Comparative Study on Finite-Precision Controller Realizations in Different Representation Schemes *

Jun Wu†, Sheng Chen‡ and Jian Chu†

† National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou, 310027, P. R. China

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

Abstract: A computationally tractable finite word length (FWL) closed-loop stability measure is derived which is applicable to fixed-point, floating-point and block-floating-point representation schemes. Both the dynamic range and precision of an arithmetic scheme are considered in this new unified measure. For each arithmetic scheme, the optimal controller realization problem is defined and a numerical optimization approach is adopted to solve it. Two examples are used to illustrate the design procedure and to compare the optimal controller realizations in different representation schemes.

Keywords — digital controller, finite word length, arithmetic scheme, closed-loop stability, optimization.

1 Introduction

In recent years, there has been a growing interest in digital controller implementation which reduces the FWL effects on closed-loop stability. It is well known that a control law can be accomplished with different realizations and that the parameters of a controller realization are represented by a digital processor of finite bit length in a particular format, namely fixed-point, floating-point or block-float-point format. Previous works [1]–[4] have derived some FWL closed-loop stability measures for these three formats, respectively, and defined the corresponding optimal controller realization problems based on these measures. However, all these previous measures are only linked to the precision bit lengths of the respective representation schemes used and do not consider the dynamic range bit lengths. Arguably, a better approach is to consider some measure which has a direct link to the total bit length required. The main contribution of this paper is to derive a unified FWL closed-loop stability measure that can accommodate both the dynamic range and precision requirements and is applicable to all the three schemes.

*J. Wu and S. Chen wish to thank the support of the UK Royal Society under a KC Wong fellowship (RL/ART/CN/XFI/KCW/11949). J. Wu and J. Chu wish to thank the support of National Natural Science Foundation of China (Grant Ref.60174026), Zhejiang Provincial Natural Science Foundation of China (Grant Ref.699085) and Doctor Degree Programs Foundation of China (Grant Ref.1999033571).

2 Number Representation Schemes

When $x \in \mathcal{R}$ is represented in the fixed-point scheme of bit length $\beta = 1 + \beta_g + \beta_f$, the bits are assigned as follows: one bit for the sign, β_g bits for the integer part and β_f bits for the fraction part. Assuming that no overflow occurs, which means that $|x| \leq 2^{\beta_g}$, x is perturbed to

$$\mathcal{Q}_1(x) = x + \delta_1, \quad |\delta_1| < 2^{-(\beta_f+1)}. \tag{1}$$

Any $x \in \mathcal{R}$ can be expressed uniquely as $x = (-1)^s \times w \times 2^e$, where $s \in \{0, 1\}$ is the sign of x, $w \in [0.5,\ 1)$ is the mantissa of x, $e = \lfloor \log_2 |x| \rfloor + 1 \in \mathcal{Z}$ is the exponent of x, $\mathcal{Z}$ denotes the set of integers and the *floor* function $\lfloor x \rfloor$ is the closest integer less than or equal to x. When x is stored in the floating-point format of bit length $\beta = 1 + \beta_w + \beta_e$, the bits consists of three parts: one bit for s, β_w bits for w and β_e bits for e. Let $\underline{e}$ and $\overline{e}$ be the lower and upper limits of the exponent, respectively. Clearly, $\overline{e} - \underline{e} = 2^{\beta_e} - 1$. Denote the set of integers $\underline{e} \leq e \leq \overline{e}$ as $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$. Assuming that no underflow or overflow occurs, which means that the exponent of x is within $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$, x is perturbed to

$$\mathcal{Q}_2(x) = x + x\delta_2, \ \ |\delta_2| < 2^{-(\beta_w+1)}. \tag{2}$$

In the block-floating-point format, a set of real numbers $\mathcal{S}$ is first divided into some blocks. For an illustrative purpose, consider the case of dividing $\mathcal{S}$ into the two non-empty and non-overlapped subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $\eta_1 \in \mathcal{S}_1$ be the element in $\mathcal{S}_1$ that has the largest absolute value, and $\eta_2 \in \mathcal{S}_2$ be the element in $\mathcal{S}_2$ that has the largest absolute value. Then, any $x \in \mathcal{S}$ can be expressed uniquely as $x = (-1)^s \times u \times 2^h$, where $u \in [0,\ 1)$ is the block mantissa of x, and the block exponent of x is

$$h \triangleq \begin{cases} \lfloor \log_2 |\eta_1| \rfloor + 1, & \text{for} \quad x \in \mathcal{S}_1, \\ \lfloor \log_2 |\eta_2| \rfloor + 1, & \text{for} \quad x \in \mathcal{S}_2. \end{cases} \tag{3}$$

