Optimal Floating-Point Realizations of Finite-Precision Digital Controllers

Jun Wu [†], Sheng Chen [‡], 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, Southampton SO17 1BJ, U.K.

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

Abstract

The paper investigates the closed-loop stability issue of finite-precision realizations for digital controllers implemented in floating-point arithmetic. Unlike the existing methods which only address the effect of the mantissa bits in floating-point format to the sensitivity of closed-loop stability, the sensitivity of closed-loop stability is analyzed with respect to both the mantissa and exponent bits of floating-point format. A computationally tractable finite word length (FWL) closed-loop stability measure is defined, and the optimal controller realization problem is posed as searching for a floating-point realization that maximizes the proposed measure. A numerical optimization approach is adopted to solve for the resulting optimization problem. Simulation results show that the proposed design procedure yields computationally efficient controller realizations with enhanced FWL closed-loop stability performance.

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

I. Introduction

The classical digital controller design methodology often assumes that the controller is implemented exactly. Indeed, it may seem that the controller "uncertainty" resulting from finite-precision computation is so small, compared to the uncertainty within the plant, 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 FWL effect has become more critical with the growing popularity of robust controller design methods which focus sole on dealing with large plant uncertainty [1]. It is well known that a control law can be implemented with different realizations and different realizations possess different degrees of "robustness" to FWL errors. This property can be utilized to design "optimal" controller realizations [2],[3].

Many previous studies have focused on finding optimal controller realizations using fixed-point arithmetic [4]–[10]. However, FWL closed-loop stability measures in all these previous works only consider the fractional part of fixed-point format. Maximizing these measures, while minimizing the bits required for the fractional part, may actually increase the bits required for the integer part of fixed-point format [7],[8]. Arguably, a better approach would be to consider some measure which is linked to the total bit length required. There has been little work studying explicitly the closed-loop stability issue of FWL floating-point digital controller implementations. An exception is the work [11], in which a weighted closed-loop eigenvalue sensitivity index was defined for floating-point digital controller realizations. This FWL measure, however, only considers the mantissa part of floating-point arithmetic, under an assumption that the exponent bits are unlimited. The main contribution of this paper is to derive a new FWL closed-loop stability measure that explicitly considers both the mantissa and exponent parts of floating-point arithmetic.

II. Floating-Point Representation

Any real number $x \in \mathcal{R}$ can be represented uniquely by:

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

where $s \in \{0,1\}$ is for 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 consist of three parts: one bit for s, β_w bits for w and β_e bits for e. Obviously, $\beta = 1 + \beta_w + \beta_e$. The set of all the possible floating-point

Contact author: S. Chen, Tel/Fax: +44 (0)23 8059 6660/4508, Email: sqc@ecs.soton.ac.uk

numbers that can be presented by the bit length β is given by

$$\mathcal{F} \triangleq \{(-1)^s \left(0.5 + \sum_{i=1}^{\beta_w} b_i 2^{-(i+1)}\right) \times 2^e : s \in \{0,1\},$$
$$b_i \in \{0,1\}, e \in \mathcal{Z}, \underline{e} \le e \le \overline{e}\} \cup \{0\} \qquad (2)$$

where $\underline{e}$ and $\overline{e}$ represent the lower and upper limits of the exponent, respectively, and $\overline{e} - \underline{e} = 2^{\beta_e} - 1$.

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 $\mathcal{Q} : \mathcal{R} \rightarrow \mathcal{F}$ can be defined as

$$\mathcal{Q}(x) \triangleq \begin{cases} \operatorname{sgn}(x) 2^{(e-\beta_w-1)} \lfloor 2^{(\beta_w-e+1)} |x| + 0.5 \rfloor, & x \ne 0 \\ 0, & x = 0 \end{cases} \qquad (3)$$

where the exponent $e = \lfloor \log_2 |x| \rfloor + 1$ and the floor function $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. The quantization error is defined as $\varepsilon \triangleq |x - \mathcal{Q}(x)|$. It can easily be shown that the quantization error is bounded by

$$\varepsilon < |x| 2^{-(\beta_w+1)} \,. \qquad (4)$$

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

$$\mathcal{Q}(x) = x(1+\delta), \quad |\delta| < 2^{-(\beta_w+1)} \,. \qquad (5)$$

