Proceedings of the 6th World Congress on Control
and Automation, June 21 - 23, 2006, Dalian, China

A New Approach for Finite-Precision Controller Realizations

Jun Wu and Weihua Xu
National Lab. Ind. Contr. Tech.
Ins. Adv. Process Contr.
Zhejiang Univ., Hangzhou, China
{jwu & whxu}@iipc.zju.edu.cn

Sheng Chen
School Elec. & Computer Sci.
University of Southampton
Highfield, Southampton SO17 1BJ, U.K.
sqc@ecs.soton.ac.uk

Gang Li
School Electri. & Elec. Eng.
Nanyang Tech. Univ.
Singapore
egli@ntu.edu.sg

Abstract - **A new approach is proposed to design optimal finite word length (FWL) realizations of digital controllers implemented in fixed-point arithmetic. An analytical method is first formulated to obtain a global optimal controller realization that optimizes an FWL closed-loop stability measure. A dynamic range measure is next derived for the implemented controller realization, and a numerical optimization method is developed to make the controller realization having the smallest dynamic range without sacrificing any FWL closed-loop stability robustness.**

Index Terms - Finite word length, digital control, optimization

I. Introduction

Due to the finite word length (FWL) effect, a fixed-point controller implementation may degrade the designed closed-loop performance or even destabilize the designed stable closed-loop system, if the controller implementation structure is not carefully chosen. There exist an infinite number of different realizations corresponding to a control law. 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. All the previous FWL closed-loop measures [1]–[5] 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 affect the integer part or dynamic range of fixed-point representation. This paper proposes a novel approach for designing optimal fixed-point controller realizations by simultaneously optimizing both a precision or FWL closed-loop stability measure and a dynamic range measure.

II. 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 $\Upsilon(\mathbf{z}) \overset{\triangle}{=} [\Re(\mathbf{z}) \ \Im(\mathbf{z})]$, where $\Re(\mathbf{z})$ and $\Im(\mathbf{z})$ denote the real and the imaginary parts of $\mathbf{z}$, respectively. For $\mathbf{U} \in \mathcal{C}^{m\times n}$ with elements u_{ij}, define

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

Let $\mathrm{Vec}(\cdot)$ be the column stacking operator such that $\mathrm{Vec}(\mathbf{U})$ is an mn-dimensional vector. As usual, $\mathbf{U}^T$ is the transposed matrix of $\mathbf{U}$, $\mathbf{U}^H$ is the Hermitian adjoint matrix of $\mathbf{U}$, and $\mathbf{U}^*$ is conjugate to $\mathbf{U}$. For a 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 $\lambda_i(\mathbf{M})$. If $\mathbf{M}$ is diagonalizable, the matrix $\mathbf{M}_x \overset{\triangle}{=} [\mathbf{x}_1(\mathbf{M}) \ \cdots \ \mathbf{x}_n(\mathbf{M})]$ is invertible. Define $\mathbf{M}_y = [\mathbf{y}_1(\mathbf{M}) \ \cdots \ \mathbf{y}_n(\mathbf{M})] \overset{\triangle}{=} \mathbf{M}_x^{-H}$, where $\mathbf{y}_i(\mathbf{M})$ is called the reciprocal left eigenvector corresponding to $\mathbf{x}_i(\mathbf{M})$.

Consider the discrete-time closed-loop control system consisting of a plant $\hat{P}$ and a controller $\hat{C}$. The plant $\hat{P}$ is described by the state-space description

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases} \qquad (2)$$

with $\mathbf{A} \in \mathcal{R}^{n\times n}$, $\mathbf{B} \in \mathcal{R}^{n\times p}$ and $\mathbf{C} \in \mathcal{R}^{q\times n}$. The generic controller $\hat{C}$ is described by the state-space description

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{H}\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases} \qquad (3)$$

with $\mathbf{F} \in \mathcal{R}^{m\times m}$, $\mathbf{G} \in \mathcal{R}^{m\times q}$, $\mathbf{J} \in \mathcal{R}^{p\times m}$, $\mathbf{M} \in \mathcal{R}^{p\times q}$ and $\mathbf{H} \in \mathcal{R}^{m\times p}$. Let $\mathbf{e}(k) = \mathbf{r}(k) + \mathbf{u}(k)$ where $\mathbf{r}(k)$ is the exogenous input. Then $\hat{P}$ and $\hat{C}$ form the closed-loop control system. The state-space descriptions or realizations $(\mathbf{F}, \mathbf{G}, \mathbf{J},\ \mathbf{M}, \mathbf{H})$ of the controller $\hat{C}$ are not unique. If $(\mathbf{F}_0, \mathbf{G}_0, \mathbf{J}_0, \mathbf{M}_0, \mathbf{H}_0)$ is a realization of $\hat{C}$ that has been designed using a standard controller design procedure, all the realizations of $\hat{C}$ form a realization set

