http://www.zju.edu.cn/jzus; E-mail: jzus@zju.edu.cn
ISSN 1009-3095 *Journal of Zhejiang University SCIENCE* V.4, No.6, P.651-657, Nov.-Dec., 2003

Optimization of block-floating-point realizations for digital controllers with finite-word-length considerations*

WU Jun(吴 俊)†1, HU Xie-he(胡协和)1,
CHEN Sheng(陈 生)2, CHU Jian(褚 健)1

(1 *National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, China*)
(2 *Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*)

†E-mail: jwu@iipc.zju.edu.cn

Received Nov.21, 2002; revision accepted Mar.3, 2003

Abstract: The closed-loop stability issue of finite-precision realizations was investigated for digital controllers implemented in block-floating-point format. The controller coefficient perturbation was analyzed resulting from using finite word length (FWL) block-floating-point representation scheme. A block-floating-point FWL closed-loop stability measure was derived which considers both the dynamic range and precision. To facilitate the design of optimal finite-precision controller realizations, a computationally tractable block-floating-point FWL closed-loop stability measure was then introduced and the method of computing the value of this measure for a given controller realization was developed. The optimal controller realization is defined as the solution that maximizes the corresponding measure, and a numerical optimization approach was adopted to solve the resulting optimal realization problem. A numerical example was used to illustrate the design procedure and to compare the optimal controller realization with the initial realization.

Key words: Digital controller, Finite word length, Block-floating-point, Closed-loop stability, Optimization
Document code: A **CLC number:** O224

INTRODUCTION

Due to the finite word length (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 (Keel and Bhattacharryya, 1997).

A control law can be accomplished with different realizations and the parameters of a controller realization are represented by a digital processor of finite bit length in a particular number representation format, such as fixed-point, floating-point or block-floating-point format. In a given representation format, different controller realizations have different degrees of "robustness" against FWL errors. This property can be utilized to select "optimal" realizations in the given format. In fixed-point format or floating-point format, the optimal controller realization problems were studied in Gevers and Li (1993), Istepanian and Whidborne (2001), Fialho and Georgiou (1994; 1999), Li (1998), Whidborne *et al*. (2000; 2001), Wu *et al*. (2001) and Whidborne and Gu (2001). A comparative study was presented by Istepanian *et al*. (2000) on the stability and performance using block-floating-point and fixed-point implementations for various realizations for a PID digital controller of a steel rolling mill benchmark system. However the optimal controller realization problem in block-floating-point format was not discussed there. To date the true block-floating-point FWL

* Project supported by the National Natural Science Foundation of China (No.60174026) and the Scientific Research Foundation for Returned Overseas Chinese Scholars of Zhejiang Province (No.J20020546)

closed-loop stability measure has not been seen which can be optimized to obtain the optimal block-floating-point realization. This work is aimed to study the optimal controller realization problem in block-floating-point format.

BLOCK-FLOATING-POINT

The fixed-point and floating-point formats are the two basic representation schemes for real numbers stored in digital memory and in digital 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 S. In the block-floating-point format, 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 the fixed-point scheme and the accuracy of the floating-point scheme.

For illustrative purpose and without loss of generality, consider the case of dividing S into two non-empty subsets S_1 and S_2, which satisfy $S_1 \bigcup S_2 = S$ and $S_1 \bigcap S_2$ is the empty set. Let η_1 be the element in S_1 that has the largest absolute value, and η_2 be the element in S_2 that has the largest absolute value. Then, any $x \in S$ 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

$$h \triangleq \begin{cases} \log_2 |\eta_1| \rfloor + 1, & \text{for } x \in S_1, \\ \log_2 |\eta_2| \rfloor + 1, & \text{for } x \in S_2, \end{cases} \tag{2}$$

$\cdot\rfloor$ denotes the floor function, i.e., $x\rfloor$ is the closest integer less than or equal to x. Obviously, all the elements in the same block have the same exponent value of h. When all the elements in S are stored in a digital processor of the bit length

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

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

$$\mathscr{F} \triangleq \left\{ \left(\sum_{j=1}^{\beta_u} b_j 2^{-j} - s \right) \times 2^h : s \in \{0,1\}, \right.$$
$$\left. b_j \in \{0,1\},\ h \in Z, \underline{h} \leqslant h \leqslant \bar{h} \right\} \tag{4}$$

where Z denotes the set of integers, $\underline{h}$ and $\bar{h}$ represent the lower and upper limits of the block exponent, respectively, and $\bar{h} - \underline{h} = 2^{\beta_h} - 1$. Obviously, when $h > \bar{h}$ or $h < \underline{h}$, overflow or underflow will occur in the block-floating-point representation.

