International Journal of Control
Vol. 79, No. 10, October 2006, 1297–1312

Taylor & Francis
Taylor & Francis Group

Optimal realizations of fixed-point implemented digital controllers with the smallest dynamic range

J. WU†, S. CHEN*‡, G. LI§ and J. CHU†

†National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, P.R. China
‡School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, U.K.
§School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

(*Received 19 August 2005; in final form 7 November 2005*)

A novel approach is proposed to design optimal finite word length (FWL) realizations of digital controllers implemented in fixed-point arithmetic. A minimax-based search procedure is first formulated to obtain an optimal controller realization that optimizes 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 realization may not have the smallest dynamic range. A measure is then derived to indicate the dynamic range of fixed-point implemented realization. By choosing an appropriate orthogonal transformation of this dynamic range measure of the optimal precision controller realization, a numerical optimization method is developed to make the controller realization having the smallest dynamic range without sacrificing FWL closed-loop stability robustness. The proposed approach is more efficient than a direct optimization of some combined FWL closed-loop stability and dynamic range measure via a numerical means. The proposed approach is established within a unified framework that includes both the shift and delta operator parameterizations, which makes it possible to compare the closed-loop stability characteristics of the optimal FWL controller realizations using shift and delta operators, respectively. Through analysing the simulation results of a design example, some useful insights and understandings are obtained regarding the FWL controller realizations based on shift and delta operators.

1. Introduction

In a closed-loop control system, there generally exist two kinds of uncertainty which have detrimental effects on the system performance. The first is the uncertainty within the plant. This kind of uncertainty has been extensively studied, and some effective methods, such as H_∞ method (Zhou *et al.* 1996) and l_1 method (Dahleh and Diaz-Bobillo 1995), have been established to design controllers which are capable of dealing with the plant uncertainty. When a designed control law is implemented, the second kind of uncertainty, the uncertainty within the controller, arises. It is well-known that, in practice, a controller cannot be implemented exactly. For example, when a control law is digitally implemented using a digital processor of finite word length (FWL), the finite-precision representation of the controller parameters is the main source of controller uncertainty. In comparison with the plant uncertainty, this controller uncertainty is small. This is the reason why the classical controller design methodology has ignored the controller uncertainty. However, it has increasingly been realized that the controller uncertainty due to the FWL effect cannot be ignored (Liu *et al.* 1992, Gevers and Li 1993, De Oliveira and Skelton 2001, Istepanian and Whidborne 2001). Firstly, for many industrial and mass-market consumer applications, a fixed-point implementation of digital controller is desired for its advantages in cost, simplicity, speed,

*Corresponding author. Email: sqc@ecs.soton.ac.uk

International Journal of Control
ISSN 0020–7179 print/ISSN 1366–5820 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207170600833257

memory space and power consumption. With a fixed-point processor, however, the detrimental FWL effects are markedly increased due to a reduced precision. Secondly, some canonical controller realizations are inherently ill-conditioned, and a small error will rapidly accumulate, subsequently degrading the designed closed-loop performance and even resulting in closed-loop instability. Furthermore, modern control design methods result in controllers of high order, where such FWL effects are even more pronounced, as is highlighted in the so-called fragility puzzles (Keel and Bhattacharryya 1997). Thus it is imperative that great care must be exercised when implementing digital controllers.

As the first and the most critical requirement for a closed-loop control system is its stability, most researches in digital controller implementation have focused on the FWL effects on closed-loop stability (Fialho and Georgiou 1994, Li 1998, Chen *et al*. 1999, Whidborne *et al*. 2000a, 2001, Fialho and Georgiou 2001, Wu *et al*. 2001). A basic idea underpinning all these researches follows. There exist an infinite number of different realizations corresponding to a control law. Although these controller realizations are equivalent, if infinite-precision implementation can be assumed, they are no longer equivalent under practical finite-precision implementation. It is recognized that, subject to the FWL effect, certain controller realizations exhibit superior ''robustness'' of closed-loop stability, compared to others. This observation can be utilized to select ''optimal'' realizations that optimize some given FWL closed-loop stability measures. Various FWL closed-loop stability measures have been investigated, and these include the complex stability radius measure (Fialho and Georgiou 2001, Chen *et al*. 2002), a variety of pole sensitivity measures (Mantey 1968, Li 1998, Chen *et al*. 1999, Whidborne *et al*. 2001, Wu *et al*. 2001) and the l_1 based stability measure (Whidborne *et al*. 2000a). This approach has also been extended to study the closed-loop stability issues of FWL controller realizations using the delta operator formulation (Chen *et al*. 2000, Wu *et al*. 2000). All these measures in the previous works, designed for fixed-point implementation, have a limitation in that they are only linked to the fractional part of fixed-point representation. Optimizing these measures, while minimizing 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 realizations 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 to design optimal fixed-point controller realizations is to consider both a precision or FWL closed-loop stability measure and a dynamic range measure together. In a recent study (Wu *et al*. 2003), this approach is adopted for fixed-point, floating-point or block-floating-point implemented digital controllers. A potential drawback of this previous approach is high computational complexity, particularly for high-order controllers. This is because numerical methods have to be used which can only rely on function values for optimization search. In this study, we adopt a very different ''two-procedure'' approach. Firstly, we optimize the FWL closed-loop stability measure proposed by Li (1998) to obtain an optimal realization. Secondly, we then optimize a dynamic range measure for this optimal realization. This second-step is based on an invariant property of the controller realization under orthogonal transformation. It is known that the value of the FWL closed-loop stability measure is invariant under an orthogonal transformation of controller realization (Gevers and Li 1993) and this property was utilized by Gevers and Li (1993) to obtain sparse realizations. We exploit this extra freedom of realization to minimize the dynamic range of the controller realization.

This two-procedure approach is attractive for the following reasons. In the first procedure, the minimax theorem and subgradient algorithms are used to search for a global optimal solution of the given FWL closed-loop stability measure (Wu *et al*. 2002). Thus, provided that there are sufficient bits for accommodating the dynamic range, the realization obtained is global optimal and is most robust to the FWL effect. In the second procedure, based on an appropriate orthogonal transformation, numerical optimization is carried out in a much smaller space than the full realization space, and the resulting realization remains to be an optimal realization with respect to the first-procedure measure. That is, the final realization has the smallest dynamic range under the constraint that it also has the maximum FWL stability robustness. 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 realizations using these two operators.

The remainder of this paper is organized as follows. Section 2 formulates the problem to be dealt with in the framework that unifies both the shift and delta operator parameterizations of a general controller structure. Section 3 introduces an FWL closed-loop stability measure and develops a procedure which optimizes the given FWL closed-loop stability measure to obtain an optimal realization. This section is based on an extension of our previous work by Wu *et al*. (2002, 2005) to the current unified control system framework.

In § 4, a criterion is introduced which measures the dynamic range of a fixed-point realization, and a method is developed to minimize this dynamic range criterion over the set that contains all the orthogonal transformations of the optimal realization obtained using the procedure described in the previous section. A comparison with a direct numerical optimization of the combined FWL closed-loop stability and dynamic range measure (Wu *et al.* 2003) is also given in this section. In § 5, a design example is used to demonstrate the effectiveness of the proposed optimization strategy and to compare the FWL closed-loop stability characteristics of optimal controller realizations using the shift and delta operators. The simulation results are analysed to reveal useful insights to these two different operator parameterizations of controller. The paper concludes in § 6.

2. Notations and the problem formulation

Let $\mathcal{R}$ denote the field of real numbers, $\mathcal{C}$ the field of complex numbers, and $\mathbf{e}_i$ the ith real coordinate vector. For any $\mathbf{z} \in \mathcal{C}^n$, define

$$\Upsilon(\mathbf{z}) \triangleq [\Re(\mathbf{z}) \quad \Im(\mathbf{z}))], \tag{1}$$

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

$$\|\mathbf{U}\|_M \triangleq \max_{\substack{i\in\{1,\ldots,m\}\\ j\in\{1,\ldots,n\}}} |u_{ij}|, \tag{2}$$

$$\|\mathbf{U}\|_F \triangleq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|u_{ij}|^2}. \tag{3}$$

Let Vec(·) be the column stacking operator such that Vec($\mathbf{U}$) is an mn-dimenstional 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 real-valued square matrix $M \in \mathcal{R}^{n\times n}$, 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 $\lambda_i(\mathbf{M})$. If $\mathbf{M}$ is diagonalizable, the matrix

$$\mathbf{M}_x \triangleq [\mathbf{x}_1(\mathbf{M}) \quad \mathbf{x}_2(\mathbf{M}) \cdots \mathbf{x}_n(\mathbf{M})] \tag{4}$$

is invertible. Define

$$\mathbf{M}_y = \left[\mathbf{y}_1(\mathbf{M}) \quad \mathbf{y}_2(\mathbf{M}) \cdots \mathbf{y}_n(\mathbf{M})\right] \triangleq \mathbf{M}_x^{-H} \tag{5}$$

where $\mathbf{y}_i(\mathbf{M})$ is called the reciprocal left eigenvector corresponding to $\mathbf{x}_i(\mathbf{M})$.

A discrete-time linear system can be described using either the usual forward shift operator z or the so-called delta operator δ. The latter is defined as (Middleton and Goodwin 1990)

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

where h is a positive real constant (the constant h is originally limited to the sampling period by Middleton and Goodwin (1990) but this constraint is removed by Gevers and Li (1993)). 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 realization. For the notational conciseness and to avoid separate derivations for the two operators, we introduce a "generalized" operator ρ for the discrete-time system. It is understood that $\rho = z$ or δ, depending on which operator is actually used. The state-space description of the general discrete-time system using the generalized operator ρ is

$$\begin{cases} \rho\mathbf{x}_g(k) = \mathbf{F}_{g,\rho}\mathbf{x}_g(k) + \mathbf{G}_{g,\rho}\mathbf{u}_{g,1}(k) + \mathbf{H}_{g,\rho}\mathbf{u}_{g,2}(k) \\ \mathbf{y}_g(k) = \mathbf{J}_{g,\rho}\mathbf{x}_g(k) + \mathbf{M}_{g,\rho}\mathbf{u}_{g,1}(k), \end{cases} \tag{7}$$

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

$$\begin{aligned} &\mathbf{F}_{g,\delta} = \frac{F_{g,z} - \mathbf{I}}{h}, \quad \mathbf{G}_{g,\delta} = \frac{\mathbf{G}_{g,z}}{h}, \quad \mathbf{J}_{g,\delta} = \mathbf{J}_{g,z}, \\ &\mathbf{M}_{g,\delta} = \mathbf{M}_{g,z}, \quad \mathbf{H}_{g,\delta} = \frac{\mathbf{H}_{g,z}}{h}, \end{aligned} \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension. In particular, the operator z is interpreted as $z\mathbf{x}_g(k) = \mathbf{x}_g(k+1)$. The following theorem relates the eigenvalues of $\mathbf{F}_{g,z}$ to those of $\mathbf{F}_{g,\delta}$.