$$\begin{aligned} \mathcal{S} \ \overset{\triangle}{=} \ & \{(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0, \\ & \mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0, \mathbf{H} = \mathbf{T}^{-1}\mathbf{H}_0\} \end{aligned} \qquad (4)$$

where $\mathbf{T} \in \mathcal{R}^{m\times m}$ is any real-valued nonsingular matrix, called a transformation. Define

$$\mathbf{w} \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}) \\ \mathrm{Vec}(\mathbf{G}) \\ \mathrm{Vec}(\mathbf{J}) \\ \mathrm{Vec}(\mathbf{M}) \\ \mathrm{Vec}(\mathbf{H}) \end{bmatrix}, \quad \mathbf{w}_0 \triangleq \begin{bmatrix} \mathrm{Vec}(\mathbf{F}_0) \\ \mathrm{Vec}(\mathbf{G}_0) \\ \mathrm{Vec}(\mathbf{J}_0) \\ \mathrm{Vec}(\mathbf{M}_0) \\ \mathrm{Vec}(\mathbf{H}_0) \end{bmatrix}. \tag{5}$$

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

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{w}) &= \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC}+\mathbf{HMC} & \mathbf{F}+\mathbf{HJ} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_0) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \end{aligned} \tag{6}$$

where $\mathbf{I}$ and $\mathbf{0}$ denote the identity and zero matrices of appropriate dimensions, respectively. As the closed-loop system is designed to be stable, $1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| > 0$, $\forall i \in \{1, \cdots, m+n\}$, which implies that all the realizations $\mathbf{w} \in \mathcal{S}$ have the same set of the closed-loop eigenvalues if they are implemented with infinite precision.

When $\mathbf{w}$ 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 for the fractional part of the representation. To avoid overflow, b_g should be sufficiently large such that

$$\|\mathbf{w}\|_M \le 2^{b_g}. \tag{7}$$

$\|\mathbf{w}\|_M$ represents the dynamic range of $\mathbf{w}$ in fixed-point format. Even without overflow, $\mathbf{w}$ is perturbed into $\mathbf{w} + \boldsymbol{\Delta}$ due to the finite b_f bits in the fractional part representation. Each element of $\boldsymbol{\Delta}$ is bounded by $\pm 2^{-(b_f+1)}$, i.e. $\|\boldsymbol{\Delta}\|_M \le 2^{-(b_f+1)}$. With the perturbation $\boldsymbol{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}+\boldsymbol{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{w}+\boldsymbol{\Delta})$ crosses over the stability boundary, the closed-loop system becomes unstable. Under the condition of no overflow, it can be seen that the closed-loop stability depends only on the fractional part representation. Finding an optimal realization with maximum FWL closed-loop stability robustness however is a multi-objective optimization. Firstly, an optimal realization should optimize some FWL closed-loop stability measure, whose value 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 largest b_f bits to achieve the highest possible precision. We propose an efficient two-step approach to tackle this multi-objective problem.

III. Optimizing an FWL stability measure

We shall use λ_i to replace $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$ when doing so does not cause ambiguity. Under the condition of no overflow, how easily the FWL error $\boldsymbol{\Delta}$ can cause a stable control system to become unstable is determined by how large 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 [1], is considered in this study

$$f(\mathbf{w}) \triangleq \max_{i\in\{1,\cdots,m+n\}} \frac{\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}} \right\|_F}{1 - |\lambda_i|}. \tag{8}$$

It is natural to search for "optimal" controller realizations that minimize the measure defined in (8). This leads to the following optimal FWL controller realization problem

$$\upsilon \triangleq \min_{\mathbf{w}\in\mathcal{S}} f(\mathbf{w}). \tag{9}$$

Given the realization $\mathbf{w}_0$, from the definition of $\mathcal{S}$ (4), $\mathbf{w}$ depends on the transformation matrix $\mathbf{T}$. Thus the optimization problem (9) is equivalent to

$$\upsilon = \min_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det \mathbf{T}\neq 0}} f(\mathbf{w}(\mathbf{T})). \tag{10}$$

We have developed an analytical global optimal solution for the optimization problem (10), which is outlined here.

