho\rm opt}, \mathbf{H}_{\rho\rm opt})$) obtained in Section III, define

$$\begin{aligned} \mathcal{S}_{\rho\rm opt} \triangleq \{&(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho) : \mathbf{F}_\rho = \mathbf{V}^{-1}\mathbf{F}_{\rho\rm opt}\mathbf{V}, \\ &\mathbf{G}_\rho = \mathbf{V}^{-1}\mathbf{G}_{\rho\rm opt}, \mathbf{J}_\rho = \mathbf{J}_{\rho\rm opt}\mathbf{V}, \mathbf{M}_\rho = \mathbf{M}_{\rho\rm opt}, \\ &\mathbf{H}_\rho = \mathbf{V}^{-1}\mathbf{H}_{\rho\rm opt}, \mathbf{V} \in \mathcal{R}^{m\times m}, \mathbf{V}^T\mathbf{V} = \mathbf{I}\}. \end{aligned} \tag{43}$$

Denote the generic realisation in $\mathcal{S}_{\rho\rm opt}$ as $\mathbf{w}_{\rho\rm opt}(\mathbf{V})$. It can be seen from Theorem 5 that, for any orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$, the realisation $\mathbf{w}_{\rho\rm opt}(\mathbf{V})$ remains to be a minimum solution of the optimisation problem (13). Thus, we can search in $\mathcal{S}_{\rho\rm opt}$ for an optimal realisation with the smallest dynamic range. Formally, this is defined by the following optimisation problem

$$\mu \triangleq \min_{\substack{\mathbf{V}\in\mathcal{R}^{m\times m} \\ \mathbf{V}^T\mathbf{V}=\mathbf{I}}} d(\mathbf{w}_{\rho\rm opt}(\mathbf{V})). \tag{44}$$

In order to remove the constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ in the optimisation problem (44), we can represent an orthogonal $\mathbf{V}$ by its independent parameters based on Givens rotation parameterisation [19]. Specifically, when $m = 2$, any orthogonal $\mathbf{V}$ can be written as

$$\mathbf{V} = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \kappa \end{bmatrix}, \tag{45}$$

with $\theta_1 \in [-\pi, \pi)$ and $\kappa \in \{-1, 1\}$. Next, for $m = 3$, an arbitrary orthogonal $\mathbf{V} \in \mathcal{R}^{3\times 3}$ can be represented by

$$\begin{aligned} \mathbf{V} = &\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 \\ \sin\theta_2\sin\theta_1 & \sin\theta_2\cos\theta_1 & \cos\theta_2 \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & \sin\theta_3 & \cos\theta_3 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \kappa \end{bmatrix}, \end{aligned} \tag{46}$$

with $\theta_1, \theta_2, \theta_3 \in [-\pi, \pi)$ and $\kappa \in \{-1, 1\}$. For $m = 4$, we have

$$\begin{aligned} \mathbf{V} = &\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 & 0 \\ \cos\theta_3\sin\theta_2\sin\theta_1 & \cos\theta_3\sin\theta_2\cos\theta_1 & \cos\theta_3\cos\theta_2 & -\sin\theta_3 \\ \sin\theta_3\sin\theta_2\sin\theta_1 & \sin\theta_3\sin\theta_2\cos\theta_1 & \sin\theta_3\cos\theta_2 & \cos\theta_3 \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_4 & -\sin\theta_4 & 0 \\ 0 & \cos\theta_5\sin\theta_4 & \cos\theta_5\cos\theta_4 & -\sin\theta_5 \\ 0 & \sin\theta_5\sin\theta_4 & \sin\theta_5\cos\theta_4 & \cos\theta_5 \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\theta_6 & -\sin\theta_6 \\ 0 & 0 & \sin\theta_6 & \cos\theta_6 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \kappa \end{bmatrix}, \end{aligned} \tag{47}$$

with $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 \in [-\pi, \pi)$ and $\kappa \in \{-1, 1\}$. Define

$$r = \frac{m(m-1)}{2}. \tag{48}$$

In general, an arbitrary orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$ is parameterised by $\theta_1, \cdots, \theta_r \in [-\pi, \pi)$ and $\kappa \in \{-1, +1\}$. Following from a simple observation