Theorem 1: *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*

$$\lambda_i(\mathbf{F}_{g,z}) = 1 + h\lambda_i(\mathbf{F}_{g,\delta}), \quad \forall i. \tag{9}$$

It is well known that the discrete-time 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*

$$|\lambda_i(\mathbf{F}_{g,z})| < 1, \quad \forall i. \tag{10}$$

From Theorem 1, *we have the stability condition for the same system described using* δ *operator*.

Theorem 2: *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*

$$\left|\lambda_i(\mathbf{F}_{g,\delta}) + \frac{1}{h}\right| < \frac{1}{h}, \quad \forall i. \tag{11}$$

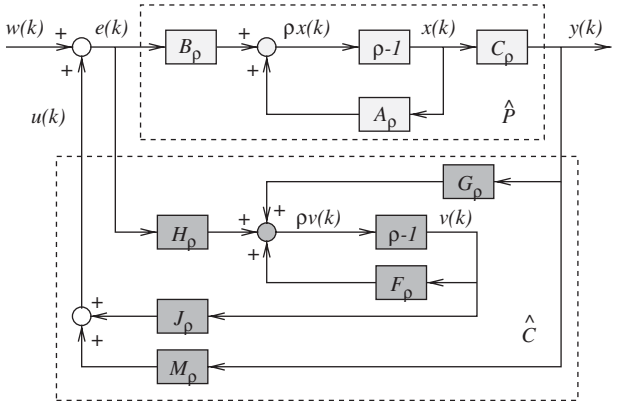

Figure 1. Discrete-time closed-loop system with a generic controller using the generalized operator ρ.

Now consider the discrete-time closed-loop control system depicted in figure 1, where the linear time-invariant plant $\hat{P}$ is described by the state-space description

$$\begin{cases} \rho\mathbf{x}(k) = \mathbf{A}_\rho\mathbf{x}(k) + \mathbf{B}_\rho\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}_\rho\mathbf{x}(k) \end{cases} \tag{12}$$

which is completely state controllable and 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 generic digital stabilizing controller $\hat{C}$ is described by the state-space description

$$\begin{cases} \rho\mathbf{v}(k) = \mathbf{F}_\rho\mathbf{v}(k) + \mathbf{G}_\rho\mathbf{y}(k) + H_\rho\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}_\rho\mathbf{v}(k) + \mathbf{M}_\rho\mathbf{y}(k) \end{cases} \tag{13}$$

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 $\mathbf{H}_\rho \in \mathcal{R}^{m\times p}$. The generic controller structure in figure 1 unifies the output feedback and observer-based controllers: $\hat{C}$ is an output feedback controller when $\mathbf{H}_\rho = 0$; a full-order observer-based 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 observer-based controller, otherwise (Kailath 1980, O'Reilly 1983).

According to a basic property of the linear system, the state-space descriptions or realizations ($\mathbf{F}_\rho$, $\mathbf{G}_\rho$, $\mathbf{J}_\rho$, $\mathbf{M}_\rho$, $\mathbf{H}_\rho$) of the controller $\hat{C}$ are not unique. In fact, let ($\mathbf{F}_{\rho 0}$, $\mathbf{G}_{\rho 0}$, $\mathbf{J}_{\rho 0}$, $\mathbf{M}_{\rho 0}$, $\mathbf{H}_{\rho 0}$) be a realization of $\hat{C}$ that has been designed using a standard controller design procedure. Then all the realizations of $\hat{C}$ form a realization set

$$\begin{aligned} \mathcal{S}_\rho \triangleq \{(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho) : \mathbf{F}_\rho = \mathbf{T}_\rho^{-1}\mathbf{F}_{\rho 0}\mathbf{T}_\rho, \mathbf{G}_\rho = \mathbf{T}_\rho^{-1}\mathbf{G}_{\rho 0}, \\ \mathbf{J}_\rho = \mathbf{J}_{\rho 0}\mathbf{T}_\rho, \mathbf{M}_\rho = \mathbf{M}_{\rho 0}, \mathbf{H}_\rho = \mathbf{T}_\rho^{-1}\mathbf{H}_{\rho 0}\} \end{aligned} \tag{14}$$

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

$$\mathbf{w}_\rho \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_\rho) \\ \mathrm{Vec}(\mathbf{G}_\rho) \\ \mathrm{Vec}(\mathbf{J}_\rho) \\ \mathrm{Vec}(\mathbf{M}_\rho) \\ \mathrm{Vec}(\mathbf{H}_\rho) \end{bmatrix}, \quad \mathbf{w}_{\rho 0} \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{G}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{J}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{M}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{H}_{\rho 0}) \end{bmatrix}. \tag{15}$$

We also refer to $\mathbf{w}_\rho$ as a realization of $\hat{C}$. The stability of the closed-loop control system depicted in figure 1 depends on the eigenvalues of the transition matrix

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{w}_\rho) &\triangleq \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{G}_\rho\mathbf{C}_\rho + \mathbf{H}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho\mathbf{J}_\rho \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_{\rho 0}\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_{\rho 0} \\ \mathbf{G}_{\rho 0}\mathbf{C}_\rho + \mathbf{H}_{\rho 0}\mathbf{M}_{\rho 0}\mathbf{C}_\rho & \mathbf{F}_{\rho 0} + \mathbf{H}_{\rho 0}\mathbf{J}_{\rho 0} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho \end{bmatrix} \\ &\triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_{\rho 0}) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_\rho \end{bmatrix}, \end{aligned} \tag{16}$$

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

$$SM(\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))) \triangleq \begin{cases} 1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_z))|, & \text{if } \rho = z, \\ \dfrac{1}{h} - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\delta))| + \dfrac{1}{h}, & \text{if } \rho = \delta. \end{cases} \tag{17}$$

From the fact that the closed-loop system is designed to be stable, it follows

$$SM(\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))) = SM(\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))) > 0, \quad \forall i \in \{1, \ldots, m+n\} \tag{18}$$

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

In practice, however, a controller realization can only be implemented with finite precision. When $\mathbf{w}_\rho$ is implemented using a fixed-point processor of the bit length b, 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

$$\|\mathbf{w}_\rho\|_M \leq 2^{b_g}. \tag{19}$$

Note that $\|\mathbf{w}_\rho\|_M$ 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,

$$\|\mathbf{\Delta}\|_M \leq 2^{-(b_f+1)}. \tag{20}$$

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, it can be seen that the closed-loop stability depends only on the perturbation $\mathbf{\Delta}$, that is, the accuracy or precision of the fractional part representation.

Intuitively, different controller realizations have different degrees of robustness to the FWL effect. It is highly desired to be able to quantify how robust a controller realization is in terms of its closed-loop stability under FWL implementation and to find some optimal realization that has the maximum robustness to the FWL effect. Because the total bit length b is divided between the dynamic range and precision of fixed-point format, this is a multi-objective optimization. Firstly, an optimal realization should optimize some FWL closed-loop stability measure. Note that the value of such a stability measure only depends on the precision or fractional part of a controller realization. Secondly, a desired realization 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. In this study, we will adopt an effective two-procedure approach to tackle this multi-objective optimization problem.

3. Optimizing an FWL closed-loop stability measure

In the remainder of this paper, λ_i is used to replace $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_\rho))$ when doing so does not cause ambiguity. Under the condition of no overflow, how easily the FWL error $\mathbf{\Delta}$ can cause a stable control system to become unstable is determined by how much the stability margin each eigenvalue λ_i has and how sensitive the closed-loop eigenvalues are to the controller parameter perturbations. The following FWL closed-loop stability measure, defined by Li (1998), is considered in this study. We adopt the inverse of the measure (thus the objective is to minimize) and remove the constant $\sqrt{N}$ given by Li (1998).

$$f(\mathbf{w}_\rho) \triangleq \max_{i\,\in\,\{1,\ldots,m+n\}} \frac{\|\partial\lambda_i/\partial\mathbf{w}_\rho\|_F}{SM(\lambda_i)}. \tag{21}$$

The measure $f(\mathbf{w}_\rho)$ describes the "robustness" of closed-loop stability of the FWL perturbation $\mathbf{\Delta}$ for the realization $\mathbf{w}_\rho$. Since different controller realizations $\mathbf{w}_\rho$ have different values of $f(\mathbf{w}_\rho)$, it is natural to search for "optimal" controller realizations that minimize the measure defined in (21). This leads to an optimal FWL controller realization problem

$$\nu \triangleq \min_{\mathbf{w}_\rho\,\in\,\mathcal{S}_\rho} f(\mathbf{w}_\rho). \tag{22}$$

Define

$$g(\mathbf{w}_\rho, i) \triangleq \frac{\|\partial\lambda_i/\partial\mathbf{w}_\rho\|_F}{SM(\lambda_i)}. \tag{23}$$

Obviously, the optimization problem (22) can be viewed as

$$\nu = \min_{\mathbf{w}_\rho\,\in\,\mathcal{S}_\rho} \max_{i\,\in\,\{1,\ldots,m+n\}} g(\mathbf{w}_\rho, i). \tag{24}$$

The following results (Owen 1982, Szép and Forgó 1985) on saddle points play an important role in obtaining global optimal solutions of minimax-formulation problems.

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

$$g(\mathbf{w}'_\rho, i) \leq g(\mathbf{w}'_\rho, i') \leq g(\mathbf{w}_\rho, i'), \quad \forall \mathbf{w}_\rho \in \mathcal{S}_\rho, \quad \forall i \in \{1,\ldots,m+n\}. \tag{25}$$

The next theorem is the well-known minimiax theorem in game theory.