A. Optimizing single-pole FWL stability measure

Define the following function linked to λ_i

$$g(\mathbf{w}, i) \triangleq \frac{\left\| \frac{\partial \lambda_i}{\partial \mathbf{w}} \right\|_F}{1 - |\lambda_i|} \tag{11}$$

and the single-pole FWL stability measure related to λ_i as

$$\eta_i \triangleq \min_{\mathbf{w}\in\mathcal{S}} g(\mathbf{w}, i) = \min_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det \mathbf{T}\neq 0}} g(\mathbf{w}(\mathbf{T}), i). \tag{12}$$

It is easy to show that $\upsilon \ge \max_{i\in\{1,\cdots,m+n\}} \eta_i$. Thus the maximum of all the single-pole measures provides a lower bound of the optimal value υ. To attain the single-pole measure η_i for the eigenvalue λ_i is equivalent to solve the minimization problem of the single-pole sensitivity

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det \mathbf{T}\neq 0}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{w}} \right\|_F. \tag{13}$$

Lemma 1: (See [1]) Let the square matrix $\mathbf{Z} = \mathbf{M}_1 + \mathbf{M}_2\mathbf{X}\mathbf{M}_3$ be diagonalizable where the real-valued matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ have appropriate dimensions and are independent of the real-valued matrix $\mathbf{X}$. Then

$$\frac{\partial \lambda_i(\mathbf{Z})}{\partial \mathbf{X}} = \mathbf{M}_2^T \mathbf{y}_i^*(\mathbf{Z}) \mathbf{x}_i^T(\mathbf{Z}) \mathbf{M}_3^T. \tag{14}$$

From (6), it can be seen that

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC}+\mathbf{HMC} & \mathbf{HJ} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{F} \left[\mathbf{0} \quad \mathbf{I} \right], \tag{15}$$

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{BJ} \\ \mathbf{HMC} & \mathbf{F}+\mathbf{HJ} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{G} \left[\mathbf{C} \quad \mathbf{0} \right], \tag{16}$$

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{0} \\ \mathbf{GC}+\mathbf{HMC} & \mathbf{F} \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{H} \end{bmatrix} \mathbf{J} \left[\mathbf{0} \quad \mathbf{I} \right], \tag{17}$$

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A} & \mathbf{BJ} \\ \mathbf{GC} & \mathbf{F}+\mathbf{HJ} \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{H} \end{bmatrix} \mathbf{M} \left[\mathbf{C} \quad \mathbf{0} \right], \tag{18}$$

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC} & \mathbf{F} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{H} \left[\mathbf{MC} \quad \mathbf{J} \right]. \tag{19}$$

$\forall i \in \{1,\cdots,m+n\}$, partition the eigenvectors of $\overline{\mathbf{A}}(\mathbf{w}_0)$, $\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_0))$ and $\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_0))$, into

$$\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w}_0)) = \begin{bmatrix} \mathbf{x}_{i,1} \\ \mathbf{x}_{i,2} \end{bmatrix}, \quad \mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_0)) = \begin{bmatrix} \mathbf{y}_{i,1} \\ \mathbf{y}_{i,2} \end{bmatrix}, \qquad (20)$$

where $\mathbf{x}_{i,1}, \mathbf{y}_{i,1} \in \mathcal{C}^n$ and $\mathbf{x}_{i,2}, \mathbf{y}_{i,2} \in \mathcal{C}^m$. It is easy to see from (6) that, $\forall i \in \{1,\cdots,m+n\}$,

$$\mathbf{x}_i(\overline{\mathbf{A}}(\mathbf{w})) = \begin{bmatrix} \mathbf{x}_{i,1} \\ \mathbf{T}^{-1}\mathbf{x}_{i,2} \end{bmatrix}, \; \mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w})) = \begin{bmatrix} \mathbf{y}_{i,1} \\ \mathbf{T}^T\mathbf{y}_{i,2} \end{bmatrix}. \qquad (21)$$

Applying Lemma 1 and (21) to (15)–(19) results in

$$\frac{\partial \lambda_i}{\partial \mathbf{F}} = \mathbf{T}^T\mathbf{y}_{i,2}^*\mathbf{x}_{i,2}^T\mathbf{T}^{-T}, \quad \frac{\partial \lambda_i}{\partial \mathbf{G}} = \mathbf{T}^T\mathbf{y}_{i,2}^*\mathbf{x}_{i,1}^T\mathbf{C}^T, \qquad (22)$$

$$\frac{\partial \lambda_i}{\partial \mathbf{J}} = \left(\mathbf{B}^T\mathbf{y}_{i,1}^* + \mathbf{H}_0^T\mathbf{y}_{i,2}^*\right)\mathbf{x}_{i,2}^T\mathbf{T}^{-T}, \qquad (23)$$

$$\frac{\partial \lambda_i}{\partial \mathbf{M}} = \left(\mathbf{B}^T\mathbf{y}_{i,1}^* + \mathbf{H}_0^T\mathbf{y}_{i,2}^*\right)\mathbf{x}_{i,1}^T\mathbf{C}^T, \qquad (24)$$

$$\frac{\partial \lambda_i}{\partial \mathbf{H}} = \mathbf{T}^T\mathbf{y}_{i,2}^*\left(\mathbf{x}_{i,1}^T\mathbf{C}^T\mathbf{M}_0^T + \mathbf{x}_{i,2}^T\mathbf{J}_0^T\right). \qquad (25)$$