When all the elements in $\mathcal{S}$ are presented in the block-floating-point format of bit length $\beta = 1 + \beta_u + \beta_h$, the bits are assigned as follows: 1 bit for the sign, β_u bits for u which is represented in fixed-point with the two's complement system, and β_h bits for h. Let $\underline{h}$ and $\overline{h}$ be the lower and upper limits of the block exponent, respectively. Obviously, $\overline{h} - \underline{h} = 2^{\beta_h} - 1$. Denote

$$r(x) \triangleq \begin{cases} 2\eta_1, & \text{for} \quad x \in \mathcal{S}_1, \\ 2\eta_2, & \text{for} \quad x \in \mathcal{S}_2. \end{cases} \tag{4}$$

Assuming no underflow or overflow, i.e. the block exponent of x is within $\mathcal{Z}_{[\underline{h},\, \overline{h}]}$, x is perturbed to

$$\mathcal{Q}_3(x) = x + r(x)\delta_3, \;\; |\delta_3| < 2^{-(\beta_u+1)}. \tag{5}$$

For the notational conciseness, we introduce the "generalized" dynamic range bit length β_r and precision bit length β_p for the three representation schemes. It is understood that $\beta_r = \beta_g$, β_e or β_h and $\beta_p = \beta_f$, β_w or β_u, depending on which format is actually used.

3 Problem Statement

The discrete-time linear time-invariant plant P is described by

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases} \tag{6}$$

with $\mathbf{A} \in \mathcal{R}^{n\times n}$, $\mathbf{B} \in \mathcal{R}^{n\times p}$ and $\mathbf{C} \in \mathcal{R}^{q\times n}$; and the generic digital controller C is described by

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{H}\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases} \tag{7}$$

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 $\mathbf{e}(k) = \mathbf{q}(k) + \mathbf{u}(k)$ with the command input $\mathbf{q}(k)$. Then P and C form a closed-loop control system. Assume that a realization $(\mathbf{F}_0, \mathbf{G}_0, \mathbf{J}_0, \mathbf{M}_0, \mathbf{H}_0)$ of C has been designed. It is well-known that the realizations of C are not unique. All the realizations of C form the realization set

$$\begin{aligned} \mathcal{S}_C \triangleq \{(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0, \\ \mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0, \mathbf{H} = \mathbf{T}^{-1}\mathbf{H}_0\} \end{aligned} \tag{8}$$

where $\mathbf{T} \in \mathcal{R}^{m\times m}$ is any nonsingular matrix. Let $\mathbf{w}_F = \mathrm{Vec}(\mathbf{F})$, where $\mathrm{Vec}(\cdot)$ denotes the column stacking operator, and $\mathbf{w}_{F_0}$, $\mathbf{w}_G$, $\mathbf{w}_{G_0}$, $\mathbf{w}_J$, $\mathbf{w}_{J_0}$, $\mathbf{w}_M$, $\mathbf{w}_{M_0}$, $\mathbf{w}_H$ and $\mathbf{w}_{H_0}$ be similarly defined. Denote

$$\begin{aligned} \mathbf{w} &= [w_1 \cdots w_N]^T \triangleq \left[\mathbf{w}_F^T \; \mathbf{w}_G^T \; \mathbf{w}_J^T \; \mathbf{w}_M^T \; \mathbf{w}_H^T\right]^T \\ \mathbf{w}_0 &\triangleq \left[\mathbf{w}_{F_0}^T \; \mathbf{w}_{G_0}^T \; \mathbf{w}_{J_0}^T \; \mathbf{w}_{M_0}^T \; \mathbf{w}_{H_0}^T\right]^T \end{aligned} \tag{9}$$

where $N = (m+p)(m+q) + mp$ and ^{T} is the transpose operator. We also refer to $\mathbf{w}$ as a realization of C.