Theorem 3: *If and only if there exists at least a saddle point* $(\mathbf{w}'_\rho, i')$ *of* $g(\mathbf{w}_\rho, i)$, *then*

$$\min_{\mathbf{w}_\rho\,\in\,\mathcal{S}_\rho} \max_{i\,\in\,\{1,\ldots,m+n\}} g(\mathbf{w}_\rho, i) = \max_{i\,\in\,\{1,\ldots,m+n\}} \min_{\mathbf{w}_\rho\,\in\,\mathcal{S}_\rho} g(\mathbf{w}_\rho, i) = g(\mathbf{w}'_\rho, i'). \tag{26}$$

Theorem 4: *Let*

$$\eta_i \triangleq \min_{\mathbf{w}_\rho\,\in\,\mathcal{S}_\rho} g(\mathbf{w}_\rho, i)\ \forall i \in \{1,\ldots,m+n\}, \tag{27}$$

$$i' \triangleq \arg \max_{i\,\in\,\{1,\ldots,m+n\}} \eta_i, \tag{28}$$

$$\mathcal{W} \triangleq \{\mathbf{w}_\rho : g(\mathbf{w}_\rho, i') = \eta_{i'}, \mathbf{w}_\rho \in \mathcal{S}_\rho\}. \tag{29}$$

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*

$$g(\mathbf{w}'_\rho, i) \leq \eta_{i'}, \quad \forall i \in \{1,\ldots,m+n\}\backslash\{i'\}. \tag{30}$$

For closed-loop system with the forward shift operator z and output feedback controllers, a minimax-based search procedure was derived in Wu *et al.* (2005) for finding a global optimal solution of (24). In this section,

the procedure is extended into the generalized operator ρ and the generic controller (13). The proposed search procedure, which consists of two stages, is outlined as follows.

3.1 *Optimizing single-pole FWL stability measure*

Given the realization $\mathbf{w}_{\rho 0}$, from the definition (14) and (15), $\mathbf{w}_\rho$ actually depends on the transformation matrix $\mathbf{T}_\rho$. In addition, $SM(\lambda_i)$ is fixed for given λ_i. Thus, to attain the single-pole measure η_i defined in (27) for the eigenvalue λ_i is equivalent to solve the minimization problem of the single-pole sensitivity

$$\min_{\mathrm{T}_\rho \in \mathcal{R}^{m\times m}_{\det \mathrm{T}_\rho \neq 0}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{w}_\rho} \right\|_F^2. \tag{31}$$

The following lemma is due to Li (1998).

Lemma 1: *Let the square matrix* $\mathbf{A} = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2$ *be diagonalizable where the real-valued matrices* $\mathbf{M}_0$, $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *have proper dimensions and are independent of the real-valued matrix* $\mathbf{X}$. *Then*

$$\frac{\partial \lambda_i(\mathbf{A})}{\partial \mathbf{X}} = \mathbf{M}_1^T y_i^*(\mathbf{A})\mathbf{x}_i^T(\mathbf{A})\mathbf{M}_2^T. \tag{32}$$

From (16), it can be seen that

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

$$\overline{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{H}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho\mathbf{J}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}\mathbf{G}_\rho\begin{bmatrix} \mathbf{C}_\rho & \mathbf{0} \end{bmatrix}, \tag{34}$$

$$\overline{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{0} \\ \mathbf{G}_\rho\mathbf{C}_\rho + \mathbf{H}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{B}_\rho \\ \mathbf{H}_\rho \end{bmatrix}\mathbf{J}_\rho\begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{35}$$

$$\overline{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{G}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho + \mathbf{H}_\rho\mathbf{J}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{B}_\rho \\ \mathbf{H}_\rho \end{bmatrix}\mathbf{M}_\rho\begin{bmatrix} \mathbf{C}_\rho & \mathbf{0} \end{bmatrix}, \tag{36}$$

$$\overline{\mathbf{A}}(\mathbf{w}_\rho) = \begin{bmatrix} \mathbf{A}_\rho + \mathbf{B}_\rho\mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{B}_\rho\mathbf{J}_\rho \\ \mathbf{G}_\rho\mathbf{C}_\rho & \mathbf{F}_\rho \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}\mathbf{H}_\rho\begin{bmatrix} \mathbf{M}_\rho\mathbf{C}_\rho & \mathbf{J}_\rho \end{bmatrix}. \tag{37}$$

Applying Lemma 1 to (33)–(37) gives rise to

$$\frac{\partial \lambda_i}{\partial \mathbf{F}_\rho} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}_\rho))\mathbf{x}_i^T(\overline{\mathbf{A}}(\mathbf{w}_\rho))\begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}, \tag{38}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{G}_\rho} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}_\rho))\mathbf{x}_i^T(\overline{\mathbf{A}}(\mathbf{w}_\rho))\begin{bmatrix} \mathbf{C}_\rho^T \\ \mathbf{I} \end{bmatrix}, \tag{39}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{J}_\rho} = \begin{bmatrix} \mathbf{B}_\rho^T & \mathbf{H}_\rho^T \end{bmatrix}\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}_\rho))\mathbf{x}_i^T(\overline{\mathbf{A}}(\mathbf{w}_\rho))\begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}, \tag{40}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{M}_\rho} = \begin{bmatrix} \mathbf{B}_\rho^T & \mathbf{H}_\rho^T \end{bmatrix}\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}_\rho))\mathbf{x}_i^T(\overline{\mathbf{A}}(\mathbf{w}_\rho))\begin{bmatrix} \mathbf{C}_\rho^T \\ \mathbf{0} \end{bmatrix}, \tag{41}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{H}_\rho} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}_\rho))\mathbf{x}_i^T(\overline{\mathbf{A}}(\mathbf{w}_\rho))\begin{bmatrix} \mathbf{C}_\rho^T\mathbf{M}_\rho^T \\ \mathbf{J}_\rho^T \end{bmatrix}. \tag{42}$$

$\forall i \in \{1, \ldots, m+n\}$, partition the eigenvectors of $\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})$, $\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))$ and $\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))$, into

$$\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \quad \mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) = \begin{bmatrix} \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \tag{43}$$

where $\mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})),\ \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \in \mathcal{C}^n$ and $\mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))$, $\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}) \in \mathcal{C}^m$. It is easy to see from (16) that, $\forall i \in \{1, \ldots, m+n\}$,

$$\begin{aligned} \mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_\rho)) &= \begin{bmatrix} \mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{T}_\rho^{-1}\mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}, \\ \mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_\rho)) &= \begin{bmatrix} \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ \mathbf{T}_\rho^T\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \end{bmatrix}. \end{aligned} \tag{44}$$

Applying (44) to (38)–(42) results in

$$\frac{\partial \lambda_i}{\partial \mathbf{F}_\rho} = \mathbf{T}_\rho^T\mathbf{y}_{i,2}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{x}_{i,2}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T}, \tag{45}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{G}_\rho} = \mathbf{T}_\rho^T\mathbf{y}_{i,2}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{x}_{i,1}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^T, \tag{46}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{J}_\rho} = \left(\mathbf{B}_\rho^T\mathbf{y}_{i,1}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T\mathbf{y}_{i,2}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,2}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{T}_\rho^{-T}, \tag{47}$$

$$\frac{\partial \lambda_i}{\partial \mathbf{M}_\rho} = \left(\mathbf{B}_\rho^T\mathbf{y}_{i,1}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T\mathbf{y}_{i,2}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\right)\mathbf{x}_{i,1}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^T, \tag{48}$$

$$\begin{aligned} \frac{\partial \lambda_i}{\partial \mathbf{H}_\rho} = {} & \mathbf{T}_\rho^T\mathbf{y}_{i,2}^*(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) \\ & \times \left(\mathbf{x}_{i,1}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{C}_\rho^T\mathbf{M}_{\rho 0}^T + \mathbf{x}_{i,2}^T(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\mathbf{J}_{\rho 0}^T\right). \end{aligned} \tag{49}$$

Let

$$\alpha_i^2 \triangleq \|\mathbf{C}_p\mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2 + \|\mathbf{M}_{\rho 0}\mathbf{C}_\rho\mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{J}_{\rho 0}\mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \tag{50}$$

$$\beta_i^2 \triangleq \|\mathbf{B}_\rho^T\mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \tag{51}$$

$$\tau_i^2 \triangleq \|\mathbf{B}_\rho^T\mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})) + \mathbf{H}_{\rho 0}^T\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2\|\mathbf{C}_\rho\mathbf{x}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0}))\|_F^2, \tag{52}$$

$$\mathbf{q}_i \triangleq \mathbf{x}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})), \tag{53}$$

$$\mathbf{z}_i \triangleq \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_{\rho 0})). \tag{54}$$

Then

$$\begin{aligned}\left\|\frac{\partial\lambda_i}{\partial\mathbf{w}_\rho}\right\|_F^2 &= \left\|\frac{\partial\lambda_i}{\partial\mathbf{F}_\rho}\right\|_F^2 + \left\|\frac{\partial\lambda_i}{\partial\mathbf{G}_\rho}\right\|_F^2 + \left\|\frac{\partial\lambda_i}{\partial\mathbf{J}_\rho}\right\|_F^2 + \left\|\frac{\partial\lambda_i}{\partial\mathbf{M}_\rho}\right\|_F^2 + \left\|\frac{\partial\lambda_i}{\partial\mathbf{H}_\rho}\right\|_F^2 \\ &= \|\mathbf{T}_\rho^{-1}\mathbf{q}_i\|_F^2\|\mathbf{T}_\rho^T\mathbf{z}_i\|_F^2 + \alpha_i^2\|\mathbf{T}_\rho^T\mathbf{z}_i\|_F^2 \\ &\quad + \beta_i^2\|\mathbf{T}_\rho^{-1}\mathbf{q}_i\|_F^2 + \tau_i^2.\end{aligned} \tag{55}$$

For the different cases of $\mathbf{q}_i$ and $\mathbf{z}_i$, the results on minimizing $\|\partial\lambda_i/\partial\mathbf{w}_\rho\|_F^2$ and the related proofs are given in Wu *et al*. (2005). Based on these results, all the solutions to (27) can be specified. The following theorem lists the result for one case of $\mathbf{q}_i$ and $\mathbf{z}_i$ to illustrate how the problem is solved.