Let

$$\alpha_i^2 \triangleq \|\mathbf{C}\mathbf{x}_{i,1}\|_F^2 + \|\mathbf{M}_0\mathbf{C}\mathbf{x}_{i,1} + \mathbf{J}_0\mathbf{x}_{i,2}\|_F^2, \qquad (26)$$

$$\beta_i^2 \triangleq \|\mathbf{B}^T\mathbf{y}_{i,1} + \mathbf{H}_0^T\mathbf{y}_{i,2}\|_F^2, \qquad (27)$$

$$\tau_i^2 \triangleq \|\mathbf{B}^T\mathbf{y}_{i,1} + \mathbf{H}_0^T\mathbf{y}_{i,2}\|_F^2\|\mathbf{C}\mathbf{x}_{i,1}\|_F^2. \qquad (28)$$

Then

$$\begin{aligned} \left\|\frac{\partial \lambda_i}{\partial \mathbf{w}}\right\|_F^2 &= \|\mathbf{T}^{-1}\mathbf{x}_{i,2}\|_F^2\|\mathbf{T}^T\mathbf{y}_{i,2}\|_F^2 + \alpha_i^2\|\mathbf{T}^T\mathbf{y}_{i,2}\|_F^2 \\ &\quad + \beta_i^2\|\mathbf{T}^{-1}\mathbf{x}_{i,2}\|_F^2 + \tau_i^2. \end{aligned} \qquad (29)$$

In order to attain η_i, one needs to minimize the function

$$\begin{aligned} \xi(\mathbf{T},\alpha,\beta,\mathbf{q},\mathbf{z}) &\triangleq \|\mathbf{T}^{-1}\mathbf{q}\|_F^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + \alpha^2\|\mathbf{T}^T\mathbf{z}\|_F^2 \\ &\quad + \beta^2\|\mathbf{T}^{-1}\mathbf{q}\|_F^2 \end{aligned} \qquad (30)$$

where nonsingular $\mathbf{T} \in \mathcal{R}^{m\times m}$, positive $\alpha, \beta \in \mathcal{R}$, and $\mathbf{q}, \mathbf{z} \in \mathcal{C}^m$ are nonzero vectors. For the different cases of $\mathbf{q}$ and $\mathbf{z}$, the results on minimizing $\xi(\mathbf{T},\alpha,\beta,\mathbf{q},\mathbf{z})$ are given in [6]. Based on these results, all the solutions to (12) can be specified. The following Theorem lists the result for one case of $\mathbf{q}$ and $\mathbf{z}$ to illustrate how the problem is solved.

Theorem 1: Given positive $\alpha, \beta \in \mathcal{R}$, $\mathbf{q}, \mathbf{z} \in \mathcal{C}^m$ and $\det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})) > 0$, we have

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det\mathbf{T}\neq 0}} \xi(\mathbf{T},\alpha,\beta,\mathbf{q},\mathbf{z}) = (|\mathbf{z}^H\mathbf{q}| + \alpha\beta)^2 - \alpha^2\beta^2, \qquad (31)$$

and $\xi(\mathbf{T},\alpha,\beta,\mathbf{q},\mathbf{z})$ achieves the minimum if and only if

$$\mathbf{T} = \mathbf{Q}\begin{bmatrix} \mathbf{\Phi}^{1/2} & \mathbf{0} \\ \mathbf{\Lambda}(\mathbf{\Phi}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}\mathbf{V} \qquad (32)$$

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

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

with nonzero $\gamma_{11}, \gamma_{22} \in \mathcal{R}$; $\mathbf{\Omega} \in \mathcal{R}^{(m-2)\times(m-2)}$ is an arbitrary nonsingular matrix; $\mathbf{V} \in \mathcal{R}^{m\times m}$ is an arbitrary orthogonal matrix;

$$\mathbf{\Phi} = \frac{\beta}{\alpha}\mathbf{\Gamma}^{-T}(\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\mathbf{\Gamma}^{-1} \qquad (34)$$

$$\mathbf{\Lambda} = \frac{\beta}{\alpha}\begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_m^T \end{bmatrix}\mathbf{Q}^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\mathbf{\Gamma}^{-1} \qquad (35)$$

with

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}, \qquad (36)$$

θ is the solution of

$$\tan\theta = \frac{a_{21} - a_{12}}{a_{11} + a_{22}}, \quad a_{11}\cos\theta - a_{12}\sin\theta > 0 \qquad (37)$$

with

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = (\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q}). \qquad (38)$$