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

III. Problem Statement

Consider the discrete-time closed-loop control system, consisting of a linear time invariant plant $P(z)$ and a digital controller $C(z)$. $P(z)$ is assumed to be strictly proper with a state-space description $(\mathbf{A}_P, \mathbf{B}_P, \mathbf{C}_P)$, where $\mathbf{A}_P \in \mathcal{R}^{m\times m}$, $\mathbf{B}_P \in \mathcal{R}^{m\times l}$ and $\mathbf{C}_P \in \mathcal{R}^{q\times m}$. Let $(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C)$ be a state-space realization of $C(z)$, with $\mathbf{A}_C \in \mathcal{R}^{n\times n}$, $\mathbf{B}_C \in \mathcal{R}^{n\times q}$, $\mathbf{C}_C \in \mathcal{R}^{l\times n}$ and $\mathbf{D}_C \in \mathcal{R}^{l\times q}$. The realizations of the controller are not unique. In fact, if $(\mathbf{A}_C^0, \mathbf{B}_C^0, \mathbf{C}_C^0, \mathbf{D}_C^0)$ is a realization of $C(z)$, all the realizations of $C(z)$ form a realization set

$$\mathcal{S}_C \triangleq \left\{(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C) : \ \mathbf{A}_C = \mathbf{T}^{-1}\mathbf{A}_C^0\mathbf{T},\right.$$
$$\left.\mathbf{B}_C = \mathbf{T}^{-1}\mathbf{B}_C^0, \mathbf{C}_C = \mathbf{C}_C^0\mathbf{T}, \mathbf{D}_C = \mathbf{D}_C^0\right\} \qquad (6)$$

where the transformation matrix $\mathbf{T} \in \mathcal{R}^{n\times n}$ is an arbitrary non-singular matrix. Denote

$$\mathbf{X} = [x_{j,k}] \triangleq \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix} . \qquad (7)$$

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

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{A}_P + \mathbf{B}_P\mathbf{D}_C\mathbf{C}_P & \mathbf{B}_P\mathbf{C}_C \\ \mathbf{B}_C\mathbf{C}_P & \mathbf{A}_C \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{A}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}$$
$$\triangleq \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2 \qquad (8)$$

where the zero matrix $\mathbf{0}$ has an appropriate dimension. All the different realizations $\mathbf{X}$ in $\mathcal{S}_C$ have exactly the same set of closed-loop poles if they are implemented in infinite precision. Since the closed-loop system has been designed to be stable, all the eigenvalues $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le i \le m+n$, are within the unit disk. Define

$$\|\mathbf{X}\|_{\max} \triangleq \max_{j,k} |x_{j,k}| \qquad (9)$$

$$g(\mathbf{X}) \triangleq \min_{j,k}\{|x_{j,k}| : x_{j,k} \ne 0\} \,. \qquad (10)$$

$\mathbf{X}$ is implemented with a floating-point processor of β_e exponent bits, β_w mantissa bits and one sign bit.

Firstly, in order to avoid underflow and overflow, both the exponents of $\|\mathbf{X}\|_{\max}$ and $g(\mathbf{X})$ must be within $\mathcal{Z}_{[\underline{e},\,\overline{e}]}$ supported by the β_e exponent bits. We define an exponent measure for the floating-point controller realization $\mathbf{X}$ as

$$\gamma(\mathbf{X}) \triangleq \log_2\left(\frac{4\|\mathbf{X}\|_{\max}}{g(\mathbf{X})}\right) . \qquad (11)$$

The following proposition is obvious.

Proposition 1: $\mathbf{X}$ can be represented in the floating-point format of β_e exponent bits without underflow or overflow, if $2^{\beta_e} \ge \log_2\left(\frac{\|\mathbf{X}\|_{\max}}{g(\mathbf{X})}\right) + 2$.

