A Unified Closed-Loop Stability Measure for Finite-Precision Digital Controller Realizations Implemented in Different Representation Schemes

Jun Wu †, Sheng Chen ‡ [0], James F. Whidborne § 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.

§ Department of Mechanical Engineering
King's College London
Strand, London WC2R 2LS, U.K.

Abstract

The closed-loop stability issue of finite-precision realizations is investigated for digital controllers implemented in three different arithmetic formats, namely fixed-point, floating-point and block-floating-point schemes. It is shown that the controller coefficient perturbations resulting from using different finite word length (FWL) representation schemes possess quite different properties. A unified FWL closed-loop stability measure is derived which is applicable to all the three arithmetic schemes. Unlike the existing works which only take into account the precision of a representation scheme with an assumption of an unlimited dynamic range, both the dynamic range and precision of an arithmetic scheme are considered in this new unified measure. To facilitate the design of optimal finite-precision controller realizations, a computationally tractable FWL closed-loop stability measure is then introduced and the method of computing the value of this measure for a given controller realization is given. For each arithmetic scheme, the optimal controller realization is defined as the solution that maximizes the corresponding measure, and a numerical optimization approach is adopted to solve for the resulting optimal realization problem. The proposed design procedure provides a unified framework for true optimal controller implementation that requires a minimum bit length with maximum robustness to the FWL effect. Numerical examples are used to illustrate the design procedure and to compare the optimal controller realizations in different representation schemes.

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

Index Terms — digital controller, finite word length, fixed-point, floating-point, block-floating-point, closed-loop stability, optimization.

1 Introduction

There has been a growing awareness that the effect of finite-precision implementation can have a serious influence on the actual performance of a designed controller. Modern digital processors have high precision, e.g. 16-bit fixed-point processor or even 32-bit floating-point processor. At the first thought, it may seem that the perturbations resulting from finite-precision computing of the digital controller are so small, compared to the uncertainty within the plant model, such that this controller "uncertainty" can simply be ignored. However, it has increasingly been realized that this is not necessarily the case. Due to the FWL effect, a casual 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. The effects of finite-precision computation have become more critical with the growing popularity of robust controller design methods which focus only on dealing with large plant uncertainty and result in controllers of much higher order and complexity than traditional classical control, as is highlighted in the so-called fragility puzzles [1],[2].

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, such as fixed-point, floating-point or block-float-point format. In other words, the implementation of a digital controller includes the choices of controller realization and number representation format. The dynamic range of number representation is defined by the integer or exponent part of the fixed-point or floating-point format, respectively, while the accuracy or precision is determined by the fractional or mantissa part, respectively. If the bit length can be infinite, that is, the dynamic range is unlimited and the precision is infinite, all the different realizations of a controller implemented in any one of the three representation schemes will be exactly equivalent. However, different controller realizations have different degrees of "robustness" to FWL errors. This property can be utilized to select "optimal" realizations that optimize some given criteria, and several works [3]–[5] have studied many aspects of FWL digital controller design, particularly the critical issue of FWL closed-loop stability robustness.

The fixed-point arithmetic has advantages of computational simplicity and efficiency over the floating-point arithmetic, but has much limited dynamic range and precision given the same bit length. It is therefore not surprising that most of the previous works have focused on finding optimal controller realizations using fixed-point arithmetic by maximizing some closed-loop stability measures [4]–[14]. With decreasing in price and increasing in availability, the use of floating-point processors in controller implementations has increased dramatically. Few studies so far have addressed digital controller implementation in floating-point format with FWL considerations. An exception is the recent work [15] which has studied explicitly the closed-loop stability issue of floating-point digital controller realizations. The block-floating-point format has well been studied in the content of digital filter designs [16],[17]. The work [18] has studied finite-precision controller realizations in a block-floating-point format. However, the approach adopted in [18] is to convert the problem into a fixed-point one, and the true block-floating-point FWL closed-loop stability measure has not been seen to date.

In all the previous works dealing with FWL digital controller realizations [4]–[15],[18], various FWL closed-loop stability measures are maximized to produce "optimal" realizations. However, there is a "flaw" here. These measures are only linked to the bits required to implement the fractional part of fixed-point representation or the mantissa part of floating-point representation, depending on which format is used. Maximizing these measures, while minimizing the bits required for the fractional or mantissa part, may actually increase the bits required for the integer or exponent part. Thus, the resulting "optimal" controller realizations are not necessarily true optimal in terms of the robustness to FWL effects. It is a common sense that a desired controller realization should be balanced, that is, has a small dynamic range. This is simply because in a digital processor the total bit length has to accommodate the dynamic range first to avoid overflow and/or underflow, and the remaining bits left are then used to implement the fractional or mantissa part of a representation scheme. Therefore, given a fixed bit length, a smaller dynamic range means a higher accuracy or precision, and an optimal controller realization should minimize the total bit length required.

The contribution of this paper is triple. Firstly, a unified FWL closed-loop stability measure is derived which is applicable to all the three arithmetic schemes. Secondly and most importantly, this new unified measure accommodates both the dynamic range and precision requirements and is directly linked to the total bit length of a representation format. Thirdly, based on the unified measure, it is convenient to compare optimal controller realizations in different representation schemes and to provide useful information in selecting a suitable controller realization and arithmetic format. The remainder of this paper is organized as follows. Section 2 briefly summarizes the three number representation schemes and highlights the different properties of perturbations resulting from different FWL representation schemes. Section 3 analyses the FWL effect of each arithmetic scheme on closed-loop stability and provides a unified measure for FWL implemented digital controllers. Section 4 defines a computationally tractable

FWL closed-loop stability measure for controller realizations implemented in any presentation scheme and provides the method of computing its value. In section 5, the optimal FWL controller realization problem is formulated based on the unified measure, and a numerical optimization technique is adopted to solve for the resulting optimization problem. Two examples are given in section 6 to demonstrate the effectiveness of the proposed design method and to compare different optimal realizations based on different arithmetic schemes. The paper concludes at section 7.

2 Representation Schemes

There are three representation schemes available for digital processors to store numerical data in memory and in registers. These are fixed-point, floating-point and block-floating-point representations. All digital processors have finite word lengths, and the precise value of a number represented in one scheme can be different from that in another scheme. Furthermore, the perturbations resulting from different FWL representation schemes have very different properties.

2.1 Fixed-Point Representation

The fixed-point format for representing $x \in \mathcal{R}$ with a bit length

$$\beta = 1 + \beta_g + \beta_f \tag{1}$$

assigns 1 bit for the sign, β_g bits for the integer part and β_f bits for the fraction part. Denote the sign of x as

$$s \triangleq \begin{cases} 0, & \text{for } \ x \geq 0, \\ 1, & \text{for } \ x < 0. \end{cases} \tag{2}$$

Then, with the two's complement system, the set of all the possible fixed-point numbers that can be presented by the bit length β is given by

$$\mathcal{F}_1 \triangleq \left\{ \sum_{j=-\beta_f}^{\beta_g - 1} b_j 2^j - 2^{\beta_g} s : \ s \in \{0,1\}, \ b_j \in \{0,1\} \right\}. \tag{3}$$

When no overflow occurs, that is, $|x| \leq 2^{\beta_g}$, the fixed-point quantization operator $\mathcal{Q}_1 : \mathcal{R} \rightarrow \mathcal{F}_1$ is defined as

$$\mathcal{Q}_1(x) \triangleq (-1)^s 2^{-\beta_f} \lfloor 2^{\beta_f} |x| + 0.5 \rfloor \tag{4}$$

where $\lfloor \cdot \rfloor$ denotes the *floor* function, i.e., $\lfloor x \rfloor$ is the closest integer less than or equal to $x \in \mathcal{R}$. The quantization error of fixed-point representation is given by

$$\varepsilon_1 \triangleq |x - \mathcal{Q}_1(x)|. \tag{5}$$

Consequently, the quantization error is bounded by

$$\varepsilon_1 < 2^{-(\beta_f+1)}. \tag{6}$$

Thus, when x is implemented in the fixed-point scheme of β_f fraction bits, assuming no overflow, it is perturbed to

$$\mathcal{Q}_1(x) = x + \delta_1, \quad |\delta_1| < 2^{-(\beta_f+1)}. \tag{7}$$

Obviously, the perturbation resulting from FWL fixed-point representation is additive.

2.2 Floating-Point Representation

It is well known that any $x \in \mathcal{R}$ can be expressed uniquely as

$$x = (-1)^s \times w \times 2^e \tag{8}$$

where $s \in \{0, 1\}$ is the sign of x, $w \in [0.5, 1)$ is the mantissa of x, $e \in \mathcal{Z}$ is the exponent of x, and $\mathcal{Z}$ denotes the set of integers. When x is stored in a digital computer of finite β bits in a floating-point format, the bits consists of three parts: one bit for s, β_w bits for w and β_e bits for e. Obviously

$$\beta = 1 + \beta_w + \beta_e. \tag{9}$$

Thus the set of all the floating-point numbers that can be represented by the bit length β is given by

$$\mathcal{F}_2 \triangleq \left\{ (-1)^s \left(0.5 + \sum_{j=1}^{\beta_w} b_j 2^{-(j+1)} \right) \times 2^e : \; s \in \{0, 1\}, \; b_j \in \{0, 1\}, \; e \in \mathcal{Z}, \; \underline{e} \le e \le \overline{e} \right\} \cup \{0\} \tag{10}$$

where $\underline{e}$ and $\overline{e}$ represent the lower and upper limits of the exponent, respectively, and $\overline{e} - \underline{e} = 2^{\beta_e} - 1$. Note that unlike the fixed-point representation, underflow can occur in the floating-point representation.