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

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{w}) &= \begin{bmatrix} \mathbf{A} + \mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC} + \mathbf{HMC} & \mathbf{F} + \mathbf{HJ} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_0) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix}. \end{aligned} \tag{10}$$

All the different realizations $\mathbf{w}$ have the same set of closed-loop poles if they are implemented with infinite precision. Since the closed-loop system is designed to be stable, the eigenvalues

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| = |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))| < 1, \;\; \forall i \in \{1, \cdots, m+n\}. \tag{11}$$

Define

$$\begin{aligned} \|\mathbf{w}\|_{\max} &\triangleq \max_{j\in\{1,\cdots,N\}} |w_j|, \\ \pi(\mathbf{w}) &\triangleq \min_{j\in\{1,\cdots,N\}} \{|w_j| : w_j \neq 0\}, \end{aligned} \tag{12}$$

and the index α of representation formats adopted

$$\alpha = \begin{cases} 1, & \text{fixed-point format,} \\ 2, & \text{floating-point format,} \\ 3, & \text{block-floating-point format.} \end{cases} \tag{13}$$

The controller realization $\mathbf{w}$ is implemented in format α of β_r dynamic range bits, β_p precision bits and one sign bit. In the remainder of this paper, it is assumed that if $\mathbf{w}$ is stored in the block-floating-point format, it is divided into "natural" blocks of $\mathbf{w}_F$, $\mathbf{w}_G$, $\mathbf{w}_J$, $\mathbf{w}_M$ and $\mathbf{w}_H$. Let $\eta_F \in \mathbf{w}_F$ be the element in $\mathbf{F}$ which has the largest absolute value. The elements η_G, η_J, η_M and η_H are similarly defined. Denote

$$\mathbf{z}(\mathbf{w}) \triangleq \begin{bmatrix} \eta_F & \eta_G & \eta_J & \eta_M & \eta_H \end{bmatrix}^T. \tag{14}$$

4 Optimization of an FWL Closed-Loop Stability Measure

Firstly, the dynamic range bit length of β_r bits must be large enough to accommodate $\mathbf{w}$. We define a dynamic range measure for realization $\mathbf{w}$ in format α as

$$\gamma(\mathbf{w}, \alpha) \triangleq \begin{cases} \|\mathbf{w}\|_{\max}, & \alpha = 1, \\ \log_2 \frac{4\|\mathbf{w}\|_{\max}}{\pi(\mathbf{w})}, & \alpha = 2, \\ \log_2 \frac{4\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}, & \alpha = 3. \end{cases} \tag{15}$$

Proposition 1 The realization $\mathbf{w}$ can be represented in the fixed-point format of β_g integer bits without overflow, if $2^{\beta_g} \geq \|\mathbf{w}\|_{\max}$; $\mathbf{w}$ can be represented in the floating-point format of β_e exponent bits without underflow or overflow, if $2^{\beta_e} \geq \log_2\left(\frac{\|\mathbf{w}\|_{\max}}{\pi(\mathbf{w})}\right) + 2$; $\mathbf{w}$ can be represented in the block-floating-point format of β_h block exponent bits without underflow or overflow, if $2^{\beta_h} \geq \log_2\left(\frac{\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}\right) + 2$.

Let β_r^{min} be the smallest dynamic-range bit length that, when used to implement $\mathbf{w}$ with format α, does not cause overflow or underflow. $\beta_r^{min}(\mathbf{w},\alpha)$ can easily be computed by: $\lceil \log_2 \|\mathbf{w}\|_{\max} \rceil$ when $\alpha = 1$, $\lceil \log_2(\lfloor \log_2 \|\mathbf{w}\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{w}) \rfloor + 1) \rceil$ when $\alpha = 2$ and $\lceil \log_2(\lfloor \log_2 \|\mathbf{z}(\mathbf{w})\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{z}(\mathbf{w})) \rfloor + 1) \rceil$ when $\alpha = 3$, where the *ceiling* function $\lceil x \rceil$ denotes the closest integer greater than or equal to $x \in \mathcal{R}$. Note that the measure $\gamma(\mathbf{w},\alpha)$ defined in (15) provides an estimate of β_r^{min} as

$$\hat{\beta}_r^{min}(\mathbf{w},\alpha) \triangleq \lceil \log_2 \gamma(\mathbf{w},\alpha) \rceil. \tag{16}$$

It can easily be seen that $\hat{\beta}_r^{min} \geq \beta_r^{min}$ and, when the fixed-point format is adopted, $\hat{\beta}_r^{min} = \beta_r^{min}$.