$$d\left(\mathbf{w}_{\rho\text{opt}}\left(\begin{bmatrix}\cos\theta_1 & -\sin\theta_1\\ \sin\theta_1 & \cos\theta_1\end{bmatrix}\right)\right) = d\left(\mathbf{w}_{\rho\text{opt}}\left(\begin{bmatrix}\cos\theta_1 & -\sin\theta_1\\ \sin\theta_1 & \cos\theta_1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\right)\right), \quad (49)$$

it can be seen that κ can be neglected in optimising $d(\mathbf{w}_{\rho\text{opt}}(\mathbf{V}))$. Thus we can represent an orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$ with only r independent parameters $\theta_1, \cdots, \theta_r$. Let

$$d_1(\theta_1, \cdots, \theta_r) \triangleq d(\mathbf{w}_{\rho\text{opt}}(\mathbf{V})). \quad (50)$$

Then the optimisation problem (44) is equivalent to the unconstrained optimisation problem

$$\mu = \min_{\theta_1,\cdots,\theta_r\in[-\pi,\pi)} d_1(\theta_1, \cdots, \theta_r). \quad (51)$$

This optimisation problem can be solved using a numerical optimisation algorithm that relies only on the function value to do search. With the optimal solution $\theta_{1\text{opt}}, \cdots, \theta_{r\text{opt}}$, we can obtain the optimal orthogonal transformation $\mathbf{V}_{\text{opt}}$ and hence the optimal realisation $\mathbf{w}_{\rho\text{opt}1} = \mathbf{w}_{\rho\text{opt}}(\mathbf{V}_{\text{opt}})$ of the smallest dynamic range.

V. A DESIGN EXAMPLE

An example considered in [2] was used to illustrate the effectiveness of the proposed design procedure for obtaining optimal FWL fixed-point controller realisations and to compare the minimum bit lengths required to implement the optimal realisations with z operator and δ operator of different h. The plant model using z operator was given by

$$\mathbf{A}_z = \begin{bmatrix}3.7156e+0 & -5.4143e+0 & 3.6525e+0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}, \quad \mathbf{B}_z = \begin{bmatrix}-9.6420e-1\\ 0\\ 0\\ 0\end{bmatrix}, \quad \mathbf{B}_z = [1 \quad 0 \quad 0 \quad 0]^T,$$

$$\mathbf{C}_z = [1.1160e-6 \quad 4.3000e-8 \quad 1.0880e-6 \quad 1.4000e-8].$$

The initial realisation of the digital controller obtained using z operator was given by

$$\mathbf{F}_{z0} = \begin{bmatrix}2.6743e+0 & -5.7446e+0 & 2.5101e+0\\ 2.8769e-1 & -2.7446e-2 & -6.9444e-1\\ -3.3773e-1 & 9.8699e-1 & -3.2925e-1\\ -8.3021e-2 & -3.1988e-3 & 9.1906e-1\end{bmatrix}, \begin{bmatrix}-9.1782e-1\\ -8.9358e-3\\ -4.2367e-3\\ -1.0415e-3\end{bmatrix},$$

$$\mathbf{G}_{z0} = [1.0959e+6 \quad 6.3827e+5 \quad 3.0262e+5 \quad 7.4392e+4]^T,$$

Realisation	$f(\mathbf{w}_z)$	$d(\mathbf{w}_z)$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
$\mathbf{w}_{z0}$	$3.9697e+6$	$1.0959e+6$	20	21	42
$\mathbf{w}_{z\text{opt}}$	$2.4246e+3$	$1.9673e+2$	8	8	17
$\mathbf{w}_{z\text{opt}1}$	$2.4246e+3$	$1.1799e+2$	8	7	16

TABLE I
COMPARISON OF VARIOUS CONTROLLER REALISATIONS USING z OPERATOR.