Denote the set of integers $\underline{e} \le e \le \overline{e}$ as $\mathcal{Z}_{[\underline{e},\, \overline{e}]}$. When no underflow or overflow occurs, that is, the exponent of x is within $\mathcal{Z}_{[\underline{e},\, \overline{e}]}$, the floating-point quantization operator $\mathcal{Q}_2 : \mathcal{R} \to \mathcal{F}_2$ is defined by

$$\mathcal{Q}_2(x) \triangleq \begin{cases} (-1)^s 2^{(e-\beta_w-1)} \lfloor 2^{(\beta_w-e+1)} |x| + 0.5 \rfloor, & \text{for } x \neq 0 \\ 0, & \text{for } x = 0 \end{cases} \tag{11}$$

where the exponent $e = \lfloor \log_2 |x| \rfloor + 1$. The quantization error of floating-point representation is defined as

$$\varepsilon_2 \triangleq |x - \mathcal{Q}_2(x)|. \tag{12}$$

It can be shown easily that the quantization error is bounded by

$$\varepsilon_2 < |x| 2^{-(\beta_w+1)}. \tag{13}$$

Thus, when x is implemented in the floating-point format of β_w mantissa bits, assuming no underflow or overflow, it is perturbed to

$$\mathcal{Q}_2(x) = x + x\delta_2, \ \ |\delta_2| < 2^{-(\beta_w+1)}. \tag{14}$$

It can be seen that the perturbation resulting from FWL floating-point representation is multiplicative, unlike the perturbation resulting from FWL fixed-point representation, which is additive.

2.3 Block-Floating-Point Representation

The fixed-point and floating-point formats are the two basic representation schemes for real numbers. For a group of real numbers stored simultaneously in a digital processor, the so-called block-floating-point format is also available. Suppose that the group of real numbers form a set $\mathcal{S}$. In the block-floating-point format, $\mathcal{S}$ is divided into some blocks. The block-floating-point scheme may be viewed as aiming to achieve a trade-off between the simplicity of fixed-point scheme and the accuracy of floating-point scheme. The difficulty is how to divide $\mathcal{S}$ into suitable blocks. If such a division is done inappropriately, the result obtained can be worse than the fixed-point scheme.

For an illustrative purpose and without the loss of generality, consider the case of dividing $\mathcal{S}$ into two non-empty subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, which satisfy $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{S}$ and $\mathcal{S}_1 \cap \mathcal{S}_2$ is the empty set. 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

$$x = (-1)^s \times u \times 2^h \tag{15}$$

where $s \in \{0, 1\}$ is the sign of x, $u \in [0,\ 1)$ is the block mantissa of x, and the block exponent of x is

$$h \triangleq \begin{cases} \lfloor \log_2 |\eta_1| \rfloor + 1, & \text{for} \ \ x \in \mathcal{S}_1, \\ \lfloor \log_2 |\eta_2| \rfloor + 1, & \text{for} \ \ x \in \mathcal{S}_2. \end{cases} \tag{16}$$

Obviously, all the elements in the same block have the same exponent value of h. When all the elements in $\mathcal{S}$ are stored in a digital processor of the bit length

$$\beta = 1 + \beta_u + \beta_h \tag{17}$$

in a block-floating-point scheme, 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. Thus the set of all the block-floating-point numbers that can be represented by the bit length β is given by

$$\mathcal{F}_3 \triangleq \left\{ \left(\sum_{j=1}^{\beta_u} b_j 2^{-j} - s \right) \times 2^h : \ s \in \{0, 1\}, \ b_j \in \{0, 1\}, \ h \in \mathcal{Z}, \ \underline{h} \le h \le \overline{h} \right\} \tag{18}$$

where $\underline{h}$ and $\overline{h}$ represent the lower and upper limits of the block exponent, respectively, and $\overline{h} - \underline{h} = 2^{\beta_h} - 1$. Similar to the floating-point representation, overflow or underflow can occur in the block-floating-point representation.

When no underflow or overflow occurs, that is, h is within $\mathcal{Z}_{[\underline{h},\, \overline{h}]}$, the block-floating-point quantization operator $\mathcal{Q}_3 : \mathcal{S} \to \mathcal{F}_3$ is defined as

$$\mathcal{Q}_3(x) \triangleq (-1)^s 2^{(h-\beta_u)} \lfloor 2^{(\beta_u - h)}|x| + 0.5 \rfloor. \tag{19}$$

The quantization error of the block-floating-point representation is defined as

$$\varepsilon_3 \triangleq |x - \mathcal{Q}_3(x)|. \tag{20}$$

Denote

$$r(x) \triangleq \begin{cases} 2\eta_1, & \text{for } \ x \in \mathcal{S}_1, \\ 2\eta_2, & \text{for } \ x \in \mathcal{S}_2. \end{cases} \tag{21}$$

It can be shown easily that the quantization error is bounded by

$$\varepsilon_3 < |r(x)| 2^{-(\beta_u+1)}. \tag{22}$$

Thus, when $x \in \mathcal{S}$ is implemented in the block-floating-point format of β_u block mantissa bits, assuming no underflow or overflow, it is perturbed to

$$\mathcal{Q}_3(x) = x + r(x)\delta_3, \ \ |\delta_3| < 2^{-(\beta_u+1)}. \tag{23}$$

It can be seen that the perturbation resulting from FWL block-floating-point representation is neither multiplicative nor additive. The perturbation depends on the set $\mathcal{S}$ and how $\mathcal{S}$ is divided into blocks.

It is easily seen that in each representation format the total bit length always consists of three parts. Sign occupies one bit. The dynamic range of representation is defined by β_g, β_e or β_h bits, and the precision of representation is determined by β_f, β_w or β_u bits, depending on which scheme is actually chosen. For the notational conciseness, we introduce the "generalized" dynamic range word length β_r and precision word 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

Consider the generic discrete-time closed-loop control system depicted in Figure 1, where the linear time-invariant plant P is described by

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases} \tag{24}$$

which is completely state controllable and observable 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 digital controller C is described by

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{H}\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases} \tag{25}$$

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 $\mathbf{H} \in \mathcal{R}^{m\times p}$. The output-feedback and observer-based controllers can be unified in this general structure: C is an output-feedback controller when $\mathbf{H} = 0$; a full-order observer-based controller when $\mathbf{F} = \mathbf{A} - \mathbf{GC}$, $\mathbf{M} = 0$ and $\mathbf{H} = \mathbf{B}$; a reduced-order observer-based controller, otherwise [19],[20].

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

$$\mathcal{S}_C \triangleq \{(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0, \mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0, \mathbf{H} = \mathbf{T}^{-1}\mathbf{H}_0\} \tag{26}$$

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

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{w}_F \\ \mathbf{w}_G \\ \mathbf{w}_J \\ \mathbf{w}_M \\ \mathbf{w}_H \end{bmatrix} \text{ and } \mathbf{w}_0 \triangleq \begin{bmatrix} \mathbf{w}_{F_0} \\ \mathbf{w}_{G_0} \\ \mathbf{w}_{J_0} \\ \mathbf{w}_{M_0} \\ \mathbf{w}_{H_0} \end{bmatrix} \tag{27}$$

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

$$\overline{\mathbf{A}}(\mathbf{w}) = \begin{bmatrix} \mathbf{A} + \mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC} + \mathbf{HMC} & \mathbf{F} + \mathbf{HJ} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_0) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{28}$$

where $\mathbf{0}$ and $\mathbf{I}$ denote the zero and identity matrices of appropriate dimensions, respectively. 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

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| = |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))| < 1, \ \forall i \in \{1, \cdots, m+n\}. \tag{29}$$

Define

$$\|\mathbf{w}\|_{\max} \triangleq \max_{j\in\{1,\cdots,N\}} |w_j|, \tag{30}$$

$$\pi(\mathbf{w}) \triangleq \min_{j\in\{1,\cdots,N\}} \{|w_j| : w_j \neq 0\}, \tag{31}$$

and the index α of representation formats

$$\alpha = \begin{cases} 1, & \text{fixed-point format is adopted,} \\ 2, & \text{floating-point format is adopted,} \\ 3, & \text{block-floating-point format is adopted.} \end{cases} \tag{32}$$

Obviously, the format α indicates the actual representation format used. 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

$$\mathbf{z}(\mathbf{w}) \triangleq [\, \eta_F \quad \eta_G \quad \eta_J \quad \eta_M \quad \eta_H \,]^T \tag{33}$$

with ^{T} being the transpose operator.

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

$$\gamma(\mathbf{w}, \alpha) \triangleq \begin{cases} \|\mathbf{w}\|_{\max}, & \alpha = 1, \\ \log_2 \frac{4\|\mathbf{w}\|_{\max}}{\pi(\mathbf{w})}, & \alpha = 2, \\ \log_2 \frac{4\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}, & \alpha = 3. \end{cases} \tag{34}$$

The rationale of this dynamic range measure becomes clear in the following (obvious) propositions.

Proposition 1 The realization $\mathbf{w}$ can be represented in the fixed-point format of β_g integer bits without overflow, if $2^{\beta_g} \geq \|\mathbf{w}\|_{\max}$.