For a vector $\mathbf{x}$, let $\mathbf{d}(\mathbf{x})$ be the vector of the same dimension whose elements are all 1s and denote

$$\tau(\mathbf{x}) \triangleq \begin{cases} 0, & \mathbf{x} \text{ is a zero vector,} \\ 1, & \mathbf{x} \text{ is a nonzero vector.} \end{cases} \tag{17}$$

For two vectors $\mathbf{x} = [x_j]$ and $\mathbf{y} = [y_j]$ of the same dimension, define the Hadamard product of $\mathbf{x}$ and $\mathbf{y}$ as $\mathbf{x} \circ \mathbf{y} \triangleq [x_j y_j]$. When the dynamic range of representation format α is sufficient, according to the results of Section 2, $\mathbf{w}$ is perturbed to $\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}$ due to the effect of finite β_p where

$$\mathbf{r}(\mathbf{w},\alpha) \triangleq \begin{cases} \begin{bmatrix} \tau(\mathbf{w}_F)\mathbf{d}(\mathbf{w}_F) \\ \tau(\mathbf{w}_G)\mathbf{d}(\mathbf{w}_G) \\ \tau(\mathbf{w}_J)\mathbf{d}(\mathbf{w}_J) \\ \tau(\mathbf{w}_M)\mathbf{d}(\mathbf{w}_M) \\ \tau(\mathbf{w}_H)\mathbf{d}(\mathbf{w}_H) \end{bmatrix}, & \alpha = 1, \\ \mathbf{w}, & \alpha = 2, \\ \begin{bmatrix} 2\eta_F\mathbf{d}(\mathbf{w}_F) \\ 2\eta_G\mathbf{d}(\mathbf{w}_G) \\ 2\eta_J\mathbf{d}(\mathbf{w}_J) \\ 2\eta_M\mathbf{d}(\mathbf{w}_M) \\ 2\eta_H\mathbf{d}(\mathbf{w}_H) \end{bmatrix}, & \alpha = 3. \end{cases} \tag{18}$$

Each element δ_j of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_p+1)}$, that is, $\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_p+1)}$. With the perturbation $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision implemented $\mathbf{w}$ in format α. It is therefore critical to know when the FWL error will cause closed-loop instability. From a first-order approximation, $\forall i \in \{1, \cdots, m+n\}$

$$\begin{aligned} &|\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| \\ &\quad \approx \sum_{j=1}^{N} \left. \frac{\partial |\lambda_i|}{\partial \delta_j} \right|_{\mathbf{\Delta}=\mathbf{0}} \delta_j \,. \end{aligned} \tag{19}$$

For the derivative $\frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} = \left[\frac{\partial |\lambda_i|}{\partial \delta_j} \right]$, define

$$\left\| \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right\|_1 \triangleq \sum_{j=1}^{N} \left| \frac{\partial |\lambda_i|}{\partial \delta_j} \right| . \tag{20}$$

Then

$$\begin{aligned} &|\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| \\ &\quad \leq \|\mathbf{\Delta}\|_{\max} \left\| \left. \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right|_{\mathbf{\Delta}=\mathbf{0}} \right\|_1 . \end{aligned} \tag{21}$$

This leads to the following precision measure for realization $\mathbf{w}$ in format α

$$\mu(\mathbf{w},\alpha) \triangleq \min_{i \in \{1,\cdots,m+n\}} \frac{1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\left\| \left. \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right|_{\mathbf{\Delta}=\mathbf{0}} \right\|_1} . \tag{22}$$

Obviously, if $\|\mathbf{\Delta}\|_{\max} < \mu(\mathbf{w},\alpha)$, then $|\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}))| < 1$ which means that the closed-loop remains stable under the FWL error $\mathbf{\Delta}$. In other words, for a given $\mathbf{w}$ implemented in format α with a sufficient dynamic range, the closed-loop can tolerate those FWL perturbations $\mathbf{\Delta}$ whose norms $\|\mathbf{\Delta}\|_{\max}$ are less than $\mu(\mathbf{w},\alpha)$. It is easy to see that

$$\left. \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right|_{\mathbf{\Delta}=\mathbf{0}} = \mathbf{r}(\mathbf{w},\alpha) \circ \frac{\partial |\lambda_i|}{\partial \mathbf{w}}, \tag{23}$$

and from the results of [2], it can be shown that the value of $\mu(\mathbf{w},\alpha)$ can be computed explicitly.