Theorem 5: *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*

$$\min_{\substack{\mathbf{T}_\rho \in \mathcal{R}^{m\times m} \\ \det\mathbf{T}_\rho \neq 0}} \left\|\frac{\partial\lambda_i}{\partial\mathbf{w}_\rho}\right\|_F^2 = \left(|\mathbf{z}_i^H\mathbf{q}_i| + \alpha_i\beta_i\right)^2 - \alpha_i^2\beta_i^2 + \tau_i^2, \tag{56}$$

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

$$\mathbf{T}_\rho = \mathbf{Q}\begin{bmatrix}\mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \boldsymbol{\Omega}\end{bmatrix}\mathbf{V} \tag{57}$$

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

$$\Upsilon(\mathbf{z}_i) = \mathbf{Q}\begin{bmatrix}\gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0\end{bmatrix} \tag{58}$$

with non-zero $\gamma_{11}, \gamma_{22} \in \mathcal{R}$,

$$\mathbf{H} \triangleq \frac{\beta_i}{\alpha_i}\begin{bmatrix}\gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22}\end{bmatrix}^{-T}(\Upsilon(\mathbf{z}_i))^T\Upsilon(\mathbf{q}_i)\begin{bmatrix}\cos\theta & \sin\theta \\ -\sin\theta & \cos\theta\end{bmatrix} \times \begin{bmatrix}\gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22}\end{bmatrix}^{-1} \tag{59}$$

$$\mathbf{F} \triangleq \frac{\beta_i}{\alpha_i}\begin{bmatrix}\mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_m^T\end{bmatrix}\mathbf{Q}^T\Upsilon(\mathbf{q}_i)\begin{bmatrix}\cos\theta & \sin\theta \\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22}\end{bmatrix}^{-1}, \tag{60}$$

θ *is the solution of*

$$\begin{cases}\tan\theta = \dfrac{a_{21} - a_{12}}{a_{11} + a_{22}} \\ a_{11}\cos\theta - a_{12}\sin\theta > 0\end{cases} \tag{61}$$

with

$$\begin{bmatrix}a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} \triangleq (\Upsilon(\mathbf{z}_i))^T\Upsilon(\mathbf{q}_i), \tag{62}$$

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

3.2 *Global optimal controller realizations*

In § 3.1, the problem of attaining the single-pole FWL stability measure η_i is solved and hence the index i' is readily given from $\eta_{i'} = \max_{i\in\{1,\ldots,m+n\}}\eta_i$. Without the loss of generality, it is assumed that $\lambda_{i'}$ is a complex-valued eigenvalue and $\det((\Upsilon(\mathbf{z}_{i'}))^T\Upsilon(\mathbf{q}_{i'})) > 0$. From Theorem 5, all the transformation matrices achieving $\eta_{i'}$ form the set

$$\mathcal{T} \triangleq \left\{\mathbf{T}_\rho \middle| \mathbf{T}_\rho = \mathbf{Q}\begin{bmatrix}\mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \boldsymbol{\Omega}\end{bmatrix}\mathbf{V}\right\} \tag{63}$$

where $\mathbf{Q}$, $\mathbf{H}$ and $\mathbf{F}$ are determined according to $\alpha_{i'}$, $\beta_{i'}$, $\mathbf{q}_{i'}$, $\mathbf{z}_{i'}$ as well as Theorem 5, $\Omega \in \mathcal{R}^{(m-2)\times(m-2)}$ is an arbitrary non-singular matrix and $\mathbf{V} \in \mathcal{R}^{m\times m}$ is an arbitrary orthogonal matrix. The realization set $\mathcal{W}$ defined in (29) is described on the transformation set $\mathcal{T}$ as

$$\mathcal{W} = \left\{\mathbf{w}_\rho : \mathbf{w}_\rho = \mathbf{w}_\rho(\mathrm{T}_\rho) = \begin{bmatrix}\mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{F}_{\rho 0}\mathbf{T}_\rho) \\ \mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{G}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{J}_{\rho 0}\mathbf{T}_\rho) \\ \mathrm{Vec}(\mathbf{M}_{\rho 0}) \\ \mathrm{Vec}(\mathbf{T}_\rho^{-1}\mathbf{H}_{\rho 0})\end{bmatrix}, \mathbf{T}_\rho \in \mathcal{T}\right\}. \tag{64}$$

From (23), (55) and the definition of $\|\cdot\|_F$, it can be seen 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 non-singular $\mathbf{T}_\rho \in \mathcal{R}^{m\times m}$. This means

that $\mathbf{V}$ plays no role in computing $g(\mathbf{w}_\rho, i)$ and hence we simply set $\mathbf{V}=\mathbf{I}$ in this section. Therefore

$$\mathbf{T}_\rho = \mathbf{T}_\rho(\mathbf{\Omega}) = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}, \tag{65}$$

are explored for a non-singular $\Omega_{\text{opt}} \in \mathcal{R}^{(m-2)\times(m-2)}$ such that $g(\mathbf{w}_\rho(\mathbf{T}_\rho(\mathbf{\Omega}_{\text{opt}})), i) \le \eta_{i'}$, $\forall i$. We can seek Ω_{opt} using a subgradient algorithm presented in Wu *et al.* (2005). The basic steps of this subgradient algorithm is listed here for completeness.

Initialization: Arbitrarily select a non-singular $\Omega \in \mathcal{R}^{(m-2)\times(m-2)}$ to obtain an initial point $\mathbf{w}_\rho(\mathbf{T}_\rho\ (\mathbf{\Omega}))$, set N to a sufficiently large integer and τ a small positive number, and set $N_t = 1$.

Step 1: Find out $e = \arg\max_{i\ \in\ \{1,\ldots,m+n\}} g(\mathbf{w}_\rho, i)$. If $g(\mathbf{w}_\rho, e) = \eta_{i'}$, which means that (30) holds, then $\Omega_{\text{opt}} = \Omega$ and terminate the routine. If $g(\mathbf{w}_\rho,\ e) > \eta_{i'}$ but $N_t \ge N$, which means that no saddle point is found after a large number of iterations, then the routine is also terminated for practical consideration.

Step 2: $\Omega = \Omega - \tau(\partial g(\mathbf{w}_\rho, e)/\partial\Omega)\|\partial g(\mathbf{w}_\rho, e)/\partial\Omega\|_F^{-1}$, $N_t = N_t + 1$, and go to Step 1.

Comment: When the routine does not find a saddle point, it still provides an excellent guess from which a direct numerical optimization algorithm can be used to find a (local) optimal solution. This is discussed in detail in Wu *et al.* (2005).

4. Optimal realization with the smallest dynamic range

In §3, we construct a controller realization $\mathbf{w}_{\rho\text{opt}} = \mathbf{w}_\rho(\mathbf{T}_\rho(\mathbf{\Omega}_{\text{opt}}))$ that achieves the minimum value of FWL closed-loop stability measure (21). Since the FWL stability measure (21) is concerned with the FWL error $\mathbf{\Delta}$ that depends only on the fraction bit length b_f, an optimal realization that minimizes this precision measure is not guaranteed to have a small dynamic range. In this section, we consider how to modify the optimal controller realization obtained in §3 to achieve the smallest dynamic range under the constraint that it remains to be a minimum solution of the optimization problem (22). From the discussion in §2, specifically, according to (19), $\|\mathbf{w}_\rho\|_M$ indicates the dynamic range of $\mathbf{w}_\rho$. Therefore, it is appropriate to use it as the dynamic range measure of a realization, that is,

$$d(\mathbf{w}_\rho) \triangleq \|\mathbf{w}_\rho\|_M. \tag{66}$$

Recalling the discussion on $\mathbf{V}$ in §3.2, it is straightforward to have the following theorem.

Theorem 6: *For two realizations* $\mathbf{w}_{\rho 1}$ *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* $\mathbf{V} \in \mathcal{R}^{\text{m}\times\text{m}}$ *such that*

$$\begin{aligned} &\mathbf{F}_{\rho 2} = \mathbf{V}^{-1}\mathbf{F}_{\rho 1}\mathbf{V}, \quad \mathbf{G}_{\rho 2} = \mathbf{V}^{-1}\mathbf{G}_{\rho 1}, \quad \mathbf{J}_{\rho 2} = \mathbf{J}_{\rho 1}\mathbf{V}, \\ &\mathbf{M}_{\rho 2} = \mathbf{M}_{\rho 1}, \quad \mathbf{H}_{\rho 2} = \mathbf{V}^{-1}\mathbf{H}_{\rho 1}, \end{aligned} \tag{67}$$

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

Given $\mathbf{w}_{\rho\text{opt}}$ (*that is* ($\mathbf{F}_{\rho\text{opt}}$, $\mathbf{G}_{\rho\text{opt}}$, $\mathbf{J}_{\rho\text{opt}}$, $\mathbf{M}_{\rho\text{opt}}$, $\mathbf{H}_{\rho\text{opt}}$)) *obtained in* §3, *define*

$$\begin{aligned} \mathcal{S}_{\rho\text{opt}} \triangleq \big\{&(\mathbf{F}_\rho, \mathbf{G}_\rho, \mathbf{J}_\rho, \mathbf{M}_\rho, \mathbf{H}_\rho) : \mathbf{F}_\rho = \mathbf{V}^{-1}\mathbf{F}_{\rho\text{opt}}\mathbf{V}, \\ &\mathbf{G}_\rho = \mathbf{V}^{-1}\mathbf{G}_{\rho\text{opt}}, \mathbf{J}_\rho = \mathbf{J}_{\rho\text{opt}}\mathbf{V}, \mathbf{M}_\rho = \mathbf{M}_{\rho\text{opt}}, \\ &\mathbf{H}_\rho = \mathbf{V}^{-1}\mathbf{H}_{\rho\text{opt}}, \mathbf{V} \in \mathcal{R}^{m\times m}, \mathbf{V}^T\mathbf{V} = \mathbf{I}\big\}. \end{aligned} \tag{68}$$

Denote the generic realization in $\mathcal{S}_{\rho\text{opt}}$ as $\mathbf{w}_{\rho\text{opt}}(\mathbf{V})$. It can be seen from Theorem 6 that, for any orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$, the realization $\mathbf{w}_{\rho\text{opt}}(\mathbf{V})$ remains to be a minimum solution of the optimization problem (22). Thus, we can search in $\mathcal{S}_{\rho\text{opt}}$ for an optimal realization with the smallest dynamic range. Formally, this is defined by the following optimization problem:

$$\mu \triangleq \min_{\substack{\mathbf{V}\ \in\ \mathcal{R}^{m\times m} \\ \mathbf{V}^T\mathbf{V}=I}} d(\mathbf{w}_{\rho\text{opt}}(\mathbf{V})). \tag{69}$$

In order to remove the constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ in the optimization problem (69), we derive a method for representing an orthogonal $\mathbf{V}$ parameterized by its independent parameters. Firstly, when $m = 2$, it is plain to see that any orthogonal $\mathbf{V}$ can be written as

$$V = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \kappa \end{bmatrix}, \quad \theta_1 \in [-\pi, \pi), \quad \kappa \in \{-1, 1\}. \tag{70}$$

Next, for $m = 3$, constructing an orthogonal $\mathbf{V}$ with its independent parameters can follow the following steps.