Let β_e^{min} be the smallest exponent bit length that, when used to implement $\mathbf{X}$, can avoid underflow and overflow. It can be computed as

$$\beta_e^{min} = -\lfloor -\log_2(\lfloor \log_2 \|\mathbf{X}\|_{\max}\rfloor - \lfloor \log_2 g(\mathbf{X})\rfloor + 1)\rfloor \,. \qquad (12)$$

The measure $\gamma(\mathbf{X})$ provides an estimate of β_e^{min} as

$$\hat{\beta}_e^{min} \triangleq -\lfloor -\log_2 \gamma(\mathbf{X})\rfloor \,. \qquad (13)$$

It is clear that $\hat{\beta}_e^{min} \ge \beta_e^{min}$.

Secondly, when there is no underflow or overflow, $\mathbf{X}$ is perturbed to $\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}$ due to finite β_w, where $\mathbf{X} \circ \mathbf{\Delta} \triangleq [x_{j,k}\delta_{j,k}]$ is the Hadamard product of $\mathbf{X}$ and $\mathbf{\Delta} = [\delta_{j,k}]$. Each element of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_w+1)}$, that is,

$$\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_w+1)} \,. \qquad (14)$$

With the perturbation $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision floating-point implemented $\mathbf{X}$. It is 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 mantissa measure for the floating-point realization $\mathbf{X}$ is defined as

$$\mu_0(\mathbf{X}) \overset{\triangle}{=} \inf\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}) \text{ is unstable}\}\,. \quad (15)$$

From this definition, the following proposition is obvious.

Proposition 2: $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is stable if $\|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{X})$.

Let β_w^{min} be the mantissa bit length such that $\forall \beta_w \geq \beta_w^{min}$, $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is stable for the floating-point implemented $\mathbf{X}$ with β_w mantissa bits and $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is unstable for the floating-point implemented $\mathbf{X}$ with $\beta_w^{min}-1$ mantissa bits. Except through simulation, β_w^{min} is generally unknown. The mantissa measure $\mu_0(\mathbf{X})$ provides an estimate of β_w^{min} as

$$\hat{\beta}_{w0}^{min} \overset{\triangle}{=} -\lfloor \log_2 \mu_0(\mathbf{X}) \rfloor - 1\,. \quad (16)$$

It can be seen that $\hat{\beta}_{w0}^{min} \geq \beta_w^{min}$.

Define the minimum total bit length required in floating point implementation as

$$\beta^{min} \overset{\triangle}{=} \beta_e^{min} + \beta_w^{min} + 1\,. \quad (17)$$

Clearly, a floating-point implemented $\mathbf{X}$ with a bit length $\beta \geq \beta^{min}$ can guarantee no underflow, no overflow and closed-loop stability. Combining the measures $\gamma(\mathbf{X})$ and $\mu_0(\mathbf{X})$ results in the following true FWL closed-loop stability measure for the floating-point realization $\mathbf{X}$

$$\rho_0(\mathbf{X}) \overset{\triangle}{=} \mu_0(\mathbf{X})/\gamma(\mathbf{X})\,. \quad (18)$$

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

$$\hat{\beta}_0^{min} \overset{\triangle}{=} -\lfloor \log_2 \rho_0(\mathbf{X}) \rfloor + 1\,. \quad (19)$$

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

Proposition 3: A floating-point implemented $\mathbf{X}$ with a bit length β can guarantee no underflow, no overflow and closed-loop stability, if $2^{\beta-1} \geq 1/\rho_0(\mathbf{X})$.

An optimal controller realization can in theory be found by maximizing $\rho_0(\mathbf{X})$, leading to the following optimal controller realization problem

$$\upsilon_{\text{true}} \overset{\triangle}{=} \max_{\mathbf{X}\in\mathcal{S}_C} \rho_0(\mathbf{X})\,. \quad (20)$$

However, the difficulty is that computing the value of $\mu_0(\mathbf{X})$ is an unsolved open problem. In the next section, we will seek an alternative measure that not only can quantify FWL characteristics of $\mathbf{X}$ but also is computationally tractable.