Realisation	$f(\mathbf{w}_\delta)$	$d(\mathbf{w}_\delta)$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
$\mathbf{w}_{\delta 0}$	$2.7712e+5$	$1.7956e+10$	15	35	51
$\mathbf{w}_{\delta\text{opt}}$	$3.3740e-1$	$5.1236e+4$	-4	16	13
$\mathbf{w}_{\delta\text{opt}1}$	$3.3740e-1$	$2.5810e+4$	-4	15	12

TABLE II
COMPARISON OF VARIOUS CONTROLLER REALISATIONS USING δ OPERATOR WITH $h = 2^{-14}$.

$$\mathbf{J}_{z0} = [1.8180e-1 \quad -2.8313e-1 \quad 5.0006e-2 \quad 6.1722e-2],$$

$$\mathbf{M}_{z0} = 0, \quad \mathbf{H}_{z0} = [0 \quad 0 \quad 0 \quad 0]^T.$$

The procedure described in Section III was applied to obtain an optimal transformation matrix $\mathbf{T}_{z\text{opt}}$, and this led to a global optimal realisation in z operator $\mathbf{w}_{z\text{opt}} = \mathbf{w}_z(\mathbf{T}_{z\text{opt}})$ that minimised the FWL closed-loop stability measure (12). To obtain an optimal realisation in z operator with the smallest dynamic range, the optimisation problem (51) was formed given the dimension $r = 6$. The MATLAB routine *fminsearch.m* was used to solve this optimisation problem numerically, which yielded the solution $\mathbf{w}_{z\text{opt}1} = \mathbf{w}_{z\text{opt}}(\mathbf{V}_{\text{opt}})$. Table I lists the values of the FWL stability measure $f(\mathbf{w}_z)$ and the dynamic range measure $d(\mathbf{w}_z)$ together with the related minimum bit lengths $b^{\min}$, $b_g^{\min}$ and $b_f^{\min}$ for the realisations $\mathbf{w}_{z0}$, $\mathbf{w}_{z\text{opt}}$ and $\mathbf{w}_{z\text{opt}1}$, respectively.

Similarly, the optimal realisation problems in the δ operator with different values of h were constructed and solved. For example, given $h = 2^{-14}$, the discrete-time plant model using δ operator, $\mathbf{A}_\delta$, $\mathbf{B}_\delta$ and $\mathbf{C}_\delta$, as well as the initial realisation of the digital controller using the δ operator, $\mathbf{F}_{\delta 0}$, $\mathbf{G}_{\delta 0}$, $\mathbf{J}_{\delta 0}$, $\mathbf{M}_{\delta 0}$ and $\mathbf{H}_{\delta 0}$, were specified. The procedure of Section III was applied to obtain a $\mathbf{T}_{\delta\text{opt}}$, which led to the controller realisation $\mathbf{w}_{\delta\text{opt}} = \mathbf{w}_\delta(\mathbf{T}_{\delta\text{opt}})$ as a global optimal realisation in δ operator that minimised the FWL closed-loop stability measure (12). The optimisation problem (51) was next solved to yield the solution $\mathbf{w}_{\delta\text{opt}1} = \mathbf{w}_{\delta\text{opt}}(\mathbf{V}_{\text{opt}})$, which was a global optimal realisation with the smallest dynamic range. Table II compares the values of the FWL stability and dynamic range measures and related minimum bit lengths for the controller realisations $\mathbf{w}_{\delta 0}$, $\mathbf{w}_{\delta\text{opt}}$ and $\mathbf{w}_{\delta\text{opt}1}$. For the δ operator with sufficiently small h, $b_f^{\min}$ can be negative. This simply means that the roundoff is allowed to occur into the integer part of fixed-point representation, and the perturbation error $\|\mathbf{\Delta}\|_M$ can be larger than 1. In this case, the minimum bit length $b^{\min} = b_g^{\min} + b_f^{\min} + 1$ required for fixed-point representation can be smaller than $b_g^{\min}$ that defines the dynamic range of the representation. As an example, "-4 fractional bits" means that the entire fractional part and the first lowest 4-bit integer part in fixed-point representation are omitted.