B. Global optimal controller realizations

Define $i_1 \triangleq \arg\max_{i\in\{1,\cdots,m+n\}} \eta_i$. Without the loss of generality, it is assumed that $i_1 = m+n-1$, λ_{i_1} is a complex-valued eigenvalue, $\lambda_{i_1+1} = \lambda_{i_1}^*$ and $\det((\Upsilon(\mathbf{y}_{i_1,2}))^T\Upsilon(\mathbf{x}_{i_1,2})) > 0$. From Theorem 1, all the transformation matrices achieving η_{i_1} form the set

$$\mathcal{T} \triangleq \left\{\mathbf{T}\,\middle|\,\mathbf{T} = \mathbf{Q}\begin{bmatrix} \mathbf{\Phi}^{1/2} & \mathbf{0} \\ \mathbf{\Lambda}(\mathbf{\Phi}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}\mathbf{V}\right\} \qquad (39)$$

under $\alpha = \alpha_{i_1}, \beta = \beta_{i_1}, \mathbf{q} = \mathbf{x}_{i_1,2}$ and $\mathbf{z} = \mathbf{y}_{i_1,2}$. Note that η_{i_1} is a lower bound for the optimal value υ, i.e. $\upsilon \geq \eta_{i_1}$. When the equality holds, i.e. $\upsilon = \eta_{i_1}$, all the global optimal solutions to the optimization problem (10) lie in $\mathcal{T}$. This allows us to search in $\mathcal{T}$ for a global solution. Define

$$f_1(\mathbf{w}) \triangleq \max_{i\in\{1,\cdots,m+n-2\}} g(\mathbf{w}, i) \qquad (40)$$

and

$$\upsilon_1 \triangleq \min_{\mathbf{T}\in\mathcal{T}} f_1(\mathbf{w}(\mathbf{T})). \qquad (41)$$

It is straightforward to verify the following sufficient and necessary condition for $\upsilon = \eta_{i_1}$.

Proposition 1: $\upsilon = \eta_{i_1}$ if and only if $\upsilon_1 \leq \eta_{i_1}$.

Thus, any $\mathbf{T} \in \mathcal{T}$ which satisfies $f_1(\mathbf{w}(\mathbf{T})) \leq \eta_{i_1}$ is a global optimal solution of the optimization problem (10). In the previous work [6], an algorithm was developed to construct a global optimal $\mathbf{T}_{\text{opt}}$. Here, we present an alternative algorithm to compute a $\mathbf{T}_{\text{opt}}$.

Initialization: Arbitrarily select the positive scalar weightings σ_i, $i \in \{1, \cdots, m+n-2\}$.
Step 1: Solve the weighted eigenvalue sensitivity minimization problem

$$\min_{\mathbf{T}\in\mathcal{T}} \sum_{i=1}^{m+n-2} \sigma_i \left(g(\mathbf{w}(\mathbf{T}), i)\right)^2 \tag{42}$$

to obtain a solution $\mathbf{T}_+$.
Step 2: If $\max\limits_{i\in\{1,\cdots,m+n-2\}} g(\mathbf{w}(\mathbf{T}_+), i) \leq \eta_{i_1}$, set the global optimal transformation $\mathbf{T}_{\text{opt}} = \mathbf{T}_+$, and terminate the algorithm. Otherwise, find the index

$$i_+ = \arg \max_{i\in\{1,\cdots,m+n-2\}} g(\mathbf{w}(\mathbf{T}_+), i). \tag{43}$$

Properly increase σ_{i_+} to a new value, and go to *Step 1*.
Comment 1: The minimization problem (42) can be solved using the gradient flow technique [7]. The work [4] discussed the detailed method for a similar problem to (42).

IV. Optimum with the smallest dynamic range

We consider how to modify the optimal controller realization obtained in Section III to achieve the smallest dynamic range under the constraint that it remains to be a global minimum solution of the optimization problem (10). From the discussion in Section II, $\|\mathbf{w}\|_M$ indicates the dynamic range of $\mathbf{w}$. Therefore, it is appropriate to use it as the dynamic range measure of a realization, that is, $d(\mathbf{w}) \triangleq \|\mathbf{w}\|_M$. From the definition of $f(\mathbf{w})$ and (29), it is straightforward to prove the following proposition.