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

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

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

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

Denote

$$r(x) \triangleq \begin{cases} 2^{\log_2 |\eta_1| \rfloor + 1}, & \text{for } x \in S_1, \\ 2^{\log_2 |\eta_2| \rfloor + 1}, & \text{for } x \in S_2. \end{cases} \tag{7}$$

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

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

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

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

Hence the perturbation resulting from FWL block-floating-point representation is neither multiplicative nor additive. The perturbation depends on the set S and how S is divided into blocks. It can also be seen that the dynamic range of block-floating-point representation is determined by β_h bits, and the precision of block-floating-point representation is determined by β_u bits.

PROBLEM STATEMENT

Consider the discrete-time closed-loop con-

trol 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

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

which is completely state controllable and observable with $\boldsymbol{A} \in R^{n\times n}$, $\boldsymbol{B} \in R^{n\times p}$ and $\boldsymbol{C} \in R^{q\times n}$. The digital controller C is described by

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

with $\boldsymbol{F} \in R^{m\times m}$, $\boldsymbol{G} \in R^{m\times q}$, $\boldsymbol{J} \in R^{p\times m}$ and $\boldsymbol{M} \in R^{p\times q}$.

Assume that a realization $(\boldsymbol{F}_0, \boldsymbol{G}_0, \boldsymbol{J}_0, \boldsymbol{M}_0)$ 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

$$S_C \triangleq \{(\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{J}, \boldsymbol{M}) : \boldsymbol{F} = \boldsymbol{T}^{-1}\boldsymbol{F}_0\boldsymbol{T}, \boldsymbol{G} = \boldsymbol{T}^{-1}\boldsymbol{G}_0, \boldsymbol{J} = \boldsymbol{J}_0\boldsymbol{T}, \boldsymbol{M} = \boldsymbol{M}_0\} \tag{12}$$

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

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

We also refer to $\boldsymbol{X}$ as a realization of C. The stability of the closed-loop system depends on the eigenvalues of the matrix

$$\begin{aligned} \bar{\boldsymbol{A}}(\boldsymbol{X}) &= \begin{bmatrix} \boldsymbol{A} + \boldsymbol{B}\boldsymbol{M}\boldsymbol{C} & \boldsymbol{B}\boldsymbol{J} \\ \boldsymbol{G}\boldsymbol{C} & \boldsymbol{F} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \boldsymbol{X} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \\ &\triangleq \boldsymbol{M}_0 + \boldsymbol{M}_1\boldsymbol{X}\boldsymbol{M}_2 \end{aligned} \tag{14}$$

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

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

For a matrix $\boldsymbol{W} = [w_{j,k}]$, let $\boldsymbol{U}(\boldsymbol{W})$ be the matrix of the same dimension whose elements are all 1s,

$$\| \boldsymbol{W} \|_{\max} \triangleq \max_{j,k} |w_{j,k}|, \tag{16}$$

$$\pi(\boldsymbol{W}) \triangleq \min_{j,k} \{|w_{j,k}| : w_{j,k} \neq 0\}, \tag{17}$$