Step 1: Construct the first column $[v_{11}\ v_{21}\ v_{31}]^T$ of $\mathbf{V}$. Since $v_{11}^2 + v_{21}^2 + v_{31}^2 = 1$, we let

$$v_{11} = \cos\theta_1, \tag{71}$$

$$v_{21}^2 + v_{31}^2 = \sin^2\theta_1, \tag{72}$$

where $\theta_1 \in [-\pi, \pi)$. From (72), we further let

$$v_{21} = \cos\theta_2 \sin\theta_1, \quad v_{31} = \sin\theta_2 \sin\theta_1, \tag{73}$$

where $\theta_2 \in [-\pi, \pi)$. Thus the first column of $\mathbf{V}$ is defined by the two independent parameters as

$$\begin{bmatrix} v_{11} \\ v_{21} \\ v_{31} \end{bmatrix} = \begin{bmatrix} \cos\theta_1 \\ \cos\theta_2 \sin\theta_1 \\ \sin\theta_2 \sin\theta_1 \end{bmatrix}, \quad \theta_1, \theta_2 \in [-\pi, \pi), \tag{74}$$

which is an arbitrary unit vector in $\mathcal{R}^3$.

Step 2: Construct an orthonormal basis of the subspace $\mathcal{P}_0$ that is perpendicular to $[v_{11}\, v_{21}\, v_{31}]^T$.

Step 2.1: Construct the first column $[v_{12}\, v_{22}\, v_{32}]^T$ of the orthonormal basis.

(a) θ_1 is not equal to 0 or $-\pi$. Let $\mathcal{P}_1$ be the span of $[v_{11}\ \ v_{21}\ \ v_{31}]^T$ and $[1\ \ 0\ \ 0]^T$. Construct $[v_{12}\ v_{22}\ v_{32}]^T \in \mathcal{P}_1$ as a unit vector perpendicular to $[v_{11}\ v_{21}\ v_{31}]^T$, which means that

$$\begin{cases} \begin{bmatrix} v_{12} \\ v_{22} \\ v_{32} \end{bmatrix} = k_1 \begin{bmatrix} \cos\theta_1 \\ \cos\theta_2 \sin\theta_1 \\ \sin\theta_2 \sin\theta_1 \end{bmatrix} + k_2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \\ v_{12}^2 + v_{22}^2 + v_{32}^2 = 1, \\ v_{12}\cos\theta_1 + v_{22}\cos\theta_2 \sin\theta_1 + \ldots = 0. \end{cases} \tag{75}$$

Solving the above equations, we obtain

$$\begin{cases} k_1 = -\dfrac{\cos\theta_1}{\sin\theta_1}, \\ k_2 = \dfrac{1}{\sin\theta_1}, \end{cases} \tag{76}$$

or

$$\begin{cases} k_1 = \dfrac{\cos\theta_1}{\sin\theta_1}, \\ k_2 = -\dfrac{1}{\sin\theta_1}. \end{cases} \tag{77}$$

As only one orthonormal basis is needed, without the loss of generality, we adopt (77) and set

$$\begin{bmatrix} v_{12} \\ v_{22} \\ v_{32} \end{bmatrix} = \begin{bmatrix} -\sin\theta_1 \\ \cos\theta_2 \cos\theta_1 \\ \sin\theta_2 \cos\theta_1 \end{bmatrix}. \tag{78}$$

(b) $\theta_1 = 0$ or $\theta_1 = -\pi$. Since $[-\sin\theta_1 \quad \cos\theta_2\cos\theta_1 \quad \sin\theta_2\cos\theta_1]^T$ remains to be perpendicular to $[v_{11}\ \ v_{21}\ \ v_{31}]^T$, $[v_{12},\ v_{22},\ v_{32}]^T$ can always be constructed using (78).

Step 2.2: Construct the other column $[v_{13}\ v_{23}\ v_{33}]^T$ of the orthogonal basis. Denote $\mathcal{P}_2$ the span of $[v_{11}\ v_{21}\ v_{31}]^T$ and $[v_{12}\ v_{22}\ v_{32}]^T$. Obviously, $[v_{13}\ v_{23}\ v_{33}]^T$ is perpendicular to $\mathcal{P}_2$ and hence perpendicular to $[1\ 0\ 0]^T \in \mathcal{P}_2$. This means that $v_{13} = 0$ and $[v_{23}\ v_{33}]^T$ is perpendicular to both $[v_{21}\ \ v_{31}]^T$ and $[v_{22}\ \ v_{32}]^T$. Noting

$$\begin{bmatrix} v_{21} & v_{31} \end{bmatrix}^T = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 \end{bmatrix}^T \sin\theta_1 \tag{79}$$

and

$$\begin{bmatrix} v_{22} & v_{32} \end{bmatrix}^T = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 \end{bmatrix}^T \cos\theta_1, \tag{80}$$

we can see that $[v_{23}\ v_{23}]^T$ is the orthonormal basis of the subspace perpendicular to $[\cos\theta_2\ \sin\theta_2]^T$. From the formula (70) for the case of $m = 2$, we know that it can be chosen as

$$\begin{bmatrix} v_{23} & v_{33} \end{bmatrix}^T = \begin{bmatrix} -\sin\theta_2 & \cos\theta_2 \end{bmatrix}^T. \tag{81}$$

Step 3: Rotation of the orthonormal basis in $\mathcal{P}_0$. Now, an orthogonal matrix

$$\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 \\ \sin\theta_2\sin\theta_1 & \sin\theta_2\cos\theta_1 & \cos\theta_2 \end{bmatrix} \tag{82}$$

has been constructed. Its first column is arbitrary, but its second and third columns (the orthonormal basis of $\mathcal{P}_0$) are not arbitrary. In order to represent an arbitrary orthogonal $\mathbf{V} \in \mathcal{R}^{3\times 3}$, it is only needed to rotate the orthonormal basis in $\mathcal{P}_0$. This means that, from (70) and (82), we have

$$\begin{aligned} V = & \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 \\ \sin\theta_2\sin\theta_1 & \sin\theta_2\cos\theta_1 & \cos\theta_2 \end{bmatrix} \\ & \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & \sin\theta_3 & \cos\theta_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \kappa \end{bmatrix}, \\ & \theta_1, \theta_2, \theta_3 \in [-\pi, \pi), \kappa \in \{-1, 1\}. \end{aligned} \tag{83}$$

It should be clear that this rotation is achieved by applying a sequence of Givens rotations (in this case two Givens rotations), e.g. Delmas (1998).

In the similar way, the formula representing an arbitrary orthogonal $\mathbf{V} \in \mathcal{R}^{m\times m}$ with its independent parameters can be derived for $m > 3$. For example, the

formula for $m=4$ is given by

$$\mathbf{V}=\begin{bmatrix}\cos\theta_1 & -\sin\theta_1 & 0 & 0\\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 & 0\\ \cos\theta_3\sin\theta_2\sin\theta_1 & \cos\theta_3\sin\theta_2\cos\theta_1 & \cos\theta_3\cos\theta_2 & -\sin\theta_3\\ \sin\theta_3\sin\theta_2\sin\theta_1 & \sin\theta_3\sin\theta_2\cos\theta_1 & \sin\theta_3\cos\theta_2 & \cos\theta_3\end{bmatrix}$$
$$\times\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & \cos\theta_4 & -\sin\theta_4 & 0\\ 0 & \cos\theta_5\sin\theta_4 & \cos\theta_5\cos\theta_4 & -\sin\theta_5\\ 0 & \sin\theta_5\sin\theta_4 & \sin\theta_5\cos\theta_4 & \cos\theta_5\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & \cos\theta_6 & -\sin\theta_6\\ 0 & 0 & \sin\theta_6 & \cos\theta_6\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & \kappa\end{bmatrix},$$
$$\theta_1,\theta_2,\theta_3,\theta_4,\theta_5,\theta_6\in[-\pi,\pi),\quad \kappa\in\{-1,1\}. \tag{84}$$

Define

$$r=\frac{m(m-1)}{2}. \tag{85}$$

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

$$d\left(\mathbf{w}_{\rho\text{opt}}\left(\begin{bmatrix}\cos\theta_1 & -\sin\theta_1\\ \sin\theta_1 & \cos\theta_1\end{bmatrix}\right)\right)$$
$$=d\left(\mathbf{w}_{\rho\text{opt}}\left(\begin{bmatrix}\cos\theta_1 & -\sin\theta_1\\ \sin\theta_1 & \cos\theta_1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\right)\right), \tag{86}$$

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

$$d_1(\theta_1,\ldots,\theta_r)\triangleq d(\mathbf{w}_{\rho\text{opt}}(\mathbf{V})). \tag{87}$$

Then the optimization problem (69) is equivalent to the unconstrained optimization problem

$$\mu=\min_{\theta_1,\ldots,\theta_r\in[-\pi,\pi)} d_1(\theta_1,\ldots,\theta_r). \tag{88}$$

This kind of optimization problem can be solved using a numerical optimization algorithm that relies only on the function value to do search. With the optimal solution $\theta_{1\text{opt}},\ldots,\theta_{r\text{opt}}$, we can obtain the optimal orthogonal transformation $\mathbf{V}_{\text{opt}}$ and hence the optimal realization $\mathbf{w}_{\rho\text{opt1}}=\mathbf{w}_{\rho\text{opt}}(\mathbf{V}_{\text{opt}})$ of the smallest dynamic range.