IV. A Tractable FWL Stability Measure

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

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \approx \sum_{j=1}^{l+n}\sum_{k=1}^{q+n} \left.\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right|_{\mathbf{\Delta}=\mathbf{0}} \delta_{j,k}\,. \quad (21)$$

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

$$\left\|\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right\|_{\text{sum}} \overset{\triangle}{=} \sum_{j,k}\left|\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right|\,. \quad (22)$$

Then

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \leq \|\mathbf{\Delta}\|_{\max}\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}\,. \quad (23)$$

This leads to the following mantissa measure for $\mathbf{X}$

$$\mu_1(\mathbf{X}) \overset{\triangle}{=} \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))|}{\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}}\,. \quad (24)$$

Obviously, if $\|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})$, then $|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| < 1$ which means that the closed-loop remains stable under $\mathbf{\Delta}$. In other words, for a given $\mathbf{X}$, the closed-loop can tolerate those FWL perturbations $\mathbf{\Delta}$ whose norms $\|\mathbf{\Delta}\|_{\max}$ are less than $\mu_1(\mathbf{X})$. The larger $\mu_1(\mathbf{X})$ is, the larger FWL errors the closed-loop system can tolerate. Similar to (16), from the mantissa measure $\mu_1(\mathbf{X})$, an estimate of β_w^{min} is given as

$$\hat{\beta}_{w1}^{min} \overset{\triangle}{=} -\lfloor \log_2 \mu_1(\mathbf{X}) \rfloor - 1\,. \quad (25)$$

The assumption of small $\mathbf{\Delta}$ is usually valid in floating-point implementation. Generally speaking, there is no rigorous relationship between $\mu_0(\mathbf{X})$ and $\mu_1(\mathbf{X})$, but $\mu_1(\mathbf{X})$ may be viewed as a lower bound of $\mu_0(\mathbf{X})$, since there are "stable perturbation cubes" larger than $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})\}$ while there is no "stable perturbation cube" larger than $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{X})\}$ [8],[9]. Hence, in most cases, it is reasonable to take that $\mu_1(\mathbf{X}) \leq \mu_0(\mathbf{X})$ and $\hat{\beta}_{w1}^{min} \geq \hat{\beta}_{w0}^{min}$. More importantly, unlike the measure $\mu_0(\mathbf{X})$, the value of $\mu_1(\mathbf{X})$ can be computed explicitly. It is easy to see that

$$\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}} = \frac{\partial|\lambda_i|}{\partial\mathbf{X}}\circ\mathbf{X}\,. \quad (26)$$

Let $\mathbf{p}_i$ be a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X})$ corresponding to the eigenvalue λ_i and $\mathbf{y}_i$ be the related reciprocal left eigenvector. The following lemma is due to [5].

Lemma 1: Let $\overline{\mathbf{A}}(\mathbf{X}) = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2$ given in (8) be diagonalizable. Then

$$\frac{\partial \lambda_i}{\partial \mathbf{X}} = \mathbf{M}_1^T \mathbf{y}_i^* \mathbf{p}_i^T \mathbf{M}_2^T \tag{27}$$

where the superscript $*$ denotes the conjugate operation and T the transpose operator.

The following proposition shows that, given a $\mathbf{X}$, the value of $\mu_1(\mathbf{X})$ can easily be calculated. The proof of this proposition is straightforward.

