Optimal Finite-Precision Controller Realizations in Block-Floating-Point Format

Jun Wu †, Sheng Chen ‡ [0], and Jian Chu †

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

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

Abstract

The paper analyzes the properties of the controller coefficient perturbation resulting from using finite word length (FWL) block-floating-point arithmetic and investigates the closed-loop stability issue of finite-precision realizations for digital controllers implemented in block-floating-point format. A true FWL closed-loop stability measure is derived which considers both the dynamic range and precision of number representation in block-floating-point format. To facilitate the design of optimal finite-precision controller realizations, a computationally tractable FWL block-floating-point closed-loop stability measure is introduced and the method of computing the value of this measure for a given controller realization is developed. The optimal controller realization is defined as the solution that maximizes the proposed measure, and a numerical optimization approach is adopted to solve for the resulting optimal realization problem. A numerical example is used to illustrate the proposed design procedure.

Keywords – digital controller, finite word length, block-floating-point, closed-loop stability, optimization.

1 Introduction

The classical digital controller design methodology often assumes that the controller is implemented exactly, even though in reality a control law can only be realized in finite precision. It may seem that the uncertainty resulting from finite-precision computing of the digital controller is so small, compared to the uncertainty within the plant, such that this controller "uncertainty" can simply be ignored. Increasingly, however, researchers have 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

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

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 sole on dealing with large plant uncertainty [1].

In practice, the controller parameters are represented by a digital processor of finite bit length in one of the three number representation formats, namely, fixed-point, floating-point or block-floating-point format. In a given representation format, different controller realizations have different degrees of "robustness" to FWL errors. This property can be utilized to select "optimal" realizations in a given format. The optimal controller realization problems in fixed-point and floating-point formats have been studied [2]–[10]. The block-floating-point scheme has potential advantages of combining the simplicity of fixed-point format and the accuracy of floating-point format. The previous work [11] has compared the closed-loop stability performance of various block-floating-point and fixed-point implemented realizations for a PID benchmark system. However the optimal controller realization problem in block-floating-point format was not discussed, and to date the true block-floating-point FWL closed-loop stability measure has not been seen which can then be optimized to obtain optimal block-floating-point realizations. This paper focuses on deriving the optimal controller realization problem in block-floating-point format.

2 Block-Floating-Point Representation

The fixed-point and floating-point formats are the two basic representation schemes for real numbers stored in memory and registers. 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. For an illustrative purpose, consider 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 η_1 be the element in $\mathcal{S}_1$ that has the largest absolute value, and η_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

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

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

$$h \triangleq \begin{cases} \lfloor \log_2 |\eta_1| \rfloor + 1, & \text{for } \ x \in \mathcal{S}_1 \\ \lfloor \log_2 |\eta_2| \rfloor + 1, & \text{for } \ x \in \mathcal{S}_2 \end{cases} \tag{2}$$

with the *floor* function $\lfloor x \rfloor$ denoting the closest integer less than or equal to x. When all the elements in $\mathcal{S}$ are stored in a block-floating-point digital processor of the bit length

$$\beta = 1 + \beta_u + \beta_h, \tag{3}$$

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

$$\mathcal{F} \triangleq \left\{ \left(\sum_{j=1}^{\beta_u} b_j 2^{-j} - s \right) \times 2^h : \; s \in \{0,1\}, \; b_j \in \{0,1\}, \; h \in \mathcal{Z}, \; \underline{h} \leq h \leq \overline{h} \right\} \tag{4}$$

where $\mathcal{Z}$ denotes the set of integers, $\underline{h}$ and $\overline{h}$ represent the lower and upper limits of the block exponent, respectively, and $\overline{h} - \underline{h} = 2^{\beta_h} - 1$.