Under the condition that the dynamic range is sufficient, that is, $\beta_r \geq \beta_r^{min}$, the perturbation $\|\mathbf{\Delta}\|_{\max}$ and therefore the precision bit length β_p determines whether the closed-loop remains stable. Let β_p^{min} be the smallest precision bit length that, when used to implement $\mathbf{w}$ with format α, guarantees the closed-loop stability. From the precision measure $\mu(\mathbf{w},\alpha)$, an estimate of β_p^{min} is given as

$$\hat{\beta}_p^{min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \mu(\mathbf{w},\alpha) \rfloor - 1 \,. \tag{24}$$

Define the minimum total bit length required in the implementation of $\mathbf{w}$ with format α as

$$\beta^{min} \triangleq \beta_r^{min} + \beta_p^{min} + 1 \,. \tag{25}$$

Clearly, $\mathbf{w}$ implemented with a bit length $\beta \geq \beta^{min}$ can guarantee a sufficient dynamic range and closed-loop stability. Combining the measures $\gamma(\mathbf{w},\alpha)$ and $\mu(\mathbf{w},\alpha)$ results in the following true FWL closed-loop stability measure for the given realization $\mathbf{w}$ with format α

$$\rho(\mathbf{w},\alpha) \triangleq \mu(\mathbf{w},\alpha)/\gamma(\mathbf{w},\alpha) \,. \tag{26}$$

An estimate of β^{min} is given by $\rho(\mathbf{w},\alpha)$ as

$$\hat{\beta}^{min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \rho(\mathbf{w},\alpha) \rfloor + 1 \,. \tag{27}$$

The measure $\rho(\mathbf{w},\alpha)$ provides the FWL characteristics of a realization $\mathbf{w}$ in a given format α. The optimal controller realization problem in format α is formally defined as

$$\upsilon(\alpha) \triangleq \max_{\mathbf{w} \in \mathcal{S}_C} \rho(\mathbf{w},\alpha) \,. \tag{28}$$

Define the following optimization criterion in format α:

$$\xi(\mathbf{T},\alpha) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))|}{\left\|\mathbf{r}(\mathbf{w},\alpha)\circ\frac{\partial|\lambda_i|}{\partial\mathbf{w}}\right\|_1\gamma(\mathbf{w},\alpha)}$$

$$= \rho(\mathbf{w},\alpha). \tag{29}$$

The optimal realization problem (28) can then be posed as the following optimization problem:

$$\upsilon(\alpha) = \max_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m}\\ \det(\mathbf{T})\neq 0}} \xi(\mathbf{T},\alpha). \tag{30}$$

Given $\mathbf{T}_{\rm opt}(\alpha)$, the optimal realization $\mathbf{w}_{\rm opt}(\alpha)$ can readily be computed. By setting $\alpha = 1$, 2 and 3, respectively, in the optimization problem (30), we can attain the optimal fixed-point realization $\mathbf{w}_{\rm opt}(1)$, the optimal floating-point realization $\mathbf{w}_{\rm opt}(2)$ and the optimal block-floating-point realization $\mathbf{w}_{\rm opt}(3)$.