For two matrice $\boldsymbol{W} = [w_{j,k}]$ and $\boldsymbol{Z} = [z_{j,k}]$ of the same dimension, define the Hadamard product of $\boldsymbol{W}$ and $\boldsymbol{Z}$

$$\boldsymbol{W} \circ \boldsymbol{Z} \triangleq [w_{j,k} z_{j,k}]. \tag{18}$$

We have known that the controller realization $\boldsymbol{X}$ is implemented in block-floating-point format of β_h block exponent bits, β_u block mantissa bits and one sign bit. In the remainder of this paper, it is assumed that $\boldsymbol{X}$ stored in the block-floating-point format is divided into "natural" blocks of $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{J}$ and $\boldsymbol{M}$. Let ξ_1 be the element in $\boldsymbol{F}$ which has the largest absolute value. Similarly ξ_2, ξ_3 and ξ_4 is defined in $\boldsymbol{G}$, $\boldsymbol{J}$ and $\boldsymbol{M}$ respectively. Denote

$$\boldsymbol{q}(\boldsymbol{X}) \triangleq [\xi_1 \quad \xi_2 \quad \xi_3 \quad \xi_4]^{\mathrm{T}} \tag{19}$$

with $^{\mathrm{T}}$ being the transpose operator.

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

$$\gamma(\boldsymbol{X}) \triangleq \log_2 \frac{4 \| \boldsymbol{q}(\boldsymbol{X}) \|_{\max}}{\pi(\boldsymbol{q}(\boldsymbol{X}))}. \tag{20}$$

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

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

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

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

where $\lceil \cdot \rceil$ denotes the ceil function, i.e., $\lceil x \rceil$ is the closest integer greater than or equal to x. Note that the measure $\gamma(\boldsymbol{X})$ defined in Eq. (20) provides an estimate of $\beta_h^{\min}$ as

$$\hat{\beta}_h^{\min}(\boldsymbol{X}) \triangleq \lceil \log_2 \gamma(\boldsymbol{X}) \rceil. \tag{22}$$

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

When the dynamic range is sufficient, according to Eq. (9), $\boldsymbol{X}$ is perturbed to $\boldsymbol{X} + \boldsymbol{E}(\boldsymbol{X}) \circ \boldsymbol{\Delta}$ due to the effect of finite β_u where

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

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

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

With the perturbation $\boldsymbol{\Delta}$, $\lambda_i(\bar{\boldsymbol{A}}(\boldsymbol{X}))$ is moved to $\lambda_i(\boldsymbol{A}(\boldsymbol{X} + \boldsymbol{E}(\boldsymbol{X}) \circ \boldsymbol{\Delta}))$. If an eigenvalue of $\boldsymbol{A}(\boldsymbol{X} + \boldsymbol{E}(\boldsymbol{X}) \circ \boldsymbol{\Delta})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision implemented $\boldsymbol{X}$.

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 $\boldsymbol{X}$ in block-floating-point format can be defined as

$$\mu_0(\boldsymbol{X}) \triangleq \inf\{ \| \boldsymbol{\Delta} \|_{\max} : \bar{\boldsymbol{A}}(\boldsymbol{X} + \boldsymbol{E}(\boldsymbol{X}) \circ \boldsymbol{\Delta}) \text{ is unstable} \}. \tag{25}$$

From the above definition, the following proposition is obvious.

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

Thus under the condition that the dynamic range is sufficient, that is, $\beta_h \geqslant \beta_h^{\min}$, the perturbation $\| \boldsymbol{\Delta} \|_{\max}$ and therefore the block mantissa bit length β_u determines whether the closed-loop remains stable. Let $\beta_u^{\min}$ be the block mantissa bit length such that $\forall\, \beta_u \geqslant \beta_u^{\min}$, the closed-loop system is stable with $\boldsymbol{X}$ implemented by β_u block mantissa bits and the closed-loop system is unstable with $\boldsymbol{X}$ implemented by $\beta_u^{\min} - 1$ block mantissa bits. The precision measure $\mu_0(\boldsymbol{X})$ provides an estimate of $\beta_u^{\min}$ as

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

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

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

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

Clearly, $\boldsymbol{X}$ implemented with a bit length $\beta \geqslant \beta^{\min}$ can guarantee a sufficient dynamic range and closed-loop stability. Combining the measures $\gamma(\boldsymbol{X})$ and $\mu_0(\boldsymbol{X})$ results in the following true FWL closed-loop stability measure for the given realization $\boldsymbol{X}$ inblock-floating-point format

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

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

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

It is clear that $\hat{\beta}_0^{\min} \geqslant \beta^{\min}$. The following proposition summarizes the usefulness of $\rho_0(\boldsymbol{X})$ as a measure for the FWL characteristics of $\boldsymbol{X}$ in block-floating-point format.