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

Proposition 3 The realization $\mathbf{w}$ 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 $\mathbf{w}$, does not cause overflow or underflow. The minimum required dynamic-range bit length can easily be computed by

$$\beta_r^{min}(\mathbf{w}, \alpha) = \begin{cases} \lceil \log_2 \|\mathbf{w}\|_{\max} \rceil, & \alpha = 1 \\ \lceil \log_2 (\lfloor \log_2 \|\mathbf{w}\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{w}) \rfloor + 1) \rceil, & \alpha = 2 \\ \lceil \log_2 (\lfloor \log_2 \|\mathbf{z}(\mathbf{w})\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{z}(\mathbf{w})) \rfloor + 1) \rceil, & \alpha = 3 \end{cases} \tag{35}$$

where $\lceil \cdot \rceil$ denotes the *ceil* function, i.e., $\lceil x \rceil$ is the closest integer greater than or equal to $x \in \mathcal{R}$. Note that the measure $\gamma(\mathbf{w}, \alpha)$ defined in (34) provides an estimate of β_r^{min} as

$$\hat{\beta}_r^{min}(\mathbf{w}, \alpha) \triangleq \lceil \log_2 \gamma(\mathbf{w}, \alpha) \rceil. \tag{36}$$

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

$$\tau(\mathbf{x}) \triangleq \begin{cases} 0, & \mathbf{x} \text{ is a zero vector}, \\ 1, & \mathbf{x} \text{ is a nonzero vector}. \end{cases} \tag{37}$$

For two vectors $\mathbf{x} = [x_j]$ and $\mathbf{y} = [y_j]$ of the same dimension, define the Hadamard product of $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x} \circ \mathbf{y} \triangleq [x_j y_j]. \tag{38}$$

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

$$\mathbf{r}(\mathbf{w}, \alpha) \triangleq \begin{cases} \begin{bmatrix} \tau(\mathbf{w}_F)\mathbf{d}(\mathbf{w}_F) \\ \tau(\mathbf{w}_G)\mathbf{d}(\mathbf{w}_G) \\ \tau(\mathbf{w}_J)\mathbf{d}(\mathbf{w}_J) \\ \tau(\mathbf{w}_M)\mathbf{d}(\mathbf{w}_M) \\ \tau(\mathbf{w}_H)\mathbf{d}(\mathbf{w}_H) \end{bmatrix}, & \alpha = 1, \\ \mathbf{w}, & \alpha = 2, \\ \begin{bmatrix} 2\eta_F \mathbf{d}(\mathbf{w}_F) \\ 2\eta_G \mathbf{d}(\mathbf{w}_G) \\ 2\eta_J \mathbf{d}(\mathbf{w}_J) \\ 2\eta_M \mathbf{d}(\mathbf{w}_M) \\ 2\eta_H \mathbf{d}(\mathbf{w}_H) \end{bmatrix}, & \alpha = 3. \end{cases} \tag{39}$$

Each element δ_j of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_p+1)}$, that is,

$$\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_p+1)}. \tag{40}$$

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 $\mathbf{w}$.

It is therefore critical to know when the FWL error will cause closed-loop instability. This means that we would like to know the largest open "cube" in the perturbation space within which the closed-loop system remains stable. Based on this consideration, a precision measure for realization $\mathbf{w}$ of format α can be defined as

$$\mu_0(\mathbf{w}, \alpha) \triangleq \inf\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w}, \alpha) \circ \mathbf{\Delta}) \text{ is unstable}\}. \tag{41}$$

From the above definition, the following proposition is obvious.

Proposition 4 $\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w}, \alpha) \circ \boldsymbol{\Delta})$ is stable if $\|\boldsymbol{\Delta}\|_{\max} < \mu_0(\mathbf{w}, \alpha)$.

Thus under the condition that the dynamic range is sufficient, that is, $\beta_r \geq \beta_r^{min}$, the perturbation $\|\boldsymbol{\Delta}\|_{\max}$ and therefore the precision bit length β_p determines whether the closed-loop remains stable. Let β_p^{min} be the precision bit length such that $\forall \beta_p \geq \beta_p^{min}$, the closed-loop system is stable with $\mathbf{w}$ implemented by β_p precision bits and the closed-loop system is unstable with $\mathbf{w}$ implemented by $\beta_p^{min} - 1$ precision bits. Unlike β_r^{min} which can easily be computed for a given realization $\mathbf{w}$, β_p^{min} is generally unknown and can only be determined through simulation. It should be pointed out that due to the complex nonlinear relationship between β_p and closed-loop stability, there may exist some odd cases of smaller precision bit length $\beta_p < \beta_p^{min} - 1$ which regain closed-loop stability. For example, consider the following hypothetical system. When $\beta_p \geq 9$, the closed-loop system is stable, but the closed-loop system becomes unstable with $\beta_p = 8$. However, with $\beta_p = 7$, the closed-loop regains stability. The system becomes unstable again for $\beta_p \leq 6$. For this system, β_p^{min} is 9 rather than 7. The precision measure $\mu_0(\mathbf{w}, \alpha)$ provides an estimate of β_p^{min} as

$$\hat{\beta}_{p0}^{min}(\mathbf{w}, \alpha) \triangleq -\lfloor \log_2 \mu_0(\mathbf{w}, \alpha) \rfloor - 1 \,. \tag{42}$$

It can be seen that $\hat{\beta}_{p0}^{min} \geq \beta_p^{min}$.

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

$$\beta^{min} \triangleq \beta_r^{min} + \beta_p^{min} + 1 \,. \tag{43}$$

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 $\mu_0(\mathbf{w}, \alpha)$ results in the following true FWL closed-loop stability measure for the given realization $\mathbf{w}$ with format α

$$\rho_0(\mathbf{w}, \alpha) \triangleq \mu_0(\mathbf{w}, \alpha) / \gamma(\mathbf{w}, \alpha) \,. \tag{44}$$

An estimate of β^{min} is given by $\rho_0(\mathbf{w}, \alpha)$ as

$$\hat{\beta}_0^{min}(\mathbf{w}, \alpha) \triangleq -\lfloor \log_2 \rho_0(\mathbf{w}, \alpha) \rfloor + 1 \,. \tag{45}$$

It is clear that $\hat{\beta}_0^{min} \geq \beta^{min}$. The following proposition summarizes the usefulness of $\rho_0(\mathbf{w}, \alpha)$ as a measure for the FWL characteristics of $\mathbf{w}$ in representation format α.

Proposition 5 The controller realization $\mathbf{w}$ implemented in format α with a bit length β can guarantee a sufficient dynamic range and closed-loop stability, if

$$2^{\beta - 1} \geq \rho_0(\mathbf{w}, \alpha) \,. \tag{46}$$

For a given format α, the closed-loop stability measure $\rho_0(\mathbf{w}, \alpha)$ depends on the controller realization $\mathbf{w}$ only. Consequently, an optimal realization can in theory be found by maximizing $\rho_0(\mathbf{w}, \alpha)$ over $\mathcal{S}_C$, leading to the following optimal controller realization problem

$$v_{\text{true}}(\alpha) \stackrel{\triangle}{=} \max_{\mathbf{w} \in \mathcal{S}_C} \rho_0(\mathbf{w}, \alpha) . \tag{47}$$

However, the difficulty with this approach is that computing the value of $\mu_0(\mathbf{w}, \alpha)$ is an unsolved open problem. Thus, the true FWL closed-loop stability measure $\rho_0(\mathbf{w}, \alpha)$ and the optimal realization problem (47) have limited practical significance. In the next section, an alternative measure is developed which not only can quantify FWL characteristics of $\mathbf{w}$ in format α but also is computationally tractable.

4 A Tractable FWL Closed-Loop Stability Measure

When the FWL error $\mathbf{\Delta}$ is small, from a first-order approximation, $\forall i \in \{1, \cdots, m+n\}$

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w}, \alpha) \circ \mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| \approx \sum_{j=1}^{N} \left. \frac{\partial |\lambda_i|}{\partial \delta_j} \right|_{\mathbf{\Delta}=\mathbf{0}} \delta_j . \tag{48}$$

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

$$\left\| \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right\|_1 \stackrel{\triangle}{=} \sum_{j=1}^{N} \left| \frac{\partial |\lambda_i|}{\partial \delta_j} \right| . \tag{49}$$

Then

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w}, \alpha) \circ \mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| \le \|\mathbf{\Delta}\|_{\max} \left\| \left. \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right|_{\mathbf{\Delta}=\mathbf{0}} \right\|_1 . \tag{50}$$

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

$$\mu_1(\mathbf{w}, \alpha) \stackrel{\triangle}{=} \min_{i \in \{1, \cdots, m+n\}} \frac{1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\left\| \left. \frac{\partial |\lambda_i|}{\partial \mathbf{\Delta}} \right|_{\mathbf{\Delta}=\mathbf{0}} \right\|_1} . \tag{51}$$

Obviously, if $\|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{w}, \alpha)$, 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 $\mu_1(\mathbf{w}, \alpha)$. The larger $\mu_1(\mathbf{w}, \alpha)$ is, the larger FWL errors the closed-loop system can tolerate. Similar to (42), from the precision measure $\mu_1(\mathbf{w}, \alpha)$, an estimate of β_p^{min} is given as

$$\hat{\beta}_{p1}^{min}(\mathbf{w}, \alpha) \stackrel{\triangle}{=} -\lfloor \log_2 \mu_1(\mathbf{w}, \alpha) \rfloor - 1 . \tag{52}$$

The assumption of small $\mathbf{\Delta}$ is usually valid in practical implementation of digital controllers. Generally speaking, there is no rigorous relationship between $\mu_0(\mathbf{w}, \alpha)$ and $\mu_1(\mathbf{w}, \alpha)$, but $\mu_1(\mathbf{w}, \alpha)$ is connected with a lower bound of $\mu_0(\mathbf{w}, \alpha)$ in some manners: there are "stable perturbation cubes" larger than