5 Two Design Examples and Result Comparison

In Example 1, the closed-loop system contained a plant with $n = 5$ and a reduced-order observer-based controller with $m = 2$. Based on the proposed unified FWL closed-loop stability measure, the optimization problem (30) was formed. Using the MATLAB routine *fminsearch.m*, this optimization problem was solved for $\alpha = 1$, 2 and 3, respectively, to obtain the optimal realizations $\mathbf{w}_{\rm opt}(1)$, $\mathbf{w}_{\rm opt}(2)$ and $\mathbf{w}_{\rm opt}(3)$. In Example 2, the closed-loop system contained a plant with $n = 4$ and a output-feedback controller with $m = 4$. Using the same procedure for Example 1, the optimal realizations $\mathbf{w}_{\rm opt}(1)$, $\mathbf{w}_{\rm opt}(2)$ and $\mathbf{w}_{\rm opt}(3)$ were obtained. Table 1 lists the values of the measures ρ, μ and γ in the three different representation schemes together with the corresponding estimated minimum bit lengths for $\mathbf{w}_0$ and $\mathbf{w}_{\rm opt}(\alpha)$ of Example 1. Table 2 does the same thing for Example 2. As far as the robustness of FWL closed-loop stability is concerned, given an arbitrary realization, floating-point representation is not necessarily better than fixed-point or block-floating-point one. For example, floating-point is the best format to implement the initial realization $\mathbf{w}_0$ of Example 1 while fixed-point is the best format to implement $\mathbf{w}_0$ of Example 2. However, as expected, the optimal floating-point realization $\mathbf{w}_{\rm opt}(2)$ implemented in floating-point format is always the best in terms of robustness to FWL errors. Also the results in Table 1 show that fixed-point format is better than block-floating-point format to implement $\mathbf{w}_{\rm opt}(\alpha)$ of Example 1 for $1 \le \alpha \le 3$, while the results of Table 2 indicate that the opposite is true for Example 2. This simply confirms the fact that the performance of block-floating-point scheme critically depends on how to divide $\mathbf{w}$ into blocks. With a proper division, block-floating-point scheme should beat fixed-point scheme in terms of robustness to FWL errors. Table 3 compares the true minimum required bit lengths β_r^{min}, β_p^{min} and β^{min} of $\mathbf{w}_0$ implemented in the three different schemes with those of fixed-point implemented $\mathbf{w}_{\rm opt}(1)$, floating-point implemented $\mathbf{w}_{\rm opt}(2)$ and block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ of Example 1, respectively. Table 4 does the same thing for Example 2.

6 Conclusions

We have proposed a design procedure for optimal controller realizations in different representation schemes. The procedure provides designer with useful quantitative information regarding robustness to FWL errors and estimated minimum bit length for guaranteeing closed-loop stability. This allows designer to choose an optimal controller realization in an appropriate representation scheme to achieve best computational efficiency and closed-loop performance.

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] 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.

[3] J.F. Whidborne and D. Gu, "Optimal finite-precision controller and filter realizations using floating-point arithmetic," *Research Report EM2001/07*, Department of Mechanical Engineering, King's College London, U.K., September 2001.

[4] R.S.H. Istepanian, J.F. Whidborne and P. Bauer, "Stability analysis of block floating point digital controllers," in *Proc. UKACC Int. Conf. Control* (Cambridge, U.K.), Sept. 4-7, 2000, CD-ROM, 6 pages.

	$\mathbf{w}_0$	$\mathbf{w}_{\text{opt}}(1)$	$\mathbf{w}_{\text{opt}}(2)$	$\mathbf{w}_{\text{opt}}(3)$
$\rho(\mathbf{w},1)$	$\mathbf{2.5150e-9}$	$\mathbf{1.1386e-7}$	$2.7728e-8$	$1.0861e-7$
$\hat{\beta}^{min}(\mathbf{w},1)$	**30**	**25**	27	25
$\mu(\mathbf{w},1)$	$\mathbf{2.5569e-6}$	$\mathbf{5.0795e-7}$	$2.5937e-5$	$1.7450e-7$
$\hat{\beta}_p^{min}(\mathbf{w},1)$	**18**	**20**	15	22
$\gamma(\mathbf{w},1)$	$\mathbf{1.0167e+3}$	$\mathbf{4.4612e+0}$	$9.3543e+2$	$1.6066e+0$
$\hat{\beta}_r^{min}(\mathbf{w},1)$	**10**	**3**	10	1
$\rho(\mathbf{w},2)$	$\mathbf{1.3134e-7}$	$1.9204e-5$	$\mathbf{1.9593e-5}$	$3.3365e-7$
$\hat{\beta}^{min}(\mathbf{w},2)$	**24**	17	**17**	23
$\mu(\mathbf{w},2)$	$\mathbf{3.1118e-6}$	$4.3127e-4$	$\mathbf{4.3127e-4}$	$5.4490e-6$
$\hat{\beta}_p^{min}(\mathbf{w},2)$	**18**	11	**11**	17
$\gamma(\mathbf{w},2)$	$\mathbf{2.3692e+1}$	$2.2458e+1$	$\mathbf{2.2012e+1}$	$1.6332e+1$
$\hat{\beta}_r^{min}(\mathbf{w},2)$	**5**	5	**5**	5
$\rho(\mathbf{w},3)$	$\mathbf{9.2976e-10}$	$5.3779e-9$	$2.8185e-9$	$\mathbf{1.3362e-8}$
$\hat{\beta}^{min}(\mathbf{w},3)$	**32**	29	30	**28**
$\mu(\mathbf{w},3)$	$\mathbf{2.1343e-8}$	$5.7385e-8$	$5.7266e-8$	$\mathbf{5.4549e-8}$
$\hat{\beta}_p^{min}(\mathbf{w},3)$	**25**	24	24	**24**
$\gamma(\mathbf{w},3)$	$\mathbf{2.2955e+1}$	$1.0671e+1$	$2.0318e+1$	$\mathbf{4.0823e+0}$
$\hat{\beta}_r^{min}(\mathbf{w},3)$	**5**	4	5	**3**

Table 1: Measures and estimated minimum bit lengths of example 1.