Define the integer set $\mathcal{Z}_{[\underline{h},\, \overline{h}]} \triangleq \{h : \; h \in \mathcal{Z}, \; \underline{h} \leq h \leq \overline{h}\}$. When no underflow or overflow occurs, that is, $h \in \mathcal{Z}_{[\underline{h},\, \overline{h}]}$, the block-floating-point quantization operator $\mathcal{Q} : \mathcal{S} \to \mathcal{F}$ is defined as

$$\mathcal{Q}(x) \triangleq (-1)^s 2^{(h-\beta_u)} \lfloor 2^{(\beta_u - h)} |x| + 0.5 \rfloor. \tag{5}$$

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

$$\varepsilon \triangleq |x - \mathcal{Q}(x)|. \tag{6}$$

Denote

$$r(x) \triangleq \begin{cases} 2^{\lfloor \log_2 |\eta_1| \rfloor + 1}, & \text{for } \; x \in \mathcal{S}_1, \\ 2^{\lfloor \log_2 |\eta_2| \rfloor + 1}, & \text{for } \; x \in \mathcal{S}_2. \end{cases} \tag{7}$$

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

$$\varepsilon < r(x) 2^{-(\beta_u + 1)}. \tag{8}$$

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

$$\mathcal{Q}(x) = x + r(x)\delta, \;\; |\delta| < 2^{-(\beta_u + 1)}. \tag{9}$$

Note that the perturbation resulting from FWL block-floating-point representation is neither multiplicative nor additive. It can also be seen that the dynamic range of block-floating-point representation is determined by β_h bits while the precision is determined by β_u bits.

3 Problem Statement

Consider the discrete-time closed-loop control system, consisting of a linear time-invariant plant P and a digital controller C. The plant model P is assumed to be strictly proper with a state-space description

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases} \tag{10}$$

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 $\mathbf{C} \in \mathcal{R}^{q\times n}$. The digital controller C is described by

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) \\ \mathbf{e}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases} \tag{11}$$

with $\mathbf{F} \in \mathcal{R}^{m\times m}$, $\mathbf{G} \in \mathcal{R}^{m\times q}$, $\mathbf{J} \in \mathcal{R}^{p\times m}$ and $\mathbf{M} \in \mathcal{R}^{p\times q}$. It is well-known that the realizations of C are not unique. Assume that a realization $(\mathbf{F}_0, \mathbf{G}_0, \mathbf{J}_0, \mathbf{M}_0)$ of C has been designed. Then all the realizations of C form the realization set

$$\mathcal{S}_C \stackrel{\triangle}{=} \{(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0, \mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0\} \tag{12}$$

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

$$\mathbf{X} = [x_{j,k}] \stackrel{\triangle}{=} \begin{bmatrix} \mathbf{M} & \mathbf{J} \\ \mathbf{G} & \mathbf{F} \end{bmatrix}. \tag{13}$$

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

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{A} + \mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC} & \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \stackrel{\triangle}{=} \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2 \tag{14}$$

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

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| < 1, \ \ \forall i \in \{1, \cdots, m+n\}. \tag{15}$$

However, the controller realization $\mathbf{X}$ is implemented in block-floating-point format of β_h block exponent bits, β_u block mantissa bits and one sign bit. For a matrix $\mathbf{W} = [w_{j,k}]$, define

$$\|\mathbf{W}\|_{\max} \stackrel{\triangle}{=} \max_{j,k} |w_{j,k}|, \tag{16}$$

$$\pi(\mathbf{W}) \stackrel{\triangle}{=} \min_{j,k}\{|w_{j,k}| : w_{j,k} \neq 0\} \tag{17}$$

and let $\mathbf{U}(\mathbf{W})$ be the matrix of the same dimension as $\mathbf{W}$, whose elements are all 1s. For the two matrices $\mathbf{W} = [w_{j,k}]$ and $\mathbf{Z} = [z_{j,k}]$ of the same dimension, define the Hadamard product of $\mathbf{W}$ and $\mathbf{Z}$

$$\mathbf{W} \circ \mathbf{Z} \stackrel{\triangle}{=} [w_{j,k}z_{j,k}]. \tag{18}$$

Assumed that $\mathbf{X}$ is divided into "natural" blocks of $\mathbf{F}$, $\mathbf{G}$, $\mathbf{J}$ and $\mathbf{M}$. Let ξ_1 be the element in $\mathbf{F}$ which has the largest absolute value. Similarly, ξ_2, ξ_3 and ξ_4 are defined in $\mathbf{G}$, $\mathbf{J}$ and $\mathbf{M}$, respectively. Denote

$$\mathbf{q}(\mathbf{X}) \stackrel{\triangle}{=} [\,\xi_1 \quad \xi_2 \quad \xi_3 \quad \xi_4\,]^T \tag{19}$$

with ^{T} being the transpose operator.