h	$f(\mathbf{w}_{\delta\text{opt1}})$	$d(\mathbf{w}_{\delta\text{opt1}})$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
2^{10}	$2.4825e+6$	$3.6871e+0$	18	2	21
2^{9}	$1.2413e+6$	$5.2144e+0$	17	3	21
2^{8}	$6.2063e+5$	$7.3743e+0$	16	3	20
2^{7}	$3.1032e+5$	$1.0429e+1$	15	4	20
2^{6}	$1.5516e+5$	$1.4749e+1$	14	4	19
2^{5}	$7.7579e+4$	$2.0858e+1$	13	5	19
2^{4}	$3.8790e+4$	$2.9497e+1$	12	5	18
2^{3}	$1.9395e+4$	$4.1715e+1$	11	6	18
2^{2}	$9.6977e+3$	$5.8994e+1$	10	6	17
2^{1}	$4.8490e+3$	$8.3431e+1$	9	7	17
2^{0}	$2.4246e+3$	$1.1799e+2$	8	7	16
2^{-1}	$1.2125e+3$	$1.6686e+2$	7	8	16
2^{-2}	$6.0639e+2$	$2.3598e+2$	6	8	15
2^{-3}	$3.0335e+2$	$3.3372e+2$	5	9	15
2^{-4}	$1.5183e+2$	$4.7195e+2$	4	9	14
2^{-5}	$7.6071e+1$	$6.6744e+2$	3	10	14
2^{-6}	$3.8190e+1$	$9.4391e+2$	2	10	13
2^{-7}	$1.9248e+1$	$1.3349e+3$	1	11	13
2^{-8}	$9.7758e+0$	$1.8878e+3$	0	11	12
2^{-9}	$5.0361e+0$	$2.6698e+3$	-1	12	12
2^{-10}	$2.6601e+0$	$3.7756e+3$	-2	12	11
2^{-11}	$1.4618e+0$	$5.3396e+3$	-3	13	11
2^{-12}	$8.4740e-1$	$7.6314e+3$	-3	13	11
2^{-13}	$5.2102e-1$	$1.2905e+4$	-3	14	12
2^{-14}	$3.3740e-1$	$2.5810e+4$	-4	15	12
2^{-15}	$2.2681e-1$	$5.1621e+4$	-5	16	12
2^{-16}	$1.5606e-1$	$1.0324e+5$	-6	17	12
2^{-17}	$1.0879e-1$	$2.0648e+5$	-6	18	13
2^{-18}	$7.6367e-2$	$4.1297e+5$	-6	19	14
2^{-19}	$5.3801e-2$	$8.2593e+5$	-7	20	14
2^{-20}	$3.7973e-2$	$1.6519e+6$	-7	21	15
2^{-21}	$2.6826e-2$	$3.3037e+6$	-8	22	15
2^{-22}	$1.8960e-2$	$6.6075e+6$	-8	23	16
2^{-23}	$1.3404e-2$	$1.3215e+7$	-9	24	16
2^{-24}	$9.4767e-3$	$2.6430e+7$	-9	25	17
2^{-25}	$6.7006e-3$	$5.2860e+7$	-10	26	17

TABLE III

COMPARISON OF $\mathbf{w}_{\delta\text{opt1}}$ UNDER DIFFERENT h.

Table III compares the values of $f(\mathbf{w}_{\delta\text{opt1}})$ and $d(\mathbf{w}_{\delta\text{opt1}})$ together with the related minimum bit lengths for the controller realisation $\mathbf{w}_{\delta\text{opt1}}$, giving $h = 2^{10} \sim 2^{-25}$. It can be seen that $\mathbf{w}_{z\text{opt1}}$ and $\mathbf{w}_{\delta\text{opt1}}$ of $h = 2^0 = 1$ have the identical FWL closed-loop stability characteristics, as is expected according to the definition of δ. In general, as h decreases, $f(\mathbf{w}_{\delta\text{opt1}})$ and hence $b_f^{\min}(\mathbf{w}_{\delta\text{opt1}})$ decrease, while $d(\mathbf{w}_{\delta\text{opt1}})$ and $b_g^{\min}(\mathbf{w}_{\delta\text{opt1}})$ increase. Before certain values of h (in this case, $2^{-10}, 2^{-11}, 2^{-12}$), the reduction in $b_f^{\min}$ outpaces the increase in $b_g^{\min}$ and, as a consequence, $b^{\min}$ decreases as h decreases. However, when h is smaller than these values, the increase in $b_g^{\min}$ outpaces the decrease in $b_f^{\min}$ and, consequently, $b^{\min}$ increases as h decreases. It can be concluded that there exist optimal values of h for the δ operator and the resulting optimal controller realisations $\mathbf{w}_{\delta\text{opt1}}$ achieve the maximum robustness to the FWL errors.