Proposition 4: Let $\overline{\mathbf{A}}(\mathbf{X})$ be diagonalizable. Then

$$\mu_1(\mathbf{X}) = \min_{i\in\{1,\cdots,m+n\}} \frac{|\lambda_i|(1-|\lambda_i|)}{\left\|\left(\mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T]\mathbf{M}_2^T\right)\circ \mathbf{X}\right\|_{\mathrm{sum}}} . \tag{28}$$

Replacing $\mu_0(\mathbf{X})$ with $\mu_1(\mathbf{X})$ in (18) leads to a computationally tractable FWL closed-loop stability measure

$$\rho_1(\mathbf{X}) \stackrel{\triangle}{=} \mu_1(\mathbf{X})/\gamma(\mathbf{X}) . \tag{29}$$

From the above measure, an estimate of β^{min} is given as

$$\hat{\beta}_1^{min} \stackrel{\triangle}{=} -\lfloor \log_2 \rho_1(\mathbf{X}) \rfloor + 1 . \tag{30}$$

It is useful to compare the proposed measure with the previous results, especially the most recent one given by [11]. For a complex-valued matrix $\mathbf{Y} = [y_{j,k}]$, define the Frobenius norm

$$\|\mathbf{Y}\|_{\mathrm{F}} \stackrel{\triangle}{=} \left(\sum_{j,k} y_{j,k}^* y_{j,k}\right)^{1/2} . \tag{31}$$

Under an assumption that the exponent bits are unlimited, the computationally tractable weighted closed-loop eigenvalue sensitivity index addressed in [11] is given by

$$\Upsilon(\mathbf{X}) \stackrel{\triangle}{=} \sum_{i=1}^{m+n} \alpha_i \Upsilon_i(\mathbf{X}) \tag{32}$$

where α_i are non-negative weighting scalars and $\Upsilon_i(\mathbf{X})$ are single-eigenvalue sensitivities defined by

$$\Upsilon_i(\mathbf{X}) \stackrel{\triangle}{=} \|\mathbf{X}\|_{\mathrm{F}}^2 \left\|\frac{\partial \lambda_i}{\partial \mathbf{X}}\right\|_{\mathrm{F}}^2 . \tag{33}$$

The thinking behind the above definition is as follows. From a first-order approximation, it can easily be shown that

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))-\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max}\|\mathbf{X}\|_{\mathrm{F}} \left\|\frac{\partial \lambda_i}{\partial \mathbf{X}}\right\|_{\mathrm{F}} . \tag{34}$$

Therefore, for those multiplicative perturbations bounded by $\|\mathbf{\Delta}\|_{\max}$, a small $\Upsilon_i(\mathbf{X})$ will limit the resulting change of the corresponding eigenvalue within a small range.

The first observation is that $\rho_1(\mathbf{X})$ considers both the mantissa and exponent of floating-point arithmetic and is therefore able to handle all the aspects of underflow, overflow and closed-loop stability, while $\Upsilon(\mathbf{X})$ only considers the mantissa part and is thus "incomplete". Secondly, $\Upsilon(\mathbf{X})$ deals with the sensitivity of λ_i while $\rho_1(\mathbf{X})$ ($\mu_1(\mathbf{X})$) considers the the sensitivity of $|\lambda_i|$. It is well-known that the stability of a discrete-time linear time-invariant system depends only on the module of its eigenvalues. As $\Upsilon(\mathbf{X})$ includes the unnecessary eigenvalue arguments in consideration, it is generally conservative in comparison with $\rho_1(\mathbf{X})$. Thirdly, $\rho_1(\mathbf{X})$ uses $\left\|\frac{\partial|\lambda_i|}{\partial \mathbf{X}}\circ \mathbf{X}\right\|_{\mathrm{sum}}$ while $\Upsilon(\mathbf{X})$ uses $\|\mathbf{X}\|_{\mathrm{F}} \left\|\frac{\partial\lambda_i}{\partial \mathbf{X}}\right\|_{\mathrm{F}}$ in checking the change of an eigenvalue. It is easy to see that

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))|-|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max} \left\|\frac{\partial|\lambda_i|}{\partial \mathbf{X}}\circ \mathbf{X}\right\|_{\mathrm{sum}}$$

$$\le \|\mathbf{\Delta}\|_{\max}\|\mathbf{X}\|_{\mathrm{F}} \left\|\frac{\partial \lambda_i}{\partial \mathbf{X}}\right\|_{\mathrm{F}} . \tag{35}$$