Proposition 2: For two realizations $\mathbf{w}_1$ and $\mathbf{w}_2$ (or equivalently $(\mathbf{F}_1, \mathbf{G}_1, \mathbf{J}_1, \mathbf{M}_1, \mathbf{H}_1)$ and $(\mathbf{F}_2, \mathbf{G}_2, \mathbf{J}_2, \mathbf{M}_2, \mathbf{H}_2)$), if there exists an orthogonal transformation $\mathbf{\Psi} \in \mathcal{R}^{m\times m}$ such that

$$\begin{cases} \mathbf{F}_2 = \mathbf{\Psi}^{-1}\mathbf{F}_1\mathbf{\Psi},\ \mathbf{G}_2 = \mathbf{\Psi}^{-1}\mathbf{G}_1, \\ \mathbf{J}_2 = \mathbf{J}_1\mathbf{\Psi},\ \mathbf{M}_2 = \mathbf{M}_1,\ \mathbf{H}_2 = \mathbf{\Psi}^{-1}\mathbf{H}_1. \end{cases} \tag{44}$$

then $f(\mathbf{w}_1) = f(\mathbf{w}_2)$.

Given $\mathbf{w}_{\text{opt}}$ obtained in Section III, define

$$\mathcal{S}_{\text{opt}} \triangleq \left\{ (\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H}) \;\middle|\; \begin{array}{l} \mathbf{F} = \mathbf{\Psi}^{-1}\mathbf{F}_{\text{opt}}\mathbf{\Psi} \\ \mathbf{G} = \mathbf{\Psi}^{-1}\mathbf{G}_{\text{opt}} \\ \mathbf{J} = \mathbf{J}_{\text{opt}}\mathbf{\Psi} \\ \mathbf{M} = \mathbf{M}_{\text{opt}} \\ \mathbf{H} = \mathbf{\Psi}^{-1}\mathbf{H}_{\text{opt}} \\ \mathbf{\Psi} \in \mathcal{R}^{m\times m} \\ \mathbf{\Psi}^T\mathbf{\Psi} = \mathbf{I} \end{array} \right\}. \tag{45}$$

Denote the generic realization in $\mathcal{S}_{\text{opt}}$ as $\mathbf{w}_{\text{opt}}(\mathbf{\Psi})$. It can be seen from Proposition 2 that, for any orthogonal $\mathbf{\Psi} \in \mathcal{R}^{m\times m}$, the realization $\mathbf{w}_{\text{opt}}(\mathbf{\Psi})$ remains to be a global minimum solution of the optimization problem (10). Thus, we can search in $\mathcal{S}_{\text{opt}}$ for an optimal realization with the smallest dynamic range

$$\mu \triangleq \min_{\substack{\mathbf{\Psi}\in\mathcal{R}^{m\times m} \\ \mathbf{\Psi}^T\mathbf{\Psi}=\mathbf{I}}} d(\mathbf{w}_{\text{opt}}(\mathbf{\Psi})). \tag{46}$$

In order to remove the constraint $\mathbf{\Psi}^T\mathbf{\Psi} = \mathbf{I}$ in the optimization problem (46), we derive a method for representing an orthogonal $\mathbf{\Psi}$ parameterized by its independent parameters. Firstly, when $m = 2$, it is plain to see that any orthogonal $\mathbf{\Psi}$ can be written as

$$\mathbf{\Psi} = \begin{bmatrix} \cos\theta_1 & -\kappa\sin\theta_1 \\ \sin\theta_1 & \kappa\cos\theta_1 \end{bmatrix}, \theta_1 \in [-\pi, \pi), \kappa \in \{-1, 1\}. \tag{47}$$

Next, for $m = 3$, constructing an orthogonal $\mathbf{\Psi}$ with its independent parameters can follow the following steps.
Step 1: Construct the first column $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$ of $\mathbf{\Psi}$. Since $\psi_{11}^2+\psi_{21}^2+\psi_{31}^2 = 1$, let $\psi_{11} = \cos\theta_1$ and $\psi_{21}^2+\psi_{31}^2 = \sin^2\theta_1$, where $\theta_1 \in [-\pi, \pi)$. Further let $\psi_{21} = \cos\theta_2\sin\theta_1,\ \psi_{31} = \sin\theta_2\sin\theta_1$, where $\theta_2 \in [-\pi, \pi)$. Thus the first column of $\mathbf{\Psi}$ is defined by

$$\begin{bmatrix} \psi_{11} \\ \psi_{21} \\ \psi_{31} \end{bmatrix} = \begin{bmatrix} \cos\theta_1 \\ \cos\theta_2\sin\theta_1 \\ \sin\theta_2\sin\theta_1 \end{bmatrix},\ \theta_1, \theta_2 \in [-\pi, \pi), \tag{48}$$