Firstly, the dynamic range of β_h bits must be large enough for $\mathbf{X}$. We define a dynamic range measure for realization $\mathbf{X}$ in block-floating-point format as

$$\gamma(\mathbf{X}) \stackrel{\triangle}{=} \log_2 \frac{4\|\mathbf{q}(\mathbf{X})\|_{\max}}{\pi(\mathbf{q}(\mathbf{X}))}. \tag{20}$$

The rationale of this dynamic range measure becomes clear in the following obvious proposition.

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

Let β_h^{min} be the smallest block exponent bit length that, when used to implement $\mathbf{X}$, does not cause overflow or underflow. The minimum required block exponent bit length can easily be computed by

$$\beta_h^{min}(\mathbf{X}) = \lceil \log_2(\lfloor \log_2 \|\mathbf{q}(\mathbf{X})\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{q}(\mathbf{X})) \rfloor + 1) \rceil \tag{21}$$

where the *ceil* function $\lceil x \rceil$ denotes the closest integer greater than or equal to x. The measure $\gamma(\mathbf{X})$ defined in (20) provides an estimate of β_h^{min} as

$$\hat{\beta}_h^{min}(\mathbf{X}) \stackrel{\triangle}{=} \lceil \log_2 \gamma(\mathbf{X}) \rceil. \tag{22}$$

It can easily be seen that $\hat{\beta}_h^{min} \geq \beta_h^{min}$.

Even when the dynamic range is sufficient, that is, $\beta_h \geq \beta_h^{min}$, $\mathbf{X}$ is perturbed to $\mathbf{X} + \mathbf{E}(\mathbf{X}) \circ \mathbf{\Delta}$ due to the effect of finite β_u where

$$\mathbf{E}(\mathbf{X}) \stackrel{\triangle}{=} \begin{bmatrix} 2^{\lfloor \log_2 |\xi_4| \rfloor + 1}\mathbf{U}(\mathbf{M}) & 2^{\lfloor \log_2 |\xi_3| \rfloor + 1}\mathbf{U}(\mathbf{J}) \\ 2^{\lfloor \log_2 |\xi_2| \rfloor + 1}\mathbf{U}(\mathbf{G}) & 2^{\lfloor \log_2 |\xi_1| \rfloor + 1}\mathbf{U}(\mathbf{F}) \end{bmatrix}. \tag{23}$$

Each element $\delta_{j,k}$ of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_u+1)}$, that is,

$$\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_u+1)}. \tag{24}$$

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

$$\mu_0(\mathbf{X}) \stackrel{\triangle}{=} \inf\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{X} + \mathbf{E}(\mathbf{X}) \circ \mathbf{\Delta}) \text{ is unstable}\}. \tag{25}$$

From the above definition, the following proposition is obvious.

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

Thus under the condition of a sufficient block exponent bit length, that is, $\beta_h \geq \beta_h^{min}$, the perturbation $\|\mathbf{\Delta}\|_{\max}$ and therefore the block mantissa bit length β_u determines whether the closed-loop remains stable. Let β_u^{min} be the block mantissa bit length such that $\forall \beta_u \geq \beta_u^{min}$, the closed-loop system is stable with $\mathbf{X}$ implemented by β_u block mantissa bits and the closed-loop system is unstable with $\mathbf{X}$ implemented by $\beta_u^{min}-1$ block mantissa bits. Except in simulation, this minimum block mantissa bit length β_u^{min} is generally unknown. However, the precision measure $\mu_0(\mathbf{X})$ provides an estimate of β_u^{min} as

$$\hat{\beta}_{u0}^{min}(\mathbf{X}) \triangleq -\lfloor \log_2 \mu_0(\mathbf{X}) \rfloor - 1\,. \tag{26}$$

It can easily be seen that $\hat{\beta}_{u0}^{min} \geq \beta_u^{min}$.