Obviously, $\left\|\frac{\partial|\lambda_i|}{\partial \mathbf{X}}\circ \mathbf{X}\right\|_{\mathrm{sum}}$ gives a more accurate limit than $\|\mathbf{X}\|_{\mathrm{F}} \left\|\frac{\partial\lambda_i}{\partial \mathbf{X}}\right\|_{\mathrm{F}}$ does on the change of the corresponding eigenvalue module due to the multiplicative perturbations. This again implies that $\rho_1(\mathbf{X})$ is less conservative than $\Upsilon(\mathbf{X})$ in estimating the robustness of closed-loop stability with respect to controller perturbations. The fourth observation is that $\rho_1(\mathbf{X})$ provides an estimate of β^{min}, $\hat{\beta}_1^{min}$ in (30), while $\Upsilon(\mathbf{X})$ cannot provide information on bit length to the designer. One reason is that the measure $\rho_1(\mathbf{X})$ consists of two components, with $\mu_1(\mathbf{X})$ addressing the stability margin and eigenvalue sensitivity linked to the mantissa bits, and $\gamma(\mathbf{X})$ considering the exponent bits, while $\Upsilon(\mathbf{X})$ only focuses on the eigenvalue sensitivity partially linked to the mantissa part. The other reason is that, over all the closed-loop eigenvalues, $\mu_1(\mathbf{X})$ considers the "worst" one while $\Upsilon(\mathbf{X})$ considers a "weighted average".

V. Optimization Procedure

As different realizations $\mathbf{X}$ have different values of the FWL closed-loop stability measure $\rho_1(\mathbf{X})$, it is of practical importance to find an "optimal" realization $\mathbf{X}_{\mathrm{opt}}$ that maximizes $\rho_1(\mathbf{X})$. The controller implemented with this optimal realization $\mathbf{X}_{\mathrm{opt}}$ needs a minimum bit length and has a maximum tolerance to the FWL error. This optimal controller realization problem is formally defined as

$$\upsilon \stackrel{\triangle}{=} \max_{\mathbf{X}\in\mathcal{S}_C} \rho_1(\mathbf{X}) . \tag{36}$$

Assume that a controller has been designed using some standard controller design method. This controller, denoted as

$$\mathbf{X}_0 \triangleq \begin{bmatrix} \mathbf{D}_C^0 & \mathbf{C}_C^0 \\ \mathbf{B}_C^0 & \mathbf{A}_C^0 \end{bmatrix}, \tag{37}$$

is used as the initial realization in the above optimization problem. Let $\mathbf{p}_{0i}$ be a right eigenvector and $\mathbf{y}_{0i}$ the related reciprocal left eigenvector of $\overline{\mathbf{A}}(\mathbf{X}_0)$ corresponding to the eigenvalue λ_i. The definition of $\mathcal{S}_C$ in (6) means that

$$\mathbf{X} \triangleq \mathbf{X}(\mathbf{T}) = \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{38}$$

where $\det(\mathbf{T}) \neq 0$. It can then be shown that

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{X}_0) \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{39}$$

which implies that

$$\mathbf{p}_i = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{p}_{0i}, \quad \mathbf{y}_i = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{y}_{0i}. \tag{40}$$

Hence

$$\mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T] \mathbf{M}_2^T = \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_{0i}^* \mathbf{p}_{0i}^T] \mathbf{M}_2^T$$
$$\times \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{\Phi}_i \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} = \mathbf{\Gamma}(\mathbf{T}) \tag{41}$$

with $\mathbf{\Phi}_i = \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_{0i}^* \mathbf{p}_{0i}^T] \mathbf{M}_2^T$. Define the following cost function:

$$f(\mathbf{T}) \triangleq \min_{i \in \{1,\cdots,m+n\}} \left(\frac{\|\mathbf{\Gamma}(\mathbf{T}) \circ \mathbf{X}(\mathbf{T})\|_{\mathrm{sum}}}{|\lambda_i|(1-|\lambda_i|)} \right.$$
$$\left. \times \log_2 \frac{4\|\mathbf{X}(\mathbf{T})\|_{\max}}{g(\mathbf{X}(\mathbf{T}))} \right)^{-1}. \tag{42}$$