$\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{w},\alpha)\}$ while there is no "stable perturbation cube" larger than $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{w},\alpha)\}$ [11],[13]. Hence, in most cases, it is reasonable to take that $\mu_1(\mathbf{w},\alpha) \leq \mu_0(\mathbf{w},\alpha)$ and $\hat{\beta}_{p1}^{min} \geq \hat{\beta}_{p0}^{min}$. More importantly, unlike the measure $\mu_0(\mathbf{w},\alpha)$, the value of $\mu_1(\mathbf{w},\alpha)$ can be computed explicitly. It is easy to see that

$$\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}} = \mathbf{r}(\mathbf{w},\alpha) \circ \frac{\partial|\lambda_i|}{\partial\mathbf{w}} \tag{53}$$

and from the results of [13], it can be shown that

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{F}} = [\mathbf{0} \quad \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} \tag{54}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{G}} = [\mathbf{0} \quad \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\\ \mathbf{0}\end{bmatrix} \tag{55}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{J}} = [\mathbf{B}^T \quad \mathbf{H}^T]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} \tag{56}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{M}} = [\mathbf{B}^T \quad \mathbf{H}^T]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\\ \mathbf{0}\end{bmatrix} \tag{57}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{H}} = [\mathbf{0} \quad \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\mathbf{M}^T\\ \mathbf{J}^T\end{bmatrix} \tag{58}$$

with

$$\mathbf{L}_i(\mathbf{w}) = \frac{\mathrm{Re}[\lambda_i^*(\overline{\mathbf{A}}(\mathbf{w}))\mathbf{y}_i^*(\overline{\mathbf{A}}(\mathbf{w}))\mathbf{p}_i^T(\overline{\mathbf{A}}(\mathbf{w}))]}{|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|} \tag{59}$$

where $\mathbf{p}_i(\overline{\mathbf{A}}(\mathbf{w}))$ and $\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}))$ are the right and reciprocal left eigenvectors related to the $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$, respectively, * denotes the conjugate operation and $\mathrm{Re}[\cdot]$ the real part.

Replacing $\mu_0(\mathbf{w},\alpha)$ with $\mu_1(\mathbf{w},\alpha)$ in (44) leads to a computationally tractable FWL closed-loop stability measure

$$\rho_1(\mathbf{w},\alpha) \stackrel{\triangle}{=} \mu_1(\mathbf{w},\alpha)/\gamma(\mathbf{w},\alpha)\,. \tag{60}$$

From the measure $\rho_1(\mathbf{w},\alpha)$, an estimate of β^{min} is given as

$$\hat{\beta}_1^{min}(\mathbf{w},\alpha) \stackrel{\triangle}{=} -\lfloor\log_2 \rho_1(\mathbf{w},\alpha)\rfloor + 1\,. \tag{61}$$

Compared with the existing FWL measures [4]–[15],[18], $\rho_1(\mathbf{w},\alpha)$ has at least two advantages. Firstly, $\rho_1(\mathbf{w},\alpha)$ can be used in different representation formats while the existing measures are only valid for a particular representation format. For example, the measures presented in [6]–[14] are fixed-point measures and the measure in [15] is a floating-point one. The measure $\rho_1(\mathbf{w},\alpha)$ offers a unified framework to compare the FWL characteristics of a realization $\mathbf{w}$ in different formats. Secondly and critically, unlike the existing measures which are precision measures only and imply an unlimited dynamic range, $\rho_1(\mathbf{w},\alpha)$ is made up of a dynamic range measure and a precision measure and is therefore a true FWL measure capable of handling closed-loop stability as well as the underflow and overflow aspects.

5 Optimization Procedure

In a given format α, different realizations $\mathbf{w}$ yield different values of $\rho_1(\mathbf{w}, \alpha)$. It is of practical importance to find an "optimal" realization $\mathbf{w}_{\text{opt}}(\alpha)$ that maximizes $\rho_1(\mathbf{w}, \alpha)$ for the format α. The controller implemented with this optimal realization $\mathbf{w}_{\text{opt}}(\alpha)$ in format α needs a minimum bit length and has a maximum tolerance to the FWL error. This optimal realization problem is formally defined as

$$\upsilon(\alpha) \stackrel{\triangle}{=} \max_{\mathbf{w}\in\mathcal{S}_C} \rho_1(\mathbf{w}, \alpha)\,. \tag{62}$$

As a close-form or global analytical solution of this optimal realization problem is not available, a numerical optimization approach is adopted to solve for this optimization problem.

Given an initial controller realization $\mathbf{w}_0$ obtained by some standard controller design method, $\forall i \in \{1, \cdots, m+n\}$, partition $\mathbf{p}_i(\overline{\mathbf{A}}(\mathbf{w}_0))$ and $\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_0))$ into

$$\mathbf{p}_i(\overline{\mathbf{A}}(\mathbf{w}_0)) = \begin{bmatrix} \mathbf{p}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)) \\ \mathbf{p}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)) \end{bmatrix} \tag{63}$$

$$\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}_0)) = \begin{bmatrix} \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)) \\ \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)) \end{bmatrix} \tag{64}$$

where $\mathbf{p}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)), \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)) \in \mathcal{C}^n$ and $\mathbf{p}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)), \mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)) \in \mathcal{C}^m$. It is easily seen from (28) that

$$\mathbf{p}_i(\overline{\mathbf{A}}(\mathbf{w})) = \begin{bmatrix} \mathbf{p}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)) \\ \mathbf{T}^{-1}\mathbf{p}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)) \end{bmatrix} \tag{65}$$

$$\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w})) = \begin{bmatrix} \mathbf{y}_{i,1}(\overline{\mathbf{A}}(\mathbf{w}_0)) \\ \mathbf{T}^{T}\mathbf{y}_{i,2}(\overline{\mathbf{A}}(\mathbf{w}_0)) \end{bmatrix}. \tag{66}$$

From (54)–(58), we have

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{F}} = \mathbf{T}^T\mathbf{L}_{i,2,2}(\mathbf{w}_0)\mathbf{T}^{-T} \tag{67}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{G}} = \mathbf{T}^T\mathbf{L}_{i,2,1}(\mathbf{w}_0)\mathbf{C}^T \tag{68}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{J}} = \left(\mathbf{B}^T\mathbf{L}_{i,1,2}(\mathbf{w}_0) + \mathbf{H}_0^T\mathbf{L}_{i,2,2}(\mathbf{w}_0)\right)\mathbf{T}^{-T} \tag{69}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{M}} = \left(\mathbf{B}^T\mathbf{L}_{i,1,1}(\mathbf{w}_0) + \mathbf{H}_0^T\mathbf{L}_{i,2,1}(\mathbf{w}_0)\right)\mathbf{C}^T \tag{70}$$

$$\frac{\partial|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|}{\partial\mathbf{H}} = \mathbf{T}^T\left(\mathbf{L}_{i,2,1}(\mathbf{w}_0)\mathbf{C}^T\mathbf{M}_0^T + \mathbf{L}_{i,2,2}(\mathbf{w}_0)\mathbf{J}_0^T\right) \tag{71}$$

where

$$\mathbf{L}_{i,j,l}(\mathbf{w}_0) = \frac{\mathrm{Re}[\lambda_i^*(\overline{\mathbf{A}}(\mathbf{w}_0))\mathbf{y}_{i,j}^*(\overline{\mathbf{A}}(\mathbf{w}_0))\mathbf{p}_{i,l}^T(\overline{\mathbf{A}}(\mathbf{w}_0))]}{|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))|}, \quad j,l = 1,2\,. \tag{72}$$

Considering that $\mathbf{w}$ is a function of $\mathbf{T}$, $\mathbf{r}(\mathbf{w},\alpha)$ and $\gamma(\mathbf{w},\alpha)$ therefore depend on $\mathbf{T}$ and α, and we define the following optimization criterion in format α:

$$\xi(\mathbf{T},\alpha) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))|}{\left\|\mathbf{r}(\mathbf{w},\alpha)\circ\frac{\partial|\lambda_i|}{\partial\mathbf{w}}\right\|_1 \gamma(\mathbf{w},\alpha)} = \rho_1(\mathbf{w},\alpha). \tag{73}$$

The optimal realization problem (62) can then be posed as the following optimization problem:

$$\upsilon(\alpha) = \max_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m}\\ \det(\mathbf{T})\neq 0}} \xi(\mathbf{T},\alpha). \tag{74}$$

Efficient numerical optimization methods exist for solving for this optimization problem to provide an optimal similarity transformation $\mathbf{T}_{\mathrm{opt}}(\alpha)$. With $\mathbf{T}_{\mathrm{opt}}(\alpha)$, the optimal realization $\mathbf{w}_{\mathrm{opt}}(\alpha)$ in format α can readily be computed. By setting $\alpha = 1$, 2 and 3, respectively, in the optimization problem (74), we can attain the optimal fixed-point realization $\mathbf{w}_{\mathrm{opt}}(1)$, the optimal floating-point realization $\mathbf{w}_{\mathrm{opt}}(2)$ and the optimal block-floating-point realization $\mathbf{w}_{\mathrm{opt}}(3)$ for a digital controller.

It is worth to re-iterate that the optimization problem (74) yields the true optimal controller realization, as the solution $\mathbf{T}_{\mathrm{opt}}(\alpha)$ minimizes the required β_p as well as β_r and therefore minimizes the required total bit length β. This should be compared with the existing "optimal" realization problems [4]–[15],[18], which only try to minimize the required precision bit β_p and, as a consequence, do not necessarily minimize the required total bit length β. As the optimization problem (74) is highly nonlinear, global optimization algorithms, such as the genetic algorithm [21],[22] and adaptive simulated annealing [23],[24], can be adopted. Global optimization methods are however computationally more demanding. Local optimization algorithms, such as Rosenbrock and Simplex algorithms [25],[26], are computationally simpler but run more risks of only attaining a local solution. Our experience with the optimization problem (74) suggests that using a local search algorithm is often adequate. Unlike optimizing the precision measure $\mu_1(\mathbf{w},1)$ alone [13], the dynamic range measure $\gamma(\mathbf{w},\alpha)$ in the criterion $\rho_1(\mathbf{w},\alpha)$ helps to bound the solution set and the cost function (73) appears to behave better.