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

$$\beta^{min} \triangleq \beta_h^{min} + \beta_u^{min} + 1\,. \tag{27}$$

Clearly, $\mathbf{X}$ implemented with a bit length $\beta \geq \beta^{min}$ can guarantee a sufficient dynamic range 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 given realization $\mathbf{X}$ in block-floating-point format

$$\rho_0(\mathbf{X}) \triangleq \mu_0(\mathbf{X})/\gamma(\mathbf{X})\,. \tag{28}$$

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

$$\hat{\beta}_0^{min}(\mathbf{X}) \triangleq -\lfloor \log_2 \rho_0(\mathbf{X}) \rfloor + 1\,. \tag{29}$$

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 $\mathbf{X}$ in block-floating-point format.

Proposition 3 The controller realization $\mathbf{X}$ implemented in block-floating-point format with a bit length β can guarantee a sufficient dynamic range and closed-loop stability, if

$$2^{\beta-1} \geq \frac{1}{\rho_0(\mathbf{X})}\,. \tag{30}$$

Since $\rho_0(\mathbf{X})$ depends on the controller realization $\mathbf{X}$ only, an optimal realization can in theory be found by maximizing $\rho_0(\mathbf{X})$ over $\mathcal{S}_C$, leading to the following optimal controller realization problem

$$v_{\mathrm{true}} \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} \rho_0(\mathbf{X})\,. \tag{31}$$

However, how to compute the value of $\mu_0(\mathbf{X})$ is an unsolved open problem. Thus, the true FWL closed-loop stability measure $\rho_0(\mathbf{X})$ and the optimal realization problem (31) have limited practical significance. In the next section, an alternative measure is derived which not only can quantify FWL characteristics of $\mathbf{X}$ in block-floating-point format 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, $\forall i \in \{1, \cdots, m+n\}$

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{E}(\mathbf{X})\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \approx \sum_{j,k} \left.\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right|_{\mathbf{\Delta}=\mathbf{0}} \delta_{j,k} \,. \tag{32}$$

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

$$\left\|\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right\|_{\text{sum}} \triangleq \sum_{j,k}\left|\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right| \,. \tag{33}$$

Then

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{E}(\mathbf{X})\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max}\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}} \,. \tag{34}$$

This leads to the following precision measure for realization $\mathbf{X}$ in block-floating-point format

$$\mu_1(\mathbf{X}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))|}{\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}} \,. \tag{35}$$

Obviously, if $\|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})$, then $|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{E}(\mathbf{X})\circ\mathbf{\Delta}))| < 1$ which means that the closed-loop remains stable under the FWL error $\mathbf{\Delta}$. In other words, for a given $\mathbf{X}$ implemented in block-floating-point format with a sufficient dynamic range, the closed-loop can tolerate those FWL perturbations $\mathbf{\Delta}$ whose norms $\|\mathbf{\Delta}\|_{\max}$ are less than $\mu_1(\mathbf{X})$. The larger $\mu_1(\mathbf{X})$ is, the larger FWL errors the closed-loop system can tolerate. Similar to (26), from the precision measure $\mu_1(\mathbf{X})$, an estimate of β_u^{min} is given as

$$\hat{\beta}_{u1}^{min}(\mathbf{X}) \triangleq -\lfloor\log_2\mu_1(\mathbf{X})\rfloor - 1 \,. \tag{36}$$

The assumption of small $\mathbf{\Delta}$ is usually valid in practical implementation of digital controllers. Generally speaking, there is no rigorous relationship between $\mu_0(\mathbf{X})$ and $\mu_1(\mathbf{X})$, but $\mu_1(\mathbf{X})$ is connected with a lower bound of $\mu_0(\mathbf{X})$ in some manners: 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 $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{X})\}$ [7]. Hence, in most cases, it is reasonable to take that $\mu_1(\mathbf{X}) \le \mu_0(\mathbf{X})$ and $\hat{\beta}_{u1}^{min} \ge \hat{\beta}_{u0}^{min}$. Unlike the measure $\mu_0(\mathbf{X})$, the value of $\mu_1(\mathbf{X})$ can be computed explicitly. It is easy to see that

$$\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}} = \mathbf{E}(\mathbf{X})\circ\frac{\partial|\lambda_i|}{\partial\mathbf{X}} \,. \tag{37}$$