Then the optimal controller realization problem (36) can be posed as the following optimization problem:

$$v = \max_{\substack{\mathbf{T} \in \mathcal{R}^{n \times n} \\ \det \mathbf{T} \neq 0}} f(\mathbf{T}). \tag{43}$$

Efficient numerical optimization methods exist for solving for this optimization problem to provide an optimal transformation matrix $\mathbf{T}_{\mathrm{opt}}$. With $\mathbf{T}_{\mathrm{opt}}$, the optimal realization $\mathbf{X}_{\mathrm{opt}}$ can readily be computed.

VI. A NUMERICAL EXAMPLE

The example taken from [2] was used to illustrate the proposed design procedure for obtaining optimal FWL floating-point controller realizations and to compare it with the method given in [11]. The discrete-time plant was given by

$$\begin{aligned} \mathbf{A}_P &= \left[\begin{matrix} 3.7156e+0 & -5.4143e+0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{matrix} \right. \\ &\qquad \left. \begin{matrix} 3.6525e+0 & -9.6420e-1 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{matrix} \right], \\ \mathbf{B}_P &= [1 \quad 0 \quad 0 \quad 0]^T, \\ \mathbf{C}_P &= [1.1160e-6 \quad 4.3000e-8 \\ &\qquad 1.0880e-6 \quad 1.4000e-8]. \end{aligned}$$

The initial realization of the digital controller was given by

$$\begin{aligned} \mathbf{A}_C^0 &= \left[\begin{matrix} 2.6743e+0 & -5.7446e+0 \\ 2.8769e-1 & -2.7446e-2 \\ -3.3773e-1 & 9.8699e-1 \\ -8.3021e-2 & -3.1988e-3 \end{matrix} \right. \\ &\qquad \left. \begin{matrix} 2.5101e+0 & -9.1782e-1 \\ -6.9444e-1 & -8.9358e-3 \\ -3.2925e-1 & -4.2367e-3 \\ 9.1906e-1 & -1.0415e-3 \end{matrix} \right], \\ \mathbf{B}_C^0 &= [1.0959e+6 \quad 6.3827e+5 \\ &\qquad 3.0262e+5 \quad 7.4392e+4]^T, \\ \mathbf{C}_C^0 &= [1.8180e-1 \quad -2.8313e-1 \\ &\qquad 5.0006e-2 \quad 6.1722e-2], \quad \mathbf{D}_C^0 = 0. \end{aligned}$$

Based on the proposed FWL closed-loop stability measure, the optimization problem (43) was formed and solved for using the MATLAB routine *fminsearch.m* to obtain an optimal transformation matrix

$$\begin{aligned} \mathbf{T}_{\mathrm{opt}} &= \left[\begin{matrix} 7.7275e+3 & -1.0904e+2 \\ 6.9729e+3 & 2.1370e+3 \\ 6.2844e+3 & 3.9092e+3 \\ 5.5879e+3 & 5.2862e+3 \end{matrix} \right. \\ &\qquad \left. \begin{matrix} -2.1292e+2 & 9.8042e+1 \\ 4.4988e+1 & 2.1812e+2 \\ 2.9303e+2 & 2.9240e+2 \\ 5.5027e+2 & 3.4367e+2 \end{matrix} \right]. \end{aligned}$$

An "optimal" controller realization problem was given in [11] based on the weighted closed-loop eigenvalue sensitivity index (32). We will use the index "s", rather then "opt", to denote the solution of this "optimal" realization problem. For this example, the transformation matrix obtained using the MATLAB routine *fminsearch.m* given in [11] is

$$\mathbf{T}_{\mathrm{s}} = \left[\begin{matrix} 8.1477e+3 & 0 \\ 7.0104e+3 & 2.2671e+3 \\ 6.1991e+3 & 3.9989e+3 \\ 5.6761e+3 & 5.2680e+3 \end{matrix} \right.$$

$$\left.\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ 1.1558e+2 & 0 \\ 3.5814e+2 & 1.5299e+1 \end{array}\right] .$$