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

$$2^{\beta - 1} \geqslant \frac{1}{\rho_0(\boldsymbol{X})}. \tag{30}$$

The closed-loop stability measure $\rho_0(\boldsymbol{X})$ depends on the controller realization $\boldsymbol{X}$ only. Consequently, an optimal realization can in theory be found by maximizing $\rho_0(\boldsymbol{X})$ over S_C, leading to the following optimal controller realization problem

$$\upsilon_{\text{true}} \triangleq \max_{\boldsymbol{X} \in S_C} \rho_0(\boldsymbol{X}). \tag{31}$$

However, the difficulty with this approach is that computing the value of $\mu_0(\boldsymbol{X})$ is an unsolved open problem. Thus, the true FWL closed-loop stability measure $\rho_0(\boldsymbol{X})$ and the optimal realization problem (31) have limited practical significance. In the next section, an alternative measure is developed which not only can quantify the FWL characteristics of $\boldsymbol{X}$ in block-floating-point format but also is computationally tractable.

A TRACTABLE FWL CLOSED-LOOP STABILITY MEASURE

When the FWL error $\boldsymbol{\Delta}$ is small, from a

first-order approximation, $\forall\, i \in \{1, \cdots, m+n\}$

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

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

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

Then

$$|\lambda_i(\bar{\boldsymbol{A}}(\boldsymbol{X}+\boldsymbol{E}(\boldsymbol{X})\circ\boldsymbol{\Delta}))| - |\lambda_i(\bar{\boldsymbol{A}}(\boldsymbol{X}))| \leqslant \|\boldsymbol{\Delta}\|_{\max} \left\| \frac{\partial |\lambda_i|}{\partial \boldsymbol{\Delta}}\Big|_{\boldsymbol{\Delta}=\boldsymbol{0}} \right\|_{\text{sum}}. \tag{34}$$

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

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

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

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

The assumption of small $\boldsymbol{\Delta}$ is usually valid in practical implementation of digital controllers. Generally speaking, there is no rigorous relationship between $\mu_0(\boldsymbol{X})$ and $\mu_1(\boldsymbol{X})$, but $\mu_1(\boldsymbol{X})$ is connected with a lower bound of $\mu_0(\boldsymbol{X})$ in some manner: there are "stable perturbation cubes" larger than $\{\boldsymbol{\Delta}: \|\boldsymbol{\Delta}\|_{\max} < \mu_1(\boldsymbol{X})\}$ while there is no "stable perturbation cube" larger than $\{\boldsymbol{\Delta}: \|\boldsymbol{\Delta}\|_{\max} < \mu_0(\boldsymbol{X})\}$ (Wu *et al.*, 2001). Hence, in most cases, it is reasonable to take that $\mu_1(\boldsymbol{X}) \leqslant \mu_0(\boldsymbol{X})$ and $\hat{\beta}_{u1}^{\min} \geqslant \hat{\beta}_{u0}^{\min}$. More importantly, unlike the measure $\mu_0(\boldsymbol{X})$, the value of $\mu_1(\boldsymbol{X})$ can be computed explicitly. It is easy to see that

$$\frac{\partial |\lambda_i|}{\partial \boldsymbol{\Delta}}\Big|_{\boldsymbol{\Delta}=\boldsymbol{0}} = \boldsymbol{E}(\boldsymbol{X}) \circ \frac{\partial |\lambda_i|}{\partial \boldsymbol{X}} \tag{37}$$

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

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

and

$$\boldsymbol{M}_y \triangleq [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_{m+n}] = \boldsymbol{M}_p^{-H} \tag{39}$$

where the superscript ^{H} denotes the conjugate transpose operator and $\boldsymbol{y}_i$ is called the reciprocal left eigenvector related to $\boldsymbol{p}_i$. The following lemma is due to Li (1998).

Lemma 1 Let $\bar{\boldsymbol{A}}(\boldsymbol{X}) = \boldsymbol{M}_0 + \boldsymbol{M}_1 \boldsymbol{X} \boldsymbol{M}_2$ given in Eq. (14) be diagonalizable. Then

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

where the superscript * denotes the conjugate operation.