Let $\mathbf{p}_i$ be a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X})$ corresponding to the eigenvalue λ_i. Define

$$\mathbf{M}_{\mathbf{p}} \triangleq [\,\mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_{m+n}\,] \,, \tag{38}$$

$$\mathbf{M}_{\mathbf{y}} \triangleq [\,\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_{m+n}\,] = \mathbf{M}_{\mathbf{p}}^{-H} \,, \tag{39}$$

where ^{H} denotes the conjugate transpose operator and $\mathbf{y}_i$ is called the reciprocal left eigenvector related to $\mathbf{p}_i$. 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 (14) be diagonalizable. Then

$$\frac{\partial \lambda_i}{\partial \mathbf{X}} = \mathbf{M}_1^T \mathbf{y}_i^* \mathbf{p}_i^T \mathbf{M}_2^T \tag{40}$$

where * denotes the conjugate operation.

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

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

$$\mu_1(\mathbf{X}) = \min_{i\in\{1,\cdots,m+n\}} \frac{|\lambda_i|(1-|\lambda_i|)}{\|(\mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T]\mathbf{M}_2^T) \circ \mathbf{E}(\mathbf{X})\|_{\text{sum}}} \,. \tag{41}$$

Proof: Noting $|\lambda_i| = \sqrt{\lambda_i^* \lambda_i}$ leads to

$$\frac{\partial |\lambda_i|}{\partial \mathbf{X}} = \frac{1}{2\sqrt{\lambda_i^*\lambda_i}} \left(\frac{\partial \lambda_i^*}{\partial \mathbf{X}} \lambda_i + \lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right) = \frac{1}{2|\lambda_i|} \left(\left(\frac{\partial \lambda_i}{\partial \mathbf{X}} \right)^* \lambda_i + \lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right) = \frac{1}{|\lambda_i|} \mathrm{Re} \left[\lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right] \,. \tag{42}$$

Combining (35), (37), (42) and Lemma 1 results in this proposition.

Replacing $\mu_0(\mathbf{X})$ with $\mu_1(\mathbf{X})$ in (28) leads to a computationally tractable FWL closed-loop stability measure

$$\rho_1(\mathbf{X}) \triangleq \mu_1(\mathbf{X})/\gamma(\mathbf{X}) \,. \tag{43}$$

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

$$\hat{\beta}_1^{min}(\mathbf{X}) \triangleq -\lfloor \log_2 \rho_1(\mathbf{X}) \rfloor + 1 \,. \tag{44}$$

5 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 $\mathbf{X}_{\text{opt}}$ that maximizes $\rho_1(\mathbf{X})$. The controller implemented with this optimal realization $\mathbf{X}_{\text{opt}}$ needs a minimum bit length and has a maximum tolerance to the FWL error. This optimal controller realization problem is formally defined as

$$v \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} \rho_1(\mathbf{X}) \,. \tag{45}$$

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

$$\mathbf{X}_0 \triangleq \begin{bmatrix} \mathbf{M}_0 & \mathbf{J}_0 \\ \mathbf{G}_0 & \mathbf{F}_0 \end{bmatrix} \,, \tag{46}$$

is used as the initial controller realization in the above optimal controller realization problem. Let $\mathbf{p}_{0i}$ be a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X}_0)$ corresponding to the eigenvalue λ_i, and $\mathbf{y}_{0i}$ be the reciprocal left eigenvector related to $\mathbf{p}_{0i}$. The definition of $\mathcal{S}_C$ in (12) means that

$$\mathbf{X} \triangleq \mathbf{X}(\mathbf{T}) = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{47}$$

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

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{X}_0) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{48}$$

which implies that

$$\mathbf{p}_i = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{p}_{0i}, \quad \mathbf{y}_i = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{y}_{0i} \,. \tag{49}$$

Hence

$$\mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T] \mathbf{M}_2^T = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_{0i}^* \mathbf{p}_{0i}^T] \mathbf{M}_2^T \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{\Phi}_i \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \tag{50}$$

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:

$$f(\mathbf{T}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \left(\frac{\left\| \left(\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{\Phi}_i \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \right) \circ \mathbf{E}(\mathbf{X}(\mathbf{T})) \right\|_{\mathrm{sum}}}{|\lambda_i|(1-|\lambda_i|)} \log_2 \frac{4\|\mathbf{q}(\mathbf{X}(\mathbf{T}))\|_{\max}}{\pi(\mathbf{q}(\mathbf{X}(\mathbf{T})))} \right)^{-1} \,. \tag{51}$$

Then the optimal controller realization problem (45) can be posed as the following optimization problem:

$$v = \max_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det \mathbf{T}\neq 0}} f(\mathbf{T}) \,. \tag{52}$$

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.