	$\mathbf{w}_0$	$\mathbf{w}_{\text{opt}}(1)$	$\mathbf{w}_{\text{opt}}(2)$	$\mathbf{w}_{\text{opt}}(3)$
$\rho(\mathbf{w},1)$	$\mathbf{1.2312e-10}$	$\mathbf{1.2003e-6}$	$1.0580e-7$	$1.1321e-6$
$\hat{\beta}^{min}(\mathbf{w},1)$	**34**	**21**	25	21
$\mu(\mathbf{w},1)$	$\mathbf{3.3474e-8}$	$\mathbf{2.3082e-4}$	$9.6673e-5$	$2.2287e-4$
$\hat{\beta}_p^{min}(\mathbf{w},1)$	**24**	**12**	13	12
$\gamma(\mathbf{w},1)$	$\mathbf{2.7188e+2}$	$\mathbf{1.9231e+2}$	$9.1370e+2$	$1.9687e+2$
$\hat{\beta}_r^{min}(\mathbf{w},1)$	**9**	**8**	10	8
$\rho(\mathbf{w},2)$	$\mathbf{2.9062e-11}$	$7.6826e-6$	$\mathbf{9.5931e-6}$	$8.5778e-6$
$\hat{\beta}^{min}(\mathbf{w},2)$	**37**	18	**18**	18
$\mu(\mathbf{w},2)$	$\mathbf{2.2389e-10}$	$9.5628e-5$	$\mathbf{1.5229e-4}$	$1.1822e-4$
$\hat{\beta}_p^{min}(\mathbf{w},2)$	**32**	13	**12**	13
$\gamma(\mathbf{w},2)$	$\mathbf{7.7038e+0}$	$1.2447e+1$	$\mathbf{1.5875e+1}$	$1.3782e+1$
$\hat{\beta}_r^{min}(\mathbf{w},2)$	**3**	4	**4**	4
$\rho(\mathbf{w},3)$	$\mathbf{1.4347e-11}$	$3.2975e-6$	$3.6938e-7$	$\mathbf{3.5012e-6}$
$\hat{\beta}^{min}(\mathbf{w},3)$	**38**	20	23	**20**
$\mu(\mathbf{w},3)$	$\mathbf{6.5127e-11}$	$2.7666e-5$	$2.9985e-6$	$\mathbf{3.0083e-5}$
$\hat{\beta}_p^{min}(\mathbf{w},3)$	**33**	15	18	**15**
$\gamma(\mathbf{w},3)$	$\mathbf{4.5395e+0}$	$8.3902e+0$	$8.1176e+0$	$\mathbf{8.5923e+0}$
$\hat{\beta}_r^{min}(\mathbf{w},3)$	**3**	4	4	**4**

Table 2: Measures and estimated minimum bit lengths of example 2.

Realization	Format	β^{min}	β_p^{min}	β_r^{min}
$\mathbf{w}_0$	fixed	23	12	10
$\mathbf{w}_{\text{opt}}(1)$	fixed	22	18	3
$\mathbf{w}_0$	floating	16	10	5
$\mathbf{w}_{\text{opt}}(2)$	floating	12	6	5
$\mathbf{w}_0$	block	28	22	5
$\mathbf{w}_{\text{opt}}(3)$	block	23	20	2

Table 3: True minimum bit length results of example 1.

Realization	Format	β^{min}	β_p^{min}	β_r^{min}
$\mathbf{w}_0$	fixed	31	21	9
$\mathbf{w}_{\text{opt}}(1)$	fixed	19	10	8
$\mathbf{w}_0$	floating	33	29	3
$\mathbf{w}_{\text{opt}}(2)$	floating	13	8	4
$\mathbf{w}_0$	block	33	30	2
$\mathbf{w}_{\text{opt}}(3)$	block	16	12	3

Table 4: True minimum bit length results of example 2.