4.1 *Comparison with direct optimization of a combined measure*

The proposed strategy has now been completely specified. In the first procedure, we solve the optimization problem (22) with an optimal solution $\mathbf{w}_{\rho\text{opt}}$. This realization achieves the minimum value of the FWL closed-loop stability measure defined in (21) but is not guaranteed to have a small dynamic range. In the second procedure, we solve the optimization problem (88) by a numerical means to obtain an optimal realization $\mathbf{w}_{\rho\text{opt1}}$ that has the smallest dynamic range over the set (68). Note that the set (68) contains all the orthogonal transformations of $\mathbf{w}_{\rho\text{opt}}$, and any realization in (68) is an optimal solution of the problem (22). This two-procedure approach is more effective than most of the previous works in this area, which only minimize the FWL stability measure (21) or some other similar measures by numerical means. It also becomes clear that the problem can be tackled by optimizing some combined criterion which include both the considerations for the precision or FWL stability and dynamic range of a controller realization. Define such a combined measure as (Wu *et al.* 2003)

$$\chi(\mathbf{w}_\rho)\triangleq f(\mathbf{w}_\rho)d(\mathbf{w}_\rho). \tag{89}$$

An optimal realization can be determined by minimizing $\chi(\mathbf{w}_\rho)$ over $\mathcal{S}_p$. This leads to the optimization problem

$$\varpi=\min_{\substack{\mathbf{T}_\rho\in\mathcal{R}^{m\times m}\\ \det\mathbf{T}_\rho\neq 0}}\chi(\mathbf{w}_\rho(\mathbf{T}_\rho)). \tag{90}$$

This optimization problem can be solved using a numerical optimization algorithm that uses the function value only to do search. A solution of this optimization problem is denoted by $\mathbf{w}_{\rho\text{opt2}}$.

A natural question to ask is which of the two solutions, $\mathbf{w}_{\rho\text{opt1}}$ or $\mathbf{w}_{\rho\text{opt2}}$, is better. It can easily be seen that the proposed two-procedure method in fact finds a Pareto optimal solution of the two-objective optimization problem with the two criteria $f(\mathbf{w}_\rho)$ and $d(\mathbf{w}_\rho)$. According to the multi-objective optimization theory (Pareto 1906, Zitzler and Thiele 1999), $\mathbf{w}_{\rho\text{opt1}}$ is preferred. Furthermore, note that the dimension of the search space for the optimization problem (90) is mm, and each parameter has the range $(-\infty,\infty)$. This should be compared with the optimization problem (88), where the search space has a dimension of $m(m-1)/2$ and each parameter has the range of $[-\pi,\pi)$. Also note that the optimization problem (90) is a constrained one,

although in practice the constraint $\det \mathbf{T}_\rho \neq 0$ is usually ignored during numerical search. It is obvious that the proposed two-procedure approach is computationally more attractive than this direct approach of minimizing the combined measure (89). Another potential drawback of direct minimizing $\chi(\mathbf{w}_\rho)$ numerically is that this is more prone to the problem of local minima, since the search space is much larger. One factor which makes the matter complicated is that the minimum bit length required to guarantee closed-loop stability does not have a simple linear relationship with $f(\mathbf{w}_\rho)$ and $d(\mathbf{w}_\rho)$. Note that $\mathbf{w}_{\rho\text{opt1}}$ minimizes the FWL closed-loop stability measure $f(\mathbf{w}_\rho)$, but this is not necessarily the case for $\mathbf{w}_{\rho\text{opt2}}$.

5. A design example

An example considered by Gevers and Li (1993) was used to illustrate the effectiveness of the proposed design procedure for obtaining optimal FWL fixed-point controller realizations and to compare the minimum bit lengths required to implement the optimal realizations with z operator and with δ operator of different h. The discrete-time plant model using z operator was given by

$$\mathbf{A}_z = \begin{bmatrix} 3.7156e+0 & -5.4143e+0 & 3.6525e+0 & -9.6420e-1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$
$$\mathbf{B}_z = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^T,$$
$$\mathbf{C}_z = \begin{bmatrix} 1.1160e-6 & 4.3000e-8 & 1.0880e-6 & 1.4000e-8 \end{bmatrix}.$$

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

$$\mathbf{F}_{z0} = \begin{bmatrix} 2.6743e+0 & -5.7446e+0 & 2.5101e+0 & -9.1782e-1 \\ 2.8769e-1 & -2.7446e-2 & -6.9444e-1 & -8.9358e-3 \\ -3.3773e-1 & 9.8699e-1 & -3.2925e-1 & -4.2367e-3 \\ -8.3021e-2 & -3.1988e-3 & 9.1906e-1 & -1.0415e-3 \end{bmatrix},$$
$$\mathbf{G}_{z0} = \begin{bmatrix} 1.0959e+6 & 6.3827e+5 & 3.0262e+5 & 7.4392e+4 \end{bmatrix}^T,$$
$$\mathbf{J}_{z0} = \begin{bmatrix} 1.8180e-1 & -2.8313e-1 & 5.0006e-2 & 6.1722e-2 \end{bmatrix},$$
$$\mathbf{M}_{z0} = 0, \quad \mathbf{H}_{z0} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

The procedure described in § 3 was then applied to obtain an optimal transformation matrix, which was given by

$$\mathbf{T}_{z\text{opt}} = \begin{bmatrix} -4.0558e+2 & -6.9295e+3 & -4.4853e+1 & 5.8411e+3 \\ -6.7105e+2 & -7.0344e+3 & -8.6317e+2 & 3.4389e+3 \\ -9.4359e+2 & -7.1314e+3 & -1.5943e+3 & 1.6526e+3 \\ -1.2230e+3 & -7.2202e+3 & -2.2845e+3 & 4.0879e+2 \end{bmatrix}.$$

The realization $\mathbf{w}_{\rho\text{opt}} = \mathbf{w}_z(\mathbf{T}_{z\text{opt}})$ calculated according to (14) was a global optimal realization in z operator that minimized the FWL closed-loop stability measure (21). In order to obtain an optimal realization in z operator with the smallest dynamic range, the optimization problem (88) was formed given the dimension $r = 6$. The MATLAB routine *fminsearch.m* was used to solve this optimization problem numerically, which yielded the solution

$$\theta_{1\text{opt}} = 4.3366e-1, \quad \theta_{2\text{opt}} = 2.1610e+0,$$
$$\theta_{3\text{opt}} = -2.2978e+0,$$
$$\theta_{4\text{opt}} = 1.6913e+0, \quad \theta_{5\text{opt}} = -3.0070e+0,$$
$$\theta_{6\text{opt}} = 1.8059e-1.$$

The global optimal realization with the smallest dynamic range, $\mathbf{w}_{z\text{opt1}} = \mathbf{w}_{z\text{opt}}(\mathbf{V}_{\text{opt}})$, was then calculated according to (84) and (68).

To see how robust a controller realization is to the FWL effect, the minimum bit length $b^{\min} = b_g^{\min} + b_f^{\min} + 1$ required to guarantee closed-loop stability can be examined. It is obvious that the minimum integer bit length $b_g^{\min}$ to avoid overflow for a realization $\mathbf{w}_\rho$ can directly be obtained by examining the elements of $\mathbf{w}_\rho$. The minimum fraction bit length $b_f^{\min}$ however can only be obtained through simulation. Starting from a very large b_f, we reduce b_f by one bit

Table 1. Comparison of various controller realizations using z operator.

Realization	$\chi(\mathbf{w}_z)$	$f(\mathbf{w}_z)$	$d(\mathbf{w}_z)$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
$\mathbf{w}_{z0}$	$4.3505e+12$	$3.9697e+6$	$1.0959e+6$	20	21	42
$\mathbf{w}_{z\text{opt}}$	$4.7700e+5$	$2.4246e+3$	$1.9673e+2$	8	8	17
$\mathbf{w}_{z\text{opt1}}$	$2.8608e+5$	$2.4246e+3$	$1.1799e+2$	8	7	16
$\mathbf{w}_{z\text{opt2}}$	$2.8078e+5$	$2.4411e+3$	$1.1502e+2$	10	7	18

and check closed-loop stability. The process is repeated until there appear closed-loop instability at $b_f = b_{fu}$. This gives $b_f^{\min} = b_{fu} + 1$. Table 1 lists the values of the FWL stability measure $f(\mathbf{w}_z)$, the dynamic range measure $d(\mathbf{w}_z)$ and the combined measure $\chi(\mathbf{w}_z)$ together with the related minimum bit lengths $b^{\min}$, $b_g^{\min}$ and $b_f^{\min}$ for the realizations $\mathbf{w}_{z0}$, $\mathbf{w}_{z\text{opt}}$ and $\mathbf{w}_{z\text{opt1}}$, respectively. It can be seen that the fixed-point implementation of $\mathbf{w}_{z0}$ needs at least 42 bits (20 fractional bits and 21 integer bits), while the implementation of $\mathbf{w}_{z\text{opt}}$ needs at least 17 bits (8 fractional bits and 8 integer bits). The latter achieved a reduction of 25 bits in the required bit length. It can also be seen that, as expected, $f(\mathbf{w}_{z\text{opt1}}) = f(\mathbf{w}_{z\text{opt}})$ but $d(\mathbf{w}_{z\text{opt1}})$ is smaller than $d(\mathbf{w}_{z\text{opt}})$, giving rise to further one bit reduction in $b_g^{\min}$ for $\mathbf{w}_{z\text{opt1}}$. Note that most of the existing FWL design methods, such as the one derived in Li (1998), can at the best hope to attain the realization $\mathbf{w}_{z\text{opt}}$. In fact, the method presented in Li (1998) may not always be able to achieve this optimal realization, as this method can generally attain a suboptimal solution, see Whidborne *et al*. (2000b). Thus the advantages of our proposed approach over these existing methods are selfevident.