6 A Design Example

An example is used to illustrate the design procedure based on the proposed FWL block-floating-point closed-loop stability measure. The discrete-time plant, taken from [2], 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 was given by

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

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

$$\mathbf{J}_0 = [\,-2.5548e+2 \quad -2.7185e+2 \quad -2.7188e+2 \quad 2.7188e+2\,], \quad \mathbf{M}_0 = [\,0\,].$$

Based on the proposed FWL closed-loop stability measure, the optimization problem (52) was formed. Using the MATLAB routine *fminsearch.m*, this optimization problem was solved for to obtain the optimal similarity transformation

$$\mathbf{T}_{\rm opt} = \begin{bmatrix} -1.0345e-1 & 1.2904e-1 & 3.8329e-3 & 1.0911e-2 \\ -1.1078e-1 & 1.1742e-1 & 2.9461e-3 & 8.1639e-3 \\ -2.3775e-2 & 2.3815e-2 & 4.9498e-4 & 1.8293e-3 \\ 9.2138e-2 & -1.1474e-1 & -3.4007e-3 & -9.6780e-3 \end{bmatrix}.$$

It is obvious that the true minimum block exponent bit length $\beta_h^{min}(\mathbf{X})$ for a realization $\mathbf{X}$ can directly be obtained by examining the elements of $\mathbf{X}$. The true minimum block mantissa bit length $\beta_u^{min}(\mathbf{X})$ however can only be obtained through simulation. That is, starting from a very large β_u, reduce β_u by one bit and check the closed-loop stability. The process is repeated until there appears closed-loop instability at $\beta_u = \beta_{uu}$. Then $\beta_u^{min} = \beta_{uu} + 1$. Table 1 summarizes the various measures, the corresponding estimated minimum bit lengths and the true minimum bit lengths for the controller realizations $\mathbf{X}_0$ and $\mathbf{X}_{\rm opt}$. It can be seen that $\mathbf{X}_{\rm opt}$ improves the FWL closed-loop stability measure ρ_1 by a factor of 3×10^5. To guarantee closed-loop stability, the block-floating-point implemented $\mathbf{X}_0$ needs at least 33 bits while the implementation of $\mathbf{X}_{\rm opt}$ needs at least 16 bits. The latter gives a saving of 17 bits.

7 Conclusions

The closed-loop stability issue of finite-precision realizations has been investigated for digital controller implemented in block-floating-point arithmetic. A new computationally tractable FWL closed-loop stability measure has been derived for block-floating-point controller realizations. The proposed measure takes into account both the block exponent and block mantissa parts of block-floating-point format. 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 block-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 supports of Zhejiang Provincial Natural Science Foundation of China (Grant 699085) and Doctor Degree Programs Foundation of China (Grant 1999033571).

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] 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.

[8] 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.

[9] 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.

[10] J. Wu, S. Chen, J.F. Whidborne and J. Chu, “Optimal realizations of floating-point implemented digital controllers with finite word length considerations,” submitted to *Automatica*, 2001.

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

	$\mathbf{X}_0$	$\mathbf{X}_{\mathrm{opt}}$
$\rho_1(\mathbf{X})$	$1.5154e-11$	$4.7787e-6$
$\hat{\beta}_1^{min}(\mathbf{X})$	37	19
$\mu_1(\mathbf{X})$	$6.8793e-11$	$3.6388e-5$
$\hat{\beta}_{u1}^{min}(\mathbf{X})$	33	14
$\gamma(\mathbf{X})$	$4.5395e+0$	$7.6146e+0$
$\hat{\beta}_h^{min}(\mathbf{X})$	3	3
$\beta^{min}(\mathbf{X})$	33	16
$\beta_u^{min}(\mathbf{X})$	30	12
$\beta_h^{min}(\mathbf{X})$	2	3

Table 1: Various measures, estimated minimum bit lengths and true minimum bit lengths for $\mathbf{X}_0$ and $\mathbf{X}_{\mathrm{opt}}$.