6 Design Examples and Result Comparison

Two examples were used to illustrate the design procedure based on the unified FWL closed-loop stability measure and to compare the resulting optimal realizations in different representation schemes.

Example 1. This example was taken from [13]. The closed-loop system contained a reduced-order observer-based controller. The discrete-time plant was given by

$$\mathbf{A} = \begin{bmatrix} 3.2439e-1 & -4.5451e+0 & -4.0535e+0 & -2.7003e-3 & 0 \\ 1.4518e-1 & 4.9477e-1 & -4.6945e-1 & -3.1274e-4 & 0 \\ 1.6814e-2 & 1.6491e-1 & 9.6681e-1 & -2.2114e-5 & 0 \\ 1.1889e-3 & 1.8209e-2 & 1.9829e-1 & 1.0000e+0 & 0 \\ 6.1301e-5 & 1.2609e-3 & 1.9930e-2 & 2.0000e-1 & 1.0000e+0 \end{bmatrix},$$

$$\mathbf{B} = [\, 1.4518e-1 \quad 1.6814e-2 \quad 1.1889e-3 \quad 6.1301e-5 \quad 2.4979e-6 \,]^T,$$

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1.6188e+0 & -1.5750e-1 & -4.3943e+1 \\ 1.0000e+0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The initial realization of the digital controller was given by

$$\mathbf{F}_0 = \begin{bmatrix} 0 & 1.0000e+0 \\ -9.3303e-1 & 1.9319e+0 \end{bmatrix}, \quad \mathbf{G}_0 = \begin{bmatrix} 4.1814e-2 & 2.7132e+2 \\ 3.9090e-2 & 1.0167e+3 \end{bmatrix},$$

$$\mathbf{J}_0 = [\, 3.0000e-4 \quad 5.0000e-4 \,], \quad \mathbf{M}_0 = [\, 0 \quad 6.1250e-1 \,], \quad \mathbf{H}_0 = \begin{bmatrix} 7.8047e+1 \\ 7.3849e+1 \end{bmatrix}.$$

Based on the proposed unified FWL closed-loop stability measure, the optimization problem (74) was formed. Using the MATLAB routine *fminsearch.m*, this optimization problem was solved for $\alpha = 1$, 2 and 3, respectively, to obtain the optimal similarity transformation in fixed-point format

$$\mathbf{T}_{\text{opt}}(1) = \begin{bmatrix} -5.7490e+3 & 5.6882e+3 \\ -5.4730e+3 & 5.2451e+3 \end{bmatrix},$$

the optimal similarity transformation in floating-point format

$$\mathbf{T}_{\text{opt}}(2) = \begin{bmatrix} -1.7899e+0 & -3.6899e+0 \\ -8.9324e-1 & -3.5128e+0 \end{bmatrix}$$

and the optimal similarity transformation in block-floating-point format

$$\mathbf{T}_{\text{opt}}(3) = \begin{bmatrix} -1.0126e+2 & -2.7013e+2 \\ 3.6088e+1 & -5.9669e+2 \end{bmatrix}.$$

Example 2. In this example, the discrete-time plant taken from [4] was given by

$$\mathbf{A} = \begin{bmatrix} 3.7156e+0 & -5.4143e+0 & 3.6525e+0 & -9.6420e-1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{B} = [\, 1 \quad 0 \quad 0 \quad 0 \,]^T, \quad \mathbf{C} = [\, 1.1160e-6 \quad 4.3000e-8 \quad 1.0880e-6 \quad 1.4000e-8 \,].$$

The initial realization of the digital controller, which was a modification of the initial output-feedback controller in [4] by a similarity transformation, was given by

$$\mathbf{F}_0 = \begin{bmatrix} 2.6963e+2 & -4.2709e+1 & 2.2873e+1 & 2.6184e+2 \\ 2.5561e+2 & -4.0497e+1 & 2.1052e+1 & 2.4806e+2 \\ 5.6096e+1 & -8.5715e+0 & 5.2162e+0 & 5.4920e+1 \\ -2.3907e+2 & 3.7998e+1 & -2.0338e+1 & -2.3203e+2 \end{bmatrix},$$

$$\mathbf{G}_0 = \begin{bmatrix} -4.6765e+1 \\ -4.5625e+1 \\ -9.5195e+0 \\ 4.1609e+1 \end{bmatrix},$$

$$\mathbf{J}_0 = [\,-2.5548e+2 \quad -2.7185e+2 \quad -2.7188e+2 \quad 2.7188e+2\,],$$

$$\mathbf{M}_0 = [\,0\,], \quad \mathbf{H}_0 = [\,0 \quad 0 \quad 0 \quad 0\,]^T.$$

Using the same method for Example 1, the three optimal similarity transformations obtained for $\alpha = 1$, 2 and 3 were, respectively,

$$\mathbf{T}_{\rm opt}(1) = \begin{bmatrix} -1.0375e-1 & 1.3036e-1 & 3.8757e-3 & 1.1038e-2 \\ -1.1120e-1 & 1.1701e-1 & 1.1202e-2 & 1.4748e-2 \\ -2.3751e-2 & 2.3047e-2 & 2.5598e-3 & 4.0090e-3 \\ 9.2406e-2 & -1.1588e-1 & -3.5311e-3 & -9.8824e-3 \end{bmatrix},$$

$$\mathbf{T}_{\rm opt}(2) = \begin{bmatrix} -1.0398e-1 & 1.3104e-1 & 3.8115e-3 & 1.0286e-2 \\ -1.1576e-1 & 1.1656e-1 & 3.1467e-3 & 9.0988e-3 \\ -2.4115e-2 & 2.3070e-2 & 3.0218e-4 & 1.9022e-3 \\ 9.2612e-2 & -1.1647e-1 & -3.3752e-3 & -9.1249e-3 \end{bmatrix},$$

$$\mathbf{T}_{\rm opt}(3) = \begin{bmatrix} -1.0345e-1 & 1.2904e-1 & 3.8315e-3 & 1.0912e-2 \\ -1.1079e-1 & 1.1744e-1 & 3.0183e-3 & 8.1535e-3 \\ -2.3794e-2 & 2.3860e-2 & 4.5654e-4 & 1.8405e-3 \\ 9.2140e-2 & -1.1474e-1 & -3.3981e-3 & -9.6787e-3 \end{bmatrix}.$$

Table 1 lists the values of the measures ρ_1, μ_1 and γ in the three different representation schemes together with the corresponding estimated bit lengths for the initial realization $\mathbf{w}_0$, 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 $\mathbf{w}_{\rm opt}(3)$ of Example 1. Table 2 does the same thing for Example 2. In these two tables, the various estimated bit lengths were computed from their respective measure values according to (36), (52) and (61). Some observations can readily be made from the results in Tables 1 and 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. In fact, for Example 2, we had deliberately chosen $\mathbf{w}_0$ as the transformation of the initial controller realization in [4] by a similarity transformation matrix to favor a fixed-point implementation. 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. The results also show that the proposed optimization procedure is very effective. This can be seen by comparing the values of the FWL closed-loop stability measure for $\mathbf{w}_0$ and $\mathbf{w}_{\text{opt}}(\alpha)$ implemented in a same format α.

It is obvious that the true minimum dynamic-range bit length $\beta_r^{min}(\mathbf{w}, \alpha)$ for a realization $\mathbf{w}$ in format α can directly be obtained by examining the elements of $\mathbf{w}$. The true minimum precision bit length $\beta_p^{min}(\mathbf{w}, \alpha)$ however can only be obtained through simulation. That is, starting from a very large β_p, reduce β_p by one bit and check the closed-loop stability. The process is repeated until there appears closed-loop instability at $\beta_p = \beta_{p_u}$. Then $\beta_p^{min} = \beta_{p_u} + 1$. Table 3 compares the true minimum required bit lengths β_r^{min}, β_p^{min} and β^{min} of the initial realization $\mathbf{w}_0$ implemented in the three different schemes with those of fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$, floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ and block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$ of Example 1. It can be seen that the floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ requires at least 12 bits to ensure closed-loop stability which is much better than minimum 22 bits needed by fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$ or minimum 23 bits needed by block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$. Table 4 summarizes the minimum required bit lengths β_r^{min}, β_p^{min} and β^{min} for fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$, floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ and block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$ of Example 2 together with those for $\mathbf{w}_0$ in the three formats. It can be seen that the floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ needs at least 13 bits to maintain closed-loop stability which is again better than minimum 19 bits needed by fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$ or minimum 16 bits needed by block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$.

Notice that any realization $\mathbf{w} \in \mathcal{S}_C$ implemented in infinite precision (unlimited β_r and infinite β_p) will achieve the exact performance of the infinite-precision implemented $\mathbf{w}_0$, which is the designed controller performance. For this reason, the infinite-precision implemented $\mathbf{w}_0$ is referred to as the ideal controller realization $\mathbf{w}_{\text{ideal}}$. Figure 2 compares the unit impulse response of the first plant output $y_1(k)$ of Example 1 for the ideal controller $\mathbf{w}_{\text{ideal}}$ with those of the 22-bit fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$ (3 integer bits and 18 fractional bits), the 22-bit floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ (5 exponent bits and 16 mantissa bits) and the 22-bit block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$ (2 block exponent bits and 19 block mantissa bits). Figure 3 compares the unit impulse response of $y_1(k)$ for $\mathbf{w}_{\text{ideal}}$ with those of the 24-bit fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$ ($\beta_g = 3$ and $\beta_f = 20$), the 24-bit floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ ($\beta_e = 5$ and $\beta_w = 18$) and the 24-bit block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$ ($\beta_h = 2$ and $\beta_u = 21$) of Example 1. Figure 4 compares the unit impulse response of the plant output $y(k)$ for $\mathbf{w}_{\text{ideal}}$ with those of the 18-bit fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$ ($\beta_g = 8$ and $\beta_f = 9$), the 18-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ ($\beta_e = 4$ and $\beta_w = 13$) and the 18-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ ($\beta_h = 3$ and $\beta_u = 14$) of Example 2. Figure 5 compares the unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$ with those of the 19-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ ($\beta_g = 8$ and $\beta_f = 10$), the 19-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ ($\beta_e = 4$ and $\beta_w = 14$) and the 19-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ ($\beta_h = 3$ and $\beta_u = 15$) of Example 2. It is obvious that in these figures the response with floating-point implemented $\mathbf{w}_{\rm opt}(2)$ is the closest to the ideal performance.