For a comparison with the direct optimization approach (Wu *et al*. 2003), the optimization problem (90) was formed, and the MATLAB routine *fminsearch.m* was used to solve this $mm = 16$-dimensional search problem. Using $\mathbf{w}_{z0}$ as the initial realization, the solution obtained by this numerical search was found to be much worst than $\mathbf{w}_{z\text{opt}}$. This highlighted a difficulty with this approach of directly minimizing the combined measure (89). The search space had a much higher dimension and the solution obtained was sensitive to the initial condition. Using $\mathbf{w}_{z\text{opt1}}$ as the initial realization to form (90), the following optimal transformation matrix was obtained

$$\mathbf{T}_{z\text{opt2}} = \begin{bmatrix} -3.3536e+2 & -7.5296e+3 & 1.4101e+3 & 4.7942e+3 \\ 3.4834e+2 & -6.7222e+3 & -8.1255e+2 & 4.0422e+3 \\ 7.9174e+2 & -6.0888e+3 & -2.5691e+3 & 3.5754e+3 \\ 1.0231e+3 & -5.5884e+3 & -3.9358e+3 & 3.3879e+3 \end{bmatrix}$$

which produced a corresponding optimal realization $\mathbf{w}_{z\text{opt2}}$. The values of various measures and related minimum bit lengths for $\mathbf{w}_{z\text{opt2}}$ are also listed in table 1. As expected, $\chi(\mathbf{w}_{z\text{opt2}}) < \chi(\mathbf{w}_{z\text{opt1}})$ but $f(\mathbf{w}_{z\text{opt2}}) > f(\mathbf{w}_{z\text{opt1}})$. Although $\mathbf{w}_{z\text{opt2}}$ has a smaller dynamic range than $\mathbf{w}_{z\text{opt1}}$, the amount of reduction is not enough to produce one-bit reduction in $b_g^{\min}$ for $\mathbf{w}_{z\text{opt2}}$. Also note that, although $f(\mathbf{w}_\rho)$ is linked to $b_f^{\min}$, the relationship is not a simple one. This is reflected in the result that $\mathbf{w}_{z\text{opt2}}$ requires two more bits in $b_f^{\min}$, compared with $\mathbf{w}_{z\text{opt1}}$. In this case, the proposed two-procedure approach was able to obtain a better realization $\mathbf{w}_{z\text{opt1}}$, in comparison with the direct optimization approach.

It is obvious that any realization $\mathbf{w}_\rho \in \mathcal{S}_\rho$ implemented in infinite precision will achieve exactly the same set of closed-loop eigenvalues as the infinite-precision implemented $\mathbf{w}_{\rho 0}$, which is the designed closed-loop eigenvalues. For this reason, the infinite-precision implemented $\mathbf{w}_{z0}$ is referred to as the ideal realization $\mathbf{w}_{z\text{ideal}}$. Figure 2 compares the designed eigenvalues of the closed-loop system using $\mathbf{w}_{z\text{ideal}}$ with those of the 16-bit (8 integer bits and 7 fractional bits) implemented $\mathbf{w}_{z\text{opt}}$, 16-bit (7 integer bits and 8 fractional bits) implemented $\mathbf{w}_{z\text{opt1}}$, and 16-bit (7 integer bits and 8 fractional bits) implemented $\mathbf{w}_{z\text{opt2}}$. Confirming the results of table 1, figure 2 shows that the closed-loop system with the 16-bit implemented $\mathbf{w}_{z\text{opt1}}$ is stable while the system with the 16-bit implemented $\mathbf{w}_{z\text{opt}}$ or $\mathbf{w}_{z\text{opt2}}$ is unstable.

Similarly, the optimal realization 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 was

$$\mathbf{A}_\delta = \begin{bmatrix} 4.4492e+4 & -8.8708e+4 & 5.9843e+4 & -1.5797e+4 \\ 1.6384e+4 & -1.6384e+4 & 0 & 0 \\ 0 & 1.6384e+4 & -1.6384e+4 & 0 \\ 0 & 0 & 1.6384e+4 & -1.6384e+4 \end{bmatrix},$$
$$\mathbf{B}_\delta = \begin{bmatrix} 1.6384e+4 & 0 & 0 & 0 \end{bmatrix}^T,$$
$$\mathbf{C}_\delta = \begin{bmatrix} 1.1160e-6 & 4.3000e-8 & 1.0880e-6 & 1.4000e-8 \end{bmatrix}.$$

The initial realization of the digital controller using the δ operator with $h = 2^{-14}$ was

$$\mathbf{F}_{\delta 0} = \begin{bmatrix} 2.7432e+4 & -9.4119e+4 & 4.1126e+4 & -1.5038e+4 \\ 4.7135e+3 & -1.6834e+4 & -1.1378e+4 & -1.4640e+2 \\ -5.5333e+3 & 1.6171e+4 & -2.1778e+4 & -6.9414e+1 \\ -1.3602e+3 & -5.2410e+1 & 1.5058e+4 & -1.6401e+4 \end{bmatrix},$$
$$\mathbf{G}_{\delta 0} = \begin{bmatrix} 1.7956e+10 & 1.0457e+10 & 4.9582e+9 & 1.2188e+9 \end{bmatrix}^T,$$
$$\mathbf{J}_{\delta 0} = \begin{bmatrix} 1.8180e-1 & -2.8313e-1 & 5.0006e-2 & 6.1722e-2 \end{bmatrix},$$
$$\mathbf{M}_{\delta 0} = 0, \quad H_{\delta 0} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

The procedure of § 3 was applied, which obtained

$$\mathbf{T}_{\delta \mathrm{opt}} = \begin{bmatrix} -5.1914e+4 & -8.8698e+5 & -5.7412e+3 & 7.4766e+5 \\ -8.5895e+4 & -9.0040e+5 & -1.1049e+5 & 4.4017e+5 \\ -1.2078e+5 & -9.1282e+5 & -2.0407e+5 & 2.1153e+5 \\ -1.5654e+5 & -9.2419e+5 & -2.9242e+5 & 5.2325e+4 \end{bmatrix}.$$

The controller realization $\mathbf{w}_{\delta \mathrm{opt}} = \mathbf{w}_\delta(\mathbf{T}_{\delta \mathrm{opt}})$ calculated by (14) was a global optimal realization in δ operator that minimized the FWL closed-loop stability measure (21). The optimization problem (88) was next formed, and the MATLAB routine *fminsearch.m* yielded the solution

$$\theta_{1\mathrm{opt}} = 8.0159e-1, \quad \theta_{2\mathrm{opt}} = 2.9926e+0,$$
$$\theta_{3\mathrm{opt}} = -5.5715e-2,$$
$$\theta_{4\mathrm{opt}} = 2.8273e+0, \quad \theta_{5\mathrm{opt}} = -9.7594e-1,$$
$$\theta_{6\mathrm{opt}} = -6.8654e-2.$$

The global optimal realization with the smallest dynamic range, $\mathbf{w}_{\delta \mathrm{opt}1} = \mathbf{w}_{\delta \mathrm{opt}}(\mathbf{V}_{\mathrm{opt}})$, was readily calculated according to (84) and (68). The optimization problem (90) was also formed using $\mathbf{w}_{\delta \mathrm{opt}1}$ as the initial realization, and the MATLAB routine *fminsearch.m* produced the solution

$$\mathbf{T}_{\delta \mathrm{opt}2} = \begin{bmatrix} 5.7439e+5 & -7.7021e+5 & 2.9197e+5 & 5.9710e+5 \\ 5.5729e+5 & -6.8978e+5 & -3.9745e+4 & 4.6427e+5 \\ 5.3563e+5 & -6.3592e+5 & -2.9971e+5 & 3.7509e+5 \\ 5.0873e+5 & -6.0416e+5 & -5.0002e+5 & 3.2746e+5 \end{bmatrix},$$

which yielded the corresponding optimal realization $\mathbf{w}_{\delta \mathrm{opt}2}$.

Table 2 compares the values of various measures and related minimum bit lengths for the four controller realizations $\mathbf{w}_{\delta 0}$, $\mathbf{w}_{\delta \mathrm{opt}}$, $\mathbf{w}_{\delta \mathrm{opt}1}$ and $\mathbf{w}_{\delta \mathrm{opt}2}$ with $h = 2^{-14}$. Note that 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 $\|\mathbf{\Delta}\|_M$, defined in (20), 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. From table 2, it can be seen that the fixed-point implementation of $\mathbf{w}_{\delta 0}$ needs at least 51 bits (15 fractional bits and 35 integer bits) while the implementation of $\mathbf{w}_{\delta \mathrm{opt}}$ requires at least 13 bits (−4 fractional bits and 16 integer bits). It can also be seen that $\mathbf{w}_{\delta \mathrm{opt1}}$ and $\mathbf{w}_{\delta \mathrm{opt2}}$ give further one bit reduction in $b_g^{\min}$, compared with $\mathbf{w}_{\delta \mathrm{opt}}$. In this case, the two different realization $\mathbf{w}_{\delta \mathrm{opt1}}$ and $\mathbf{w}_{\delta \mathrm{opt2}}$ seem to have

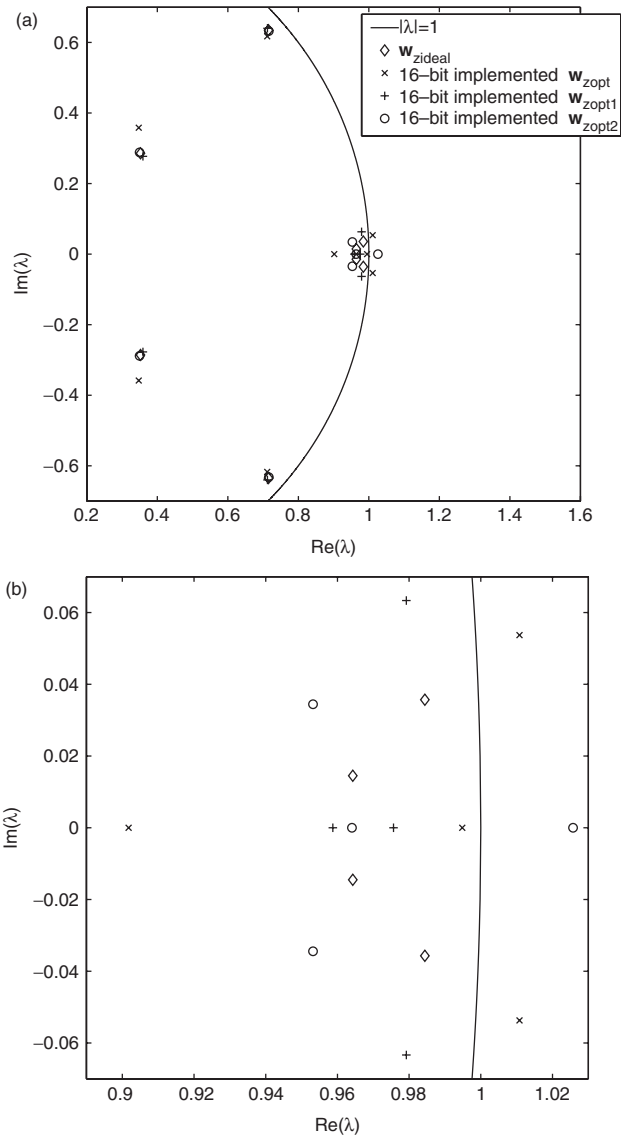

Figure 2. Closed-loop eigenvalues of $\mathbf{w}_{z\mathrm{ideal}}$, 16-bit implemented $\mathbf{w}_{z\mathrm{opt}}$, 16-bit implemented $\mathbf{w}_{z\mathrm{opt1}}$ and 16-bit implemented $\mathbf{w}_{z\mathrm{opt2}}$: (a) full plot (b) part of plot.