It is obvious that the true minimum exponent bit length β_e^{min} for a realization $\mathbf{X}$ can directly be obtained by examining the elements of $\mathbf{X}$. The true minimum mantissa bit length β_w^{min} however can only be obtained through simulation. That is, starting from a very large β_w, reduce β_w by one bit and check the closed-loop stability. The process is repeated until there appears closed-loop instability at $\beta_w = \beta_{wu}$. Then $\beta_w^{min} = \beta_{wu} + 1$. Table I summarizes the various measures, the corresponding estimated minimum bit lengths and the true minimum bit lengths for the three controller realizations $\mathbf{X}_0$, $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{opt}}$, respectively. It can be seen that the floating-point implementation of $\mathbf{X}_0$ needs at least 26 bits (20 mantissa bits and 5 exponent bits) while the implementation of $\mathbf{X}_{\mathrm{opt}}$ needs at least 13 bits (8 mantissa bits and 4 exponent bits). The reduction in the bit length required is 13 (12-bit reduction for the mantissa part and 1-bit reduction for the exponent part). Comparing $\mathbf{X}_{\mathrm{opt}}$ with $\mathbf{X}_{\mathrm{s}}$, it can be seen that $\mathbf{X}_{\mathrm{opt}}$ needs one bit less in the exponent part and one bit less in the mantissa part to maintain the closed-loop stability.

VII. Conclusions

The closed-loop stability issue of finite-precision realizations has been investigated for digital controller implemented in floating-point arithmetic. A new computationally tractable FWL closed-loop stability measure has been derived for floating-point controller realizations. Unlike the existing methods, which only consider the mantissa part of floating-point scheme, the proposed measure takes into account both the exponent and mantissa parts of floating-point format. It has been shown that this new measure yields a more accurate estimate for the FWL robustness of closed-loop stability. Based on this FWL closed-loop stability measure, the optimal controller realization problem has been formulated, which can easily be solved for using standard numerical optimization algorithms. A numerical example has demonstrated that the proposed design procedure yields computationally efficient controller realizations suitable for FWL float-point implementation in real-time applications.

Acknowledgements

J. Wu and S. Chen wish to thank the support of the U.K. 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.

TABLE I

Various measures, corresponding estimated minimum bit lengths and true minimum bit lengths for three controller realizations $\mathbf{X}_0$, $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{opt}}$.

Realization	$\mathbf{X}_0$	$\mathbf{X}_{\mathrm{s}}$	$\mathbf{X}_{\mathrm{opt}}$
ρ_1	2.6644e-9	4.7588e-6	9.5931e-6
$\hat{\beta}_1^{min}$	30	19	18
μ_1	8.5182e-8	8.7907e-5	1.5229e-4
$\hat{\beta}_{w1}^{min}$	23	13	12
γ	3.1971e+1	1.8473e+1	1.5875e+1
$\hat{\beta}_e^{min}$	5	5	4
β^{min}	26	15	13
β_w^{min}	20	9	8
β_e^{min}	5	5	4

References

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] S. Chen, R.S.H. Istepanian, J. Wu and J. Chu, “Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators,” *Systems and Control Letters*, Vol.40, No.3, pp.153–163, 2000.

[8] J. Wu, S. Chen, G. Li, R.S.H. Istepanian and J. Chu, “Shift and delta operator realizations for digital controllers with finite-word-length considerations,” *IEE Proc. Control Theory and Applications*, Vol.147, No.6, pp.664–672, 2000.

[9] 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.

[10] 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.

[11] J.F. Whidborne and D. Gu, “Optimal finite-precision controller and filter realizations using floating-point arithmetic,” *Research Report EM/2001/07*, Department of Mechanical Engineering, King’s College London, London, U.K., 2001.