7 Conclusions

We have presented a unified closed-loop stability measure for finite-precision digital controller realizations implemented in different representation schemes. The proposed computationally tractable measure takes into account both the dynamic range and precision of arithmetic schemes, and therefore provides a true measure of FWL characteristics of controller realizations. Based on this unified FWL closed-loop stability measure, the optimal controller realization problems in different representation formats have been formulated, which can easily be solved for using standard numerical optimization algorithms. Simulation results have confirmed that the optimal floating-point controller realization implemented in floating-point format is the best in terms of robustness to FWL errors. The results have also shown that with a proper division of controller coefficients into blocks, block-floating-point implemented realizations can have better robustness to FWL errors than fixed-point implemented ones but choosing an appropriate division of blocks is difficult. These results agrees with the common understanding of the number representation formats. It is well known that fixed-point format is the best in terms of hardware cost, arithmetic operation simplicity and execution speed while floating-point format is the worst. The proposed design procedure provides designer with useful quantitative information regarding finite precision computational properties, namely 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 controller performance.

Acknowledgements

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 the National Natural Science Foundation of China under Grant 60174026.

References

[1] L.H. Keel and S.P. Bhattacharyya, "Robust, fragile, or optimal?" *IEEE Trans. Automatic Control*, Vol.42, No.8, pp.1098–1105, 1997.

[2] P.M. Mäkilä, "Fragility and robustness puzzles," in *Proc. American Control Conf.* (San Diego, CA, USA), June 2-4, 1999, pp.2914–2919.

[3] D. Williamson, *Digital Control and Implementation*. Englewood-Cliffs, NJ: Prentice-Hall, 1991.

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

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

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

[7] A.G. Madievski, B.D.O. Anderson and M. Gevers, "Optimum realizations of sampled-data controllers for FWL sensitivity minimization," *Automatica*, Vol.31, No.3, pp.367–379, 1995.

[8] 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.

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

[10] 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.

[11] J. Wu, S. Chen, G. Li and J. Chu, "Optimal finite-precision state-estimate feedback controller realization of discrete-time systems," *IEEE Trans. Automatic Control*, Vol.45, No.8, pp.1550–1554, 2000.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] K. Kalliojärvi and J. Astola, "Roundoff errors in block-floating-point systems," *IEEE Trans. Signal Processing*, Vol.44, No.4, pp.783–790, 1996.

[17] K.R. Ralev and P.H. Bauer, "Realization of block floating-point digital filters and application to block implementations," *IEEE Trans. Signal Processing*, Vol.47, No.4, pp.1076–1086, 1999.

[18] 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.

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

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

[21] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison Wesley, 1989.

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

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

[24] S. Chen and B.L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, Vol.79, No.1, pp.117–128, 1999.

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

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

	$\mathbf{w}_0$	$\mathbf{w}_{\rm opt}(1)$	$\mathbf{w}_{\rm opt}(2)$	$\mathbf{w}_{\rm opt}(3)$
$\rho_1(\mathbf{w},1)$	$\mathbf{2.5150e-9}$	$\mathbf{1.1386e-7}$	$2.7728e-8$	$1.0861e-7$
$\hat\beta_1^{min}(\mathbf{w},1)$	$\mathbf{30}$	$\mathbf{25}$	27	25
$\mu_1(\mathbf{w},1)$	$\mathbf{2.5569e-6}$	$\mathbf{5.0795e-7}$	$2.5937e-5$	$1.7450e-7$
$\hat\beta_{p1}^{min}(\mathbf{w},1)$	$\mathbf{18}$	$\mathbf{20}$	15	22
$\gamma(\mathbf{w},1)$	$\mathbf{1.0167e+3}$	$\mathbf{4.4612e+0}$	$9.3543e+2$	$1.6066e+0$
$\hat\beta_r^{min}(\mathbf{w},1)$	$\mathbf{10}$	$\mathbf{3}$	10	1
$\rho_1(\mathbf{w},2)$	$\mathbf{1.3134e-7}$	$1.9204e-5$	$\mathbf{1.9593e-5}$	$3.3365e-7$
$\hat\beta_1^{min}(\mathbf{w},2)$	$\mathbf{24}$	17	$\mathbf{17}$	23
$\mu_1(\mathbf{w},2)$	$\mathbf{3.1118e-6}$	$4.3127e-4$	$\mathbf{4.3127e-4}$	$5.4490e-6$
$\hat\beta_{p1}^{min}(\mathbf{w},2)$	$\mathbf{18}$	11	$\mathbf{11}$	17
$\gamma(\mathbf{w},2)$	$\mathbf{2.3692e+1}$	$2.2458e+1$	$\mathbf{2.2012e+1}$	$1.6332e+1$
$\hat\beta_r^{min}(\mathbf{w},2)$	$\mathbf{5}$	5	$\mathbf{5}$	5
$\rho_1(\mathbf{w},3)$	$\mathbf{9.2976e-10}$	$5.3779e-9$	$2.8185e-9$	$\mathbf{1.3362e-8}$
$\hat\beta_1^{min}(\mathbf{w},3)$	$\mathbf{32}$	29	30	$\mathbf{28}$
$\mu_1(\mathbf{w},3)$	$\mathbf{2.1343e-8}$	$5.7385e-8$	$5.7266e-8$	$\mathbf{5.4549e-8}$
$\hat\beta_{p1}^{min}(\mathbf{w},3)$	$\mathbf{25}$	24	24	$\mathbf{24}$
$\gamma(\mathbf{w},3)$	$\mathbf{2.2955e+1}$	$1.0671e+1$	$2.0318e+1$	$\mathbf{4.0823e+0}$
$\hat\beta_r^{min}(\mathbf{w},3)$	$\mathbf{5}$	4	5	$\mathbf{3}$

Table 1: Values of various measures and corresponding estimated bit lengths for the four realizations in three different formats of Example 1.

	$\mathbf{w}_0$	$\mathbf{w}_{\rm opt}(1)$	$\mathbf{w}_{\rm opt}(2)$	$\mathbf{w}_{\rm opt}(3)$
$\rho_1(\mathbf{w},1)$	$\mathbf{1.2312e-10}$	$\mathbf{1.2003e-6}$	$1.0580e-7$	$1.1321e-6$
$\hat{\beta}_1^{min}(\mathbf{w},1)$	$\mathbf{34}$	$\mathbf{21}$	25	21
$\mu_1(\mathbf{w},1)$	$\mathbf{3.3474e-8}$	$\mathbf{2.3082e-4}$	$9.6673e-5$	$2.2287e-4$
$\hat{\beta}_{p1}^{min}(\mathbf{w},1)$	$\mathbf{24}$	$\mathbf{12}$	13	12
$\gamma(\mathbf{w},1)$	$\mathbf{2.7188e+2}$	$\mathbf{1.9231e+2}$	$9.1370e+2$	$1.9687e+2$
$\hat{\beta}_r^{min}(\mathbf{w},1)$	$\mathbf{9}$	$\mathbf{8}$	10	8
$\rho_1(\mathbf{w},2)$	$\mathbf{2.9062e-11}$	$7.6826e-6$	$\mathbf{9.5931e-6}$	$8.5778e-6$
$\hat{\beta}_1^{min}(\mathbf{w},2)$	$\mathbf{37}$	18	$\mathbf{18}$	18
$\mu_1(\mathbf{w},2)$	$\mathbf{2.2389e-10}$	$9.5628e-5$	$\mathbf{1.5229e-4}$	$1.1822e-4$
$\hat{\beta}_{p1}^{min}(\mathbf{w},2)$	$\mathbf{32}$	13	$\mathbf{12}$	13
$\gamma(\mathbf{w},2)$	$\mathbf{7.7038e+0}$	$1.2447e+1$	$\mathbf{1.5875e+1}$	$1.3782e+1$
$\hat{\beta}_r^{min}(\mathbf{w},2)$	$\mathbf{3}$	4	$\mathbf{4}$	4
$\rho_1(\mathbf{w},3)$	$\mathbf{1.4347e-11}$	$3.2975e-6$	$3.6938e-7$	$\mathbf{3.5012e-6}$
$\hat{\beta}_1^{min}(\mathbf{w},3)$	$\mathbf{38}$	20	23	$\mathbf{20}$
$\mu_1(\mathbf{w},3)$	$\mathbf{6.5127e-11}$	$2.7666e-5$	$2.9985e-6$	$\mathbf{3.0083e-5}$
$\hat{\beta}_{p1}^{min}(\mathbf{w},3)$	$\mathbf{33}$	15	18	$\mathbf{15}$
$\gamma(\mathbf{w},3)$	$\mathbf{4.5395e+0}$	$8.3902e+0$	$8.1176e+0$	$\mathbf{8.5923e+0}$
$\hat{\beta}_r^{min}(\mathbf{w},3)$	$\mathbf{3}$	4	4	$\mathbf{4}$

Table 2: Values of various measures and corresponding estimated bit lengths for the four realizations in three different formats of Example 2.