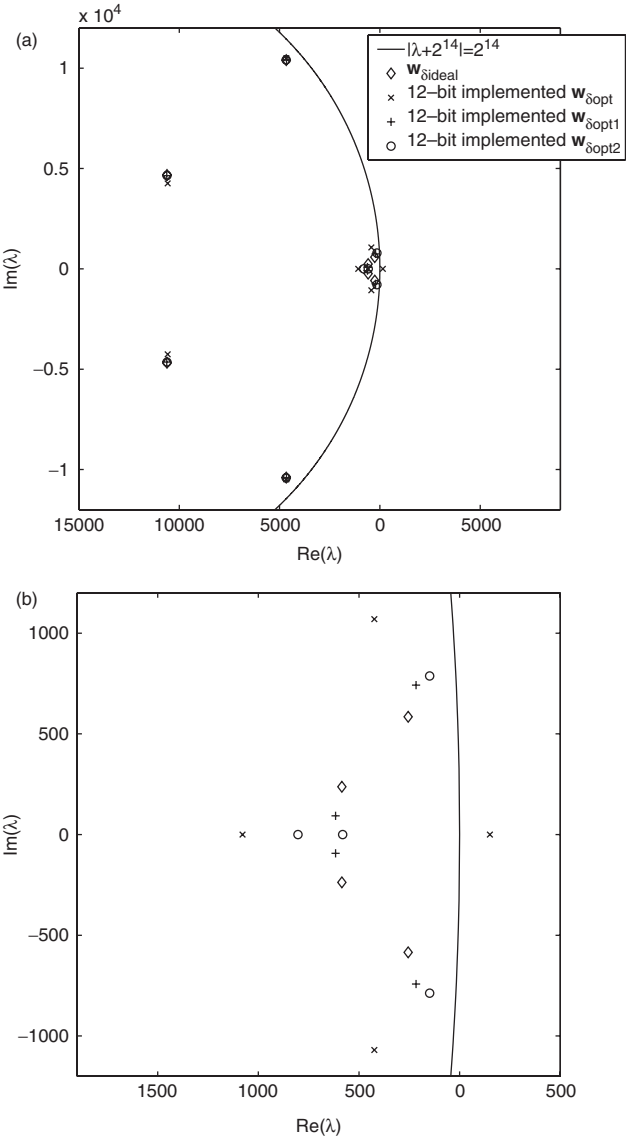

Figure 3. Closed-loop eigenvalues of $\mathbf{w}_{\delta\mathrm{ideal}}$, 12-bit implemented $\mathbf{w}_{\delta\mathrm{opt}}$, 12-bit implemented $\mathbf{w}_{\delta\mathrm{opt1}}$ and 12-bit implemented $\mathbf{w}_{\delta\mathrm{opt2}}$, given $h = 2^{-14}$: (a) full plot (b) part of plot.

Table 2. Comparison of various controller realizations using δ operator with $h = 2^{-14}$.

Realization	$\chi(\mathbf{w}_\delta)$	$f(\mathbf{w}_\delta)$	$d(\mathbf{w}_\delta)$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
$\mathbf{w}_{\delta 0}$	$4.9759e+15$	$2.7712e+5$	$1.7956e+10$	15	35	51
$\mathbf{w}_{\delta\mathrm{opt}}$	$1.7287e+4$	$3.3740e-1$	$5.1236e+4$	−4	16	13
$\mathbf{w}_{\delta\mathrm{opt1}}$	$8.7084e+3$	$3.3740e-1$	$2.5810e+4$	−4	15	12
$\mathbf{w}_{\delta\mathrm{opt2}}$	$7.3903e+3$	$3.4474e-1$	$2.1437e+4$	−4	15	12

Table 3. Comparison of $\mathbf{w}_{\delta\text{opt1}}$ under different h.

h	$f(\mathbf{w}_{\delta\text{opt1}})$	$d(\mathbf{w}_{\delta\text{opt1}})$	$b_f^{\min}(\mathbf{w}_{\delta\text{opt1}})$	$b_g^{\min}(\mathbf{w}_{\delta\text{opt1}})$	$b^{\min}(\mathbf{w}_{\delta\text{opt1}})$
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

similar robustness to the FWL error. Figure 3 compares the closed-loop eigenvalues of $\mathbf{w}_{\delta\text{ideal}}$, the infinite-precision implemented $\mathbf{w}_{\delta 0}$, with those of the 12-bit (−5 fractional bits and 16 integer bits) implemented $\mathbf{w}_{\delta\text{opt}}$, the 12-bit (−4 fractional bits and 15 integer bits) implemented $\mathbf{w}_{\delta\text{opt1}}$ and the 12-bit (−4 fractional bits and 15 integer bits) implemented $\mathbf{w}_{\delta\text{opt2}}$. As expected, the closed-loop system with the 12-bit implemented $\mathbf{w}_{\delta\text{opt1}}$ or $\mathbf{w}_{\delta\text{opt2}}$ is stable, but the closed-loop system with the 12-bit implemented $\mathbf{w}_{\delta\text{opt}}$ is unstable.

Table 3 compares the values of the FWL stability measure $f(\mathbf{w}_{\delta\text{opt1}})$ and the dynamic range measure $d(\mathbf{w}_{\delta\text{opt1}})$ together with the related minimum bit lengths for the controller realization $\mathbf{w}_{\delta\text{opt1}}$, giving $h = 2^{10} \sim 2^{-25}$. Comparing tables 1 and 3, it is 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 (6). 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 realizations $\mathbf{w}_{\delta\text{opt1}}$ achieve the maximum robustness to the FWL errors.

6. Conclusions

A novel two-procedure approach has been developed to design optimal fixed-point realizations of digital controllers with FWL considerations. The proposed strategy first finds an optimal controller realization by minimizing an FWL closed-loop stability measure. The fixed-point implementation of this realization thus requires a minimum fractional bit length to guarantee closed-loop stability. This realization is then modified via an effective numerical optimization to produce an optimal realization with the smallest dynamic range without sacrificing FWL closed-loop stability robustness. The final optimal realization thus also requires a minimum integer bit length to avoid overflow and consequently it needs 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 parameterizations of a generic controller structure. A design example has demonstrated that the proposed method provides an effective design procedure for obtaining optimal controller realizations 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 realization has much better FWL closed-loop stability characteristics than the optimal shift-operator controller realization.

Acknowledgements

J. Wu and J. Chu wish to thank the support of the National Natural Science Foundation of China (Grants No. 60374002 and No. 60421002), 973 program of China (Grant No. 2002CB312200) and program for New Century Excellent Talents in University (NCET-04-0547). S. Chen wish to thank the support of the United Kingdom Royal Academy of Engineering.

References

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

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*, 40, pp. 153–163, 2000.

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*, 45, pp. 321–329, 2002.

M.A. Dahleh and I.J. Diaz-Bobillo, *Control of Uncertain Systems: A linear Programming Approach*, Englewood Cliffs, NJ: Prentice Hall, 1995.

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

M.C. De Oliveira and R.E. Skelton, "State feedback control of linear systems in the presence of devices with finite signal-to-noise ratio", *Int. J. Control*, 74, pp. 1501–1509, 2001.

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

I.J. Fialho and T.T. Georgiou, "Computational algorithms for sparse optimal digital controller realizations", in *Digital Controller Implementation and Fragility: A modern Perspective*, R.S.H. Istepanian and J.F. Whidborne, Eds., London: Springer Verlag, 2001, pp. 105–121.

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

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

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

L.H. Keel and S.P. Bhattacharryya, "Robust, fragile, or optimal?", *IEEE Trans. Automatic Control*, 42, pp. 1098–1105, 1997.

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

K. Liu, R.E. Skelton and K. Grigoriadis, "Optimal controllers for finite wordlength implementation", *IEEE Trans. Automatic Control*, 37, pp. 294–1304, 1992.

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

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

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

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

V. Pareto, *Manuale di Economia politica*, Milan, Italy: Societa Editrice Libraria, 1906.

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

J.F. Whidborne, J. Wu and R.S.H. Istepanian, "Finite word length stability issues in an l1 framework", 73, pp. 166–176, 2000a.

J.F. Whidborne, J. Wu, R.S.H. Istepanian and J. Chu, "Comments on On the structure of digital controllers with finite word length consideration", *IEEE Trans. Automatic Control*, 45, pp. 344–344, 2000b.

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

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*, 147, pp. 664–672, 2000.

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*, 46, pp. 1162–1166, 2001.

J. Wu, S. Chen, G. Li and J. Chu, "Global optimal realizations of finite precision digital controllers", in *Proc. 41st IEEE Conf. Decision and Control*, Las Vegas, USA, Dec. 10–13, pp. 2941–2946, 2002.

J. Wu, S. Chen, J.F. Whidborne and J. Chu, "A unified close-loop stability measure for finite-precision digital controller realizations implemented in different representation schemes", *IEEE Trans. Automatic Control*, 48, pp. 816–822, 2003.

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*, 44, pp. 1787–1810, 2005.

K. Zhou, J.C. Doyle and K. Glover, *Robust Optimal Control*, Englewood Cliffs, NJ: Prentice Hall, 1996.

E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach", *IEEE Trans. Evolutionary Computation*, 3, pp. 257–271, 1999.