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 $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$.
Step 2.1: Construct the first column $[\psi_{12}\ \psi_{22}\ \psi_{32}]^T$ of the orthonormal basis.
(a) $\theta_1 \neq 0$ or $-\pi$. Let $\mathcal{P}_1$ be the span of $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$ and $[1\ 0\ 0]^T$. Construct $[\psi_{12}\ \psi_{22}\ \psi_{32}]^T \in \mathcal{P}_1$ as a unit vector perpendicular to $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$, which means that

$$\begin{cases} \begin{bmatrix} \psi_{12} \\ \psi_{22} \\ \psi_{32} \end{bmatrix} = k_1 \begin{bmatrix} \cos\theta_1 \\ \cos\theta_2\sin\theta_1 \\ \sin\theta_2\sin\theta_1 \end{bmatrix} + k_2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ \psi_{12}^2 + \psi_{22}^2 + \psi_{32}^2 = 1 \\ \psi_{12}\cos\theta_1 + (\psi_{22}\cos\theta_2 + \psi_{32}\sin\theta_2)\sin\theta_1 = 0 \end{cases} \tag{49}$$

Solving the above equations, we obtain

$$k_1 = -\frac{\cos\theta_1}{\sin\theta_1},\quad k_2 = \frac{1}{\sin\theta_1}, \tag{50}$$

or

$$k_1 = \frac{\cos\theta_1}{\sin\theta_1},\quad k_2 = -\frac{1}{\sin\theta_1}. \tag{51}$$

As only one orthonormal basis is needed, without the loss of generality, we adopt (51) and set

$$\begin{bmatrix} \psi_{12} \\ \psi_{22} \\ \psi_{32} \end{bmatrix} = \begin{bmatrix} -\sin\theta_1 \\ \cos\theta_2\cos\theta_1 \\ \sin\theta_2\cos\theta_1 \end{bmatrix}. \tag{52}$$

(b) $\theta_1 = 0$ or $-\pi$. As $[-\sin\theta_1\ \ \cos\theta_2\cos\theta_1\ \ \sin\theta_2\cos\theta_1]^T$ remains to be perpendicular to $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$, $[\psi_{12}\ \psi_{22}\ \psi_{32}]^T$ can always be constructed using (52).
Step 2.2: Construct the other column $[\psi_{13}\ \psi_{23}\ \psi_{33}]^T$ of the orthonormal basis. Denote $\mathcal{P}_2$ the span of $[\psi_{11}\ \psi_{21}\ \psi_{31}]^T$ and $[\psi_{12}\ \psi_{22}\ \psi_{32}]^T$. $[\psi_{13}\ \psi_{23}\ \psi_{33}]^T$ is perpendicular to $\mathcal{P}_2$ and hence perpendicular to $[1\ 0\ 0]^T \in \mathcal{P}_2$. This means that $\psi_{13} = 0$ and $[\psi_{23}\ \psi_{33}]^T$ is perpendicular to both $[\psi_{21}\ \psi_{31}]^T$ and $[\psi_{22}\ \psi_{32}]^T$. Noting $[\psi_{21}\ \psi_{31}]^T = [\cos\theta_2\ \ \sin\theta_2]^T\sin\theta_1$ and $[\psi_{22}\ \psi_{32}]^T = [\cos\theta_2\ \ \sin\theta_2]^T\cos\theta_1$, we can see that $[\psi_{23}\ \psi_{33}]^T$ is the orthonormal basis of the subspace perpendicular to $[\cos\theta_2\ \ \sin\theta_2]^T$. From the formula (47) for the case

of $m = 2$, we know that it can be chosen as $[\psi_{23} \ \psi_{33}]^T = [-\sin\theta_2 \ \cos\theta_2]^T$.

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

$$\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 \\ \sin\theta_2\sin\theta_1 & \sin\theta_2\cos\theta_1 & \cos\theta_2 \end{bmatrix} \tag{53}$$

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 $\mathbf{\Psi} \in \mathcal{R}^{3\times 3}$, it is only needed to rotate the orthonormal basis in $\mathcal{P}_0$. This means that, from (47) and (53), we have

$$\begin{aligned} \mathbf{\Psi} \ &= \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \cos\theta_2\sin\theta_1 & \cos\theta_2\cos\theta_1 & -\sin\theta_2 \\ \sin\theta_2\sin\theta_1 & \sin\theta_2\cos\theta_1 & \cos\theta_2 \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\kappa\sin\theta_3 \\ 0 & \sin\theta_3 & \kappa\cos\theta_3 \end{bmatrix}, \\ &\theta_1, \theta_2, \theta_3 \in [-\pi, \pi), \kappa \in \{-1, 1\}. \end{aligned} \tag{54}$$

In the similar way, the formula representing an arbitrary orthogonal $\mathbf{\Psi} \in \mathcal{R}^{m\times m}$ with its independent parameters can be derived for $m > 3$. Define $r = \frac{m(m-1)}{2}$. In general, $\mathbf{\Psi} \in \mathcal{R}^{m\times m}$ is parameterized by $\theta_1, \cdots, \theta_r \in [-\pi, \pi)$ and $\kappa \in \{-1, +1\}$. Following from a simple observation:

$$\begin{aligned} & d\left(\mathbf{w}_{\text{opt}}\left(\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\right)\right) \\ &= d\left(\mathbf{w}_{\text{opt}}\left(\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\right)\right), \end{aligned} \tag{55}$$