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

Table 3: Comparison of true minimum required bit lengths for $\mathbf{w}_0$ in three representation 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.

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

Table 4: Comparison of true minimum required bit lengths for $\mathbf{w}_0$ in three representation 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 2.

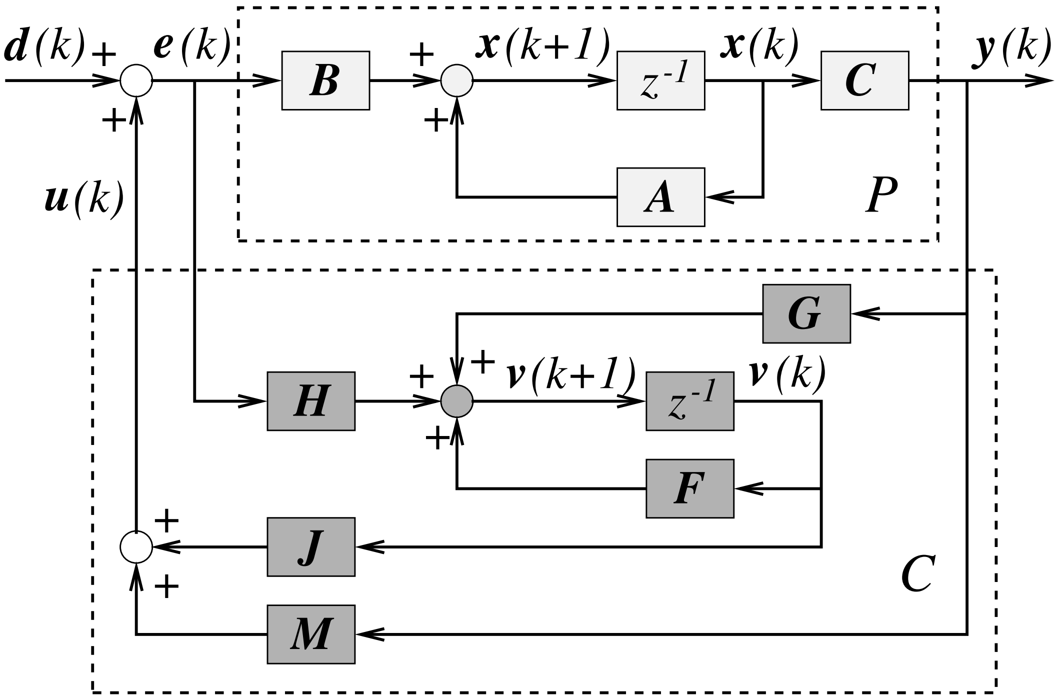

Figure 1: Discrete-time closed-loop system with a genetic digital controller.

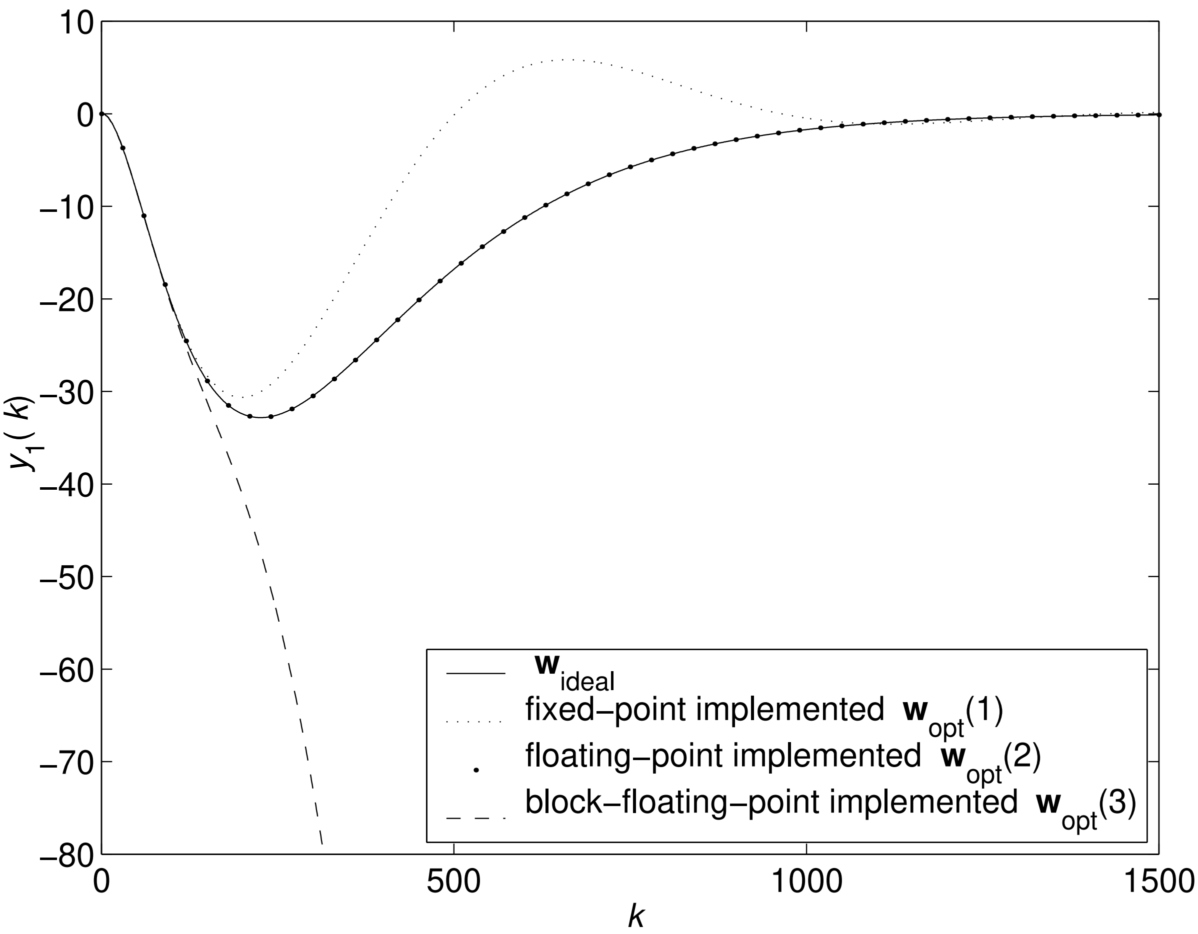

Figure 2: Unit impulse response of $y_1(k)$ for $\mathbf{w}_{\mathrm{ideal}}$, 22-bit fixed-point implemented $\mathbf{w}_{\mathrm{opt}}(1)$ (3 integer bits and 18 fractional bits), 22-bit floating-point implemented $\mathbf{w}_{\mathrm{opt}}(2)$ (5 exponent bits and 16 mantissa bits) and 22-bit block-floating-point implemented $\mathbf{w}_{\mathrm{opt}}(3)$ (2 block exponent bits and 19 block mantissa bits) of Example 1.

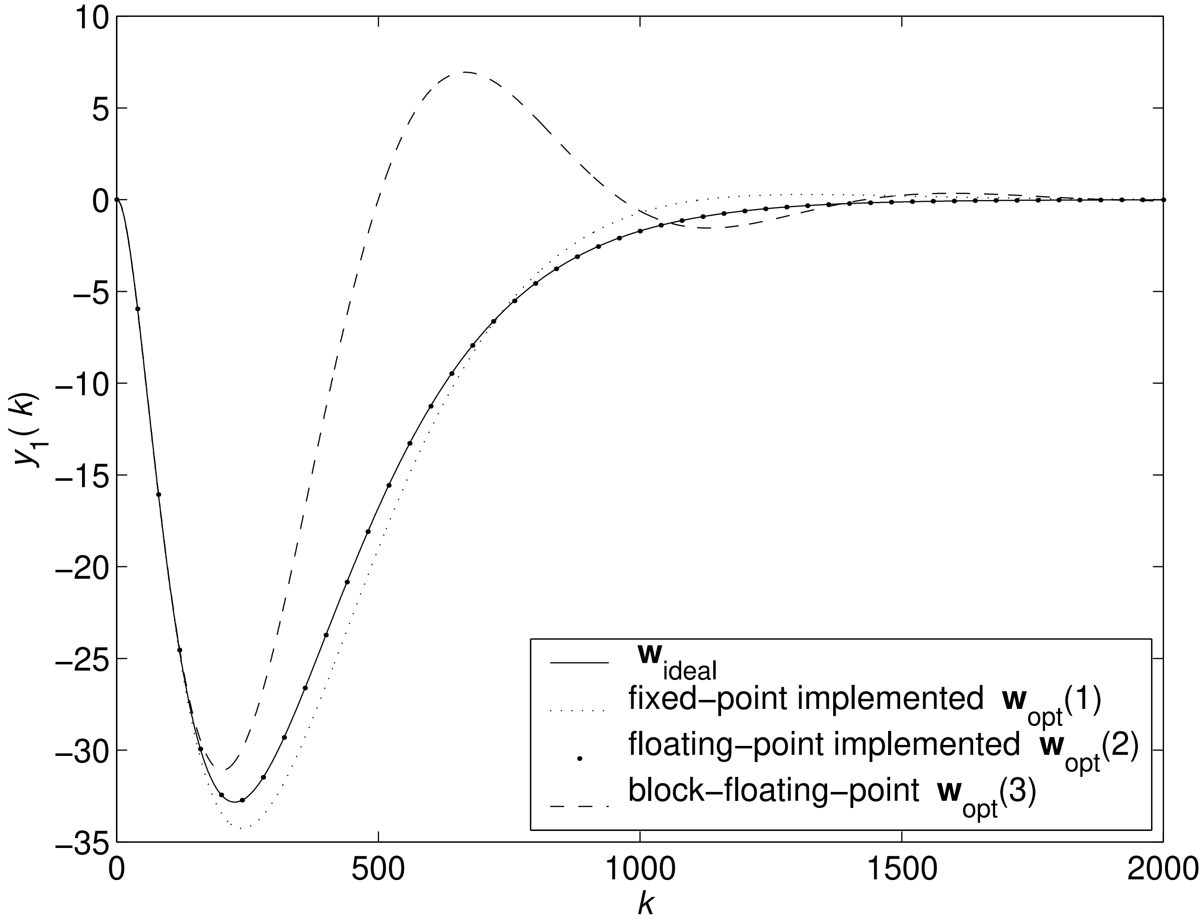

Figure 3: Unit impulse response of $y_1(k)$ for $\mathbf{w}_{\mathrm{ideal}}$, 24-bit fixed-point implemented $\mathbf{w}_{\mathrm{opt}}(1)$ (3 integer bits and 20 fractional bits), 24-bit floating-point implemented $\mathbf{w}_{\mathrm{opt}}(2)$ (5 exponent bits and 18 mantissa bits) and 24-bit block-floating-point implemented $\mathbf{w}_{\mathrm{opt}}(3)$ (2 block exponent bits and 21 block mantissa bits) of Example 1.