VI. CONCLUSIONS

A two-step approach has been developed to design optimal fixed-point realisations of digital controllers with FWL considerations. The proposed strategy first finds an optimal controller realisation by minimising an FWL closed-loop stability measure. This realisation is then modified via an effective numerical optimisation to produce an optimal realisation with the smallest dynamic range without sacrificing FWL closed-loop stability robustness. The final optimal realisation thus requires a minimum total bit length in fixed-point implementation. Our approach has been developed within the unified framework that includes both the shift and delta operator parameterisations of a generic controller structure. A design example has demonstrated that the proposed method provides an effective design procedure for obtaining optimal controller realisations that are robust to the FWL errors in fixed-point implementation. Simulation results have shown that, by choosing the value of h in the delta operator appropriately, the optimal delta-operator controller realisation has much better FWL closed-loop stability characteristics than the optimal shift-operator controller realisation.

REFERENCES

[1] P. Mantey, "Eigenvalue sensitivity and state-variable selection", *IEEE Trans. Automatic Control*, vol.13, pp.263–269, 1968.

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

[3] I.J. Fialho and T.T. Georgiou, "On stability and performance of sampled-data systems subject to wordlength constraint", *IEEE Trans. Automatic Control*, vol.39, pp.2476–2481, 1994.

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

[5] S. Chen, J. Wu, R.S.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures", *IEEE Trans. Automatic Control*, vol.44, pp.2149–2153, 1999.

[6] S. Chen, R.S.H. Istepanian, J. Wu and J. Chu, "Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators", *Systems and Control Letters*, vol.40, pp.153–163, 2000.

[7] J.F. Whidborne, J. Wu and R.S.H Istepanian, "Finite word length stability issues in an l_1 framework", *Int. J. Control*, vol.73, pp.166–176, 2000.

[8] J. Wu, S. Chen, G. Li, R.S.H. Istepanian and J. Chu, "Shift and delta operator realizations for digital controllers with finite-word-length considerations", *IEE Proc. Control Theory and Applications*, vol.147, pp.664–672, 2000.

[9] R.S.H. Istepanian and J.F. Whidborne (Eds.), *Digital Controller Implementation and Fragility: A Modern Perspective*. London: Springer Verlag, 2001.

[10] J. Wu, S. Chen, G. Li, R.S.H. Istepanian and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations", *IEEE Trans. Automatic Control*, vol.46, pp.1162–1166, 2001.

[11] J.F. Whidborne, R.S.H. Istepanian and J. Wu, "Reduction of controller fragility by pole sensitivity minimization", *IEEE Trans. Automatic Control*, vol.46, pp.320–325, 2001.

[12] S. Chen, J. Wu and G. Li, "Two approaches based on pole sensitivity and stability radius measures for finite precision digital controller realizations", *Systems and Control Letters*, vol.45, pp.321–329, 2002.

[13] J. Wu, S. Chen, G. Li and J. Chu, "A search algorithm for a class of optimal finite-precision controller realization problems with saddle points," *SIAM J. Control and Optimization*, vol.44, pp.1787–1810, 2005.

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

[15] T. Kailath, *Linear Systems*. Upper Saddle River, NJ: Prentice Hall, 1980.

[16] J. O'Reilly, *Observers for Linear Systems*. New York: Academic Press, 1983.

[17] G. Owen, *Game Theory*. New York: Academic Press, 1982.

[18] J. Szèp and F. Forgó, *Introduction to the Theory of Games*. Dordrecht, Holland: D. Reidel Publishing Company, 1985.

[19] J.-P. Delmas,"Performances analysis of a Givens parametrized adaptive eigenspace algorithm," *Signal Processing*, vol.68, pp.87–105, 1998.