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

$$d_1(\theta_1, \cdots, \theta_r) \stackrel{\triangle}{=} d(\mathbf{w}_{\text{opt}}(\mathbf{\Psi})). \tag{56}$$

Then the optimization problem (46) is equivalent to the unconstrained optimization problem

$$\mu = \min_{\theta_1, \cdots, \theta_r \in [-\pi, \pi)} d_1(\theta_1, \cdots, \theta_r). \tag{57}$$

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 $\theta_{1\text{opt}}, \cdots, \theta_{r\text{opt}}$, we can obtain the optimal orthogonal transformation $\mathbf{\Psi}_{\text{opt}}$ and hence the optimal realization $\mathbf{w}_{\text{opt1}} = \mathbf{w}_{\text{opt}}(\mathbf{\Psi}_{\text{opt}})$ of the smallest dynamic range.

V. A DESIGN EXAMPLE

The example considered in [8] was used to illustrate the proposed design procedure for obtaining optimal FWL fixed-point controller realizations. Given the discrete-time plant model ($n = 4, p = 1, q = 1$) and the initial realization of the digital controller ($m = 4$), the algorithm presented in Section III-B was first used to obtain an optimal transformation matrix $\mathbf{T}_{\text{opt}}$. The realization $\mathbf{w}_{\text{opt}} = \mathbf{w}(\mathbf{T}_{\text{opt}})$ was a global optimal realization that minimized the FWL closed-loop stability measure. In order to obtain an optimal realization with the smallest dynamic range, the optimization problem (57) was formed given the dimension $r = 6$. The MATLAB routine *fminsearch.m* was used to solve this optimization problem numerically, and the global optimal realization with the smallest dynamic range, $\mathbf{w}_{\text{opt1}} = \mathbf{w}_{\text{opt}}(\mathbf{\Psi}_{\text{opt}})$, was then calculated.

TABLE I

COMPARISON OF VARIOUS CONTROLLER REALIZATIONS

Realization	$f(\mathbf{w})$	$d(\mathbf{w})$	$b_f^{\min}$	$b_g^{\min}$	$b^{\min}$
$\mathbf{w}_0$	$3.9697e+6$	$1.0959e+6$	20	21	42
$\mathbf{w}_{\text{opt}}$	$2.4246e+3$	$1.9673e+2$	8	8	17
$\mathbf{w}_{\text{opt1}}$	$2.4246e+3$	$1.1799e+2$	8	7	16

Table I lists the values of the FWL stability measure $f(\mathbf{w})$ and the dynamic range measure $d(\mathbf{w})$ together with the related minimum bit lengths $b^{\min}$, $b_g^{\min}$ and $b_f^{\min}$ for the realizations $\mathbf{w}_0$, $\mathbf{w}_{\text{opt}}$ and $\mathbf{w}_{\text{opt1}}$, respectively. It can be seen that the fixed-point implementation of $\mathbf{w}_0$ needs at least 42 bits, while the implementation of $\mathbf{w}_{\text{opt}}$ needs at least 17 bits. The latter achieved a reduction of 25 bits in the required bit length. As expected, $f(\mathbf{w}_{\text{opt1}}) = f(\mathbf{w}_{\text{opt}})$ but $d(\mathbf{w}_{\text{opt1}})$ is smaller than $d(\mathbf{w}_{\text{opt}})$, giving rise to further one bit reduction in $b_g^{\min}$ for $\mathbf{w}_{\text{opt1}}$.

ACKNOWLEDGMENTS

J. Wu and J. Chu wish to thank the support of NSFC #60374002, #60421002, 973 program of China #2002CB312200 and NCET-04-0547. S. Chen wishes to thank the support of the United Kingdom Royal Academy of Engineering.

REFERENCES

[1] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automatic Control*, Vol.43, No.5, pp.689–693, 1998.

[2] 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, No.11, pp.2149–2153, 1999.

[3] 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, No.2, pp.166–176, 2000.

[4] J.F. Whidborne, R.S.H. Istepanian and J. Wu, "Reduction of controller fragility by pole sensitivity minimization," *IEEE Trans. Automatic Control*, Vol.46, No.2, pp.320–325, 2001.

[5] 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, No.7, pp.1162–1166, 2001.

[6] 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, 2002, pp.2941–2946.

[7] J.E. Perkins, U. Helmke and J.B. Moore, "Balanced realizations via gradient flow techniques," *Systems and Control Letters*, Vol.14, No.6, pp.369–380, 1990.

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