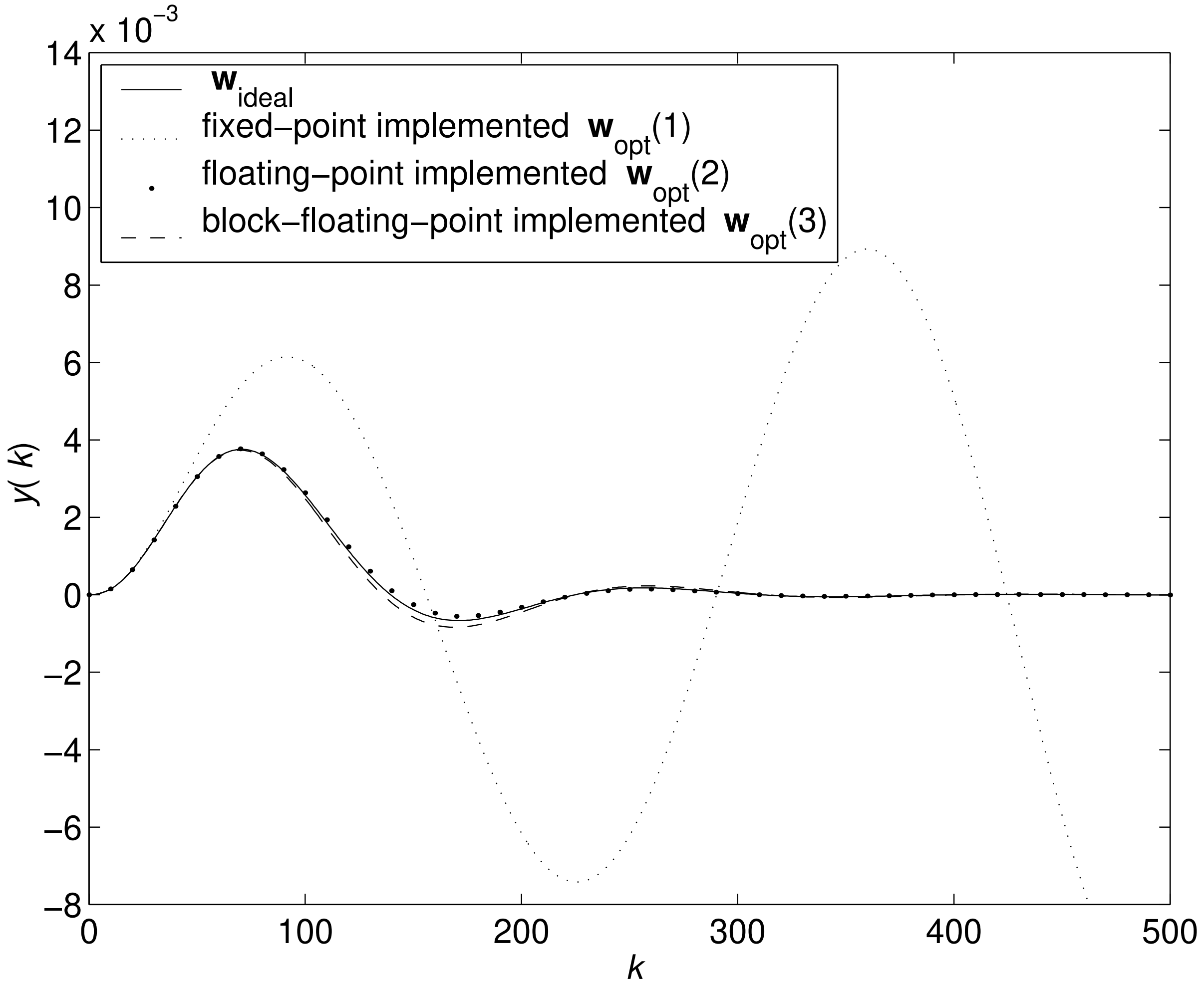

Figure 4: Unit impulse response of $y(k)$ for $\mathbf{w}_{\mathrm{ideal}}$, 18-bit fixed-point implemented $\mathbf{w}_{\mathrm{opt}}(1)$ (8 integer bits and 9 fractional bits), 18-bit floating-point implemented $\mathbf{w}_{\mathrm{opt}}(2)$ (4 exponent bits and 13 mantissa bits) and 18-bit block-floating-point implemented $\mathbf{w}_{\mathrm{opt}}(3)$ (3 block exponent bits and 14 block mantissa bits) of Example 2.

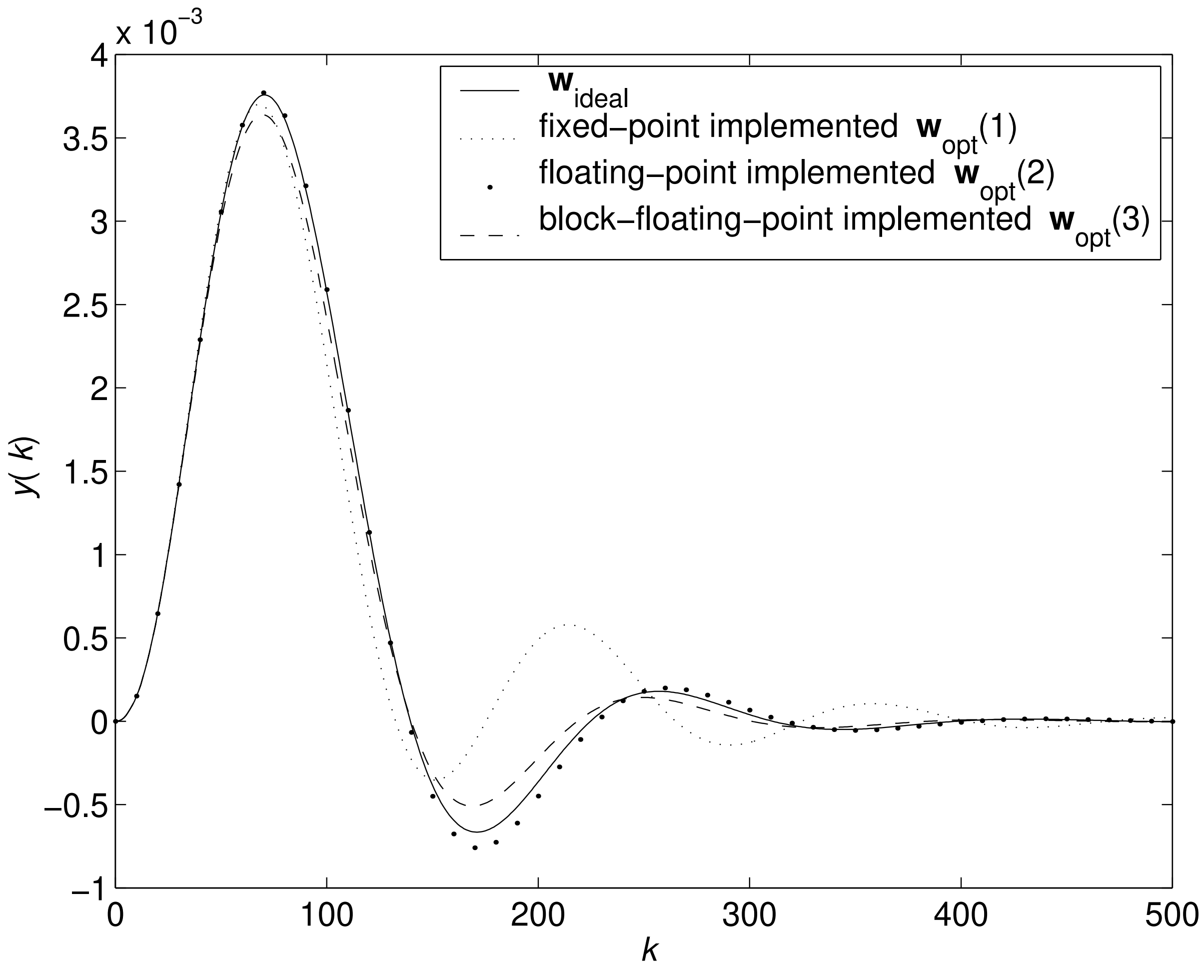

Figure 5: Unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$, 19-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ (8 integer bits and 10 fractional bits), 19-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ (4 exponent bits and 14 mantissa bits) and 19-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ (3 block exponent bits and 15 block mantissa bits) of Example 2.