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

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

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

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

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

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

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

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

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

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

OPTIMIZATION PROCEDURE

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

$$\upsilon \triangleq \max_{X \in S_C} \rho_1(\boldsymbol{X}). \tag{45}$$

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

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

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

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

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

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

which implies that

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

Hence

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

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

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

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

$$\upsilon = \max_{\substack{\boldsymbol{T} \in R^{m \times m} \\ \det \boldsymbol{T} \neq 0}} f(\boldsymbol{T}). \tag{52}$$

As the optimization problem Eq.(52) is highly nonlinear, global optimization algorithms, such as the genetic algorithm and adaptive simulated annealing, can be adopted to provide a (sub)optimal similarity transformation $\boldsymbol{T}_{\text{opt}}(\alpha)$. Global optimization methods are however computationally demanding. Local optimization algorithms, such as Rosenbrock and Simplex algorithms, are computationally simpler but run more risks of only attaining a local solution. Our experience with the optimization problem Eq.(29) suggests that local optimization methods are usually efficient in controllers of low order while global optimization methods have to be adopted in controllers of high order. It also helps to choose a "good" initial controller realization, such as Li's closed-loop sub-optimal realization (Li, 1998), as the initial guess for the optimization routine. With the solution $\boldsymbol{T}_{\text{opt}}$ of optimization problem Eq.(52), the optimal realization $\boldsymbol{X}_{\text{opt}}$ can readily be computed.

DESIGN EXAMPLE

An example is used to illustrate the design procedure based on the FWL closed-loop stability measure. In this example, the discrete-time plant is given by

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

$$\boldsymbol{B} = [1 \quad 0 \quad 0 \quad 0]^{\mathrm{T}},$$

$$\boldsymbol{C} = [1.1160\text{e}-6 \quad 4.3000\text{e}-8 \quad 1.0880\text{e}-6 \quad 1.4000\text{e}-8].$$

The initial realization of the digital controller is given by

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

$$\boldsymbol{G}_0 = [-4.6765e+1 \quad -4.5625e+1 \quad -9.5195e+0 \quad 4.1609e+1]^{\mathrm{T}},$$

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

$$\boldsymbol{M}_0 = [0].$$

Based on the proposed FWL closed-loop stability measure, the optimization problem Eq. (52) is formed. Using the MATLAB routine *fminsearch.m* for the controller of 4 order, this optimization problem is solved to obtain the optimal similarity transformation

$$\boldsymbol{T}_{\mathrm{opt}} = \begin{bmatrix} -1.0345e-001 & 1.2904e-001 & 3.8329e-003 & 1.0911e-002 \\ -1.1078e-001 & 1.1742e-001 & 2.9461e-003 & 8.1639e-003 \\ -2.3775e-002 & 2.3815e-002 & 4.9498e-004 & 1.8293e-003 \\ 9.2138e-002 & -1.1474e-001 & -3.4007e-003 & -9.6780e-003 \end{bmatrix}.$$

It is obvious that the true minimum block exponent bit length $\beta_h^{\min}(\boldsymbol{X})$ for a realization $\boldsymbol{X}$ can directly be obtained by examining the elements of $\boldsymbol{X}$. The true minimum block mantissa bit length $\beta_u^{\min}(\boldsymbol{X})$ 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 $\boldsymbol{X}_0$ and $\boldsymbol{X}_{\mathrm{opt}}$. It can be seen that $\boldsymbol{X}_{\mathrm{opt}}$ improves the measure ρ_1 by a factor of 300000 over $\boldsymbol{X}_0$ and that the block-floating-point implemented $\boldsymbol{X}_0$ needs at least 33 bits while the implementation of $\boldsymbol{X}_{\mathrm{opt}}$ needs at least 16 bits. More than half of the bit length is saved.

Table 1 Various measures and bit lengths for $\boldsymbol{X}$ and $\boldsymbol{X}_{\mathrm{opt}}$

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

References

Fialho, I.J. and Georgiou, T.T., 1994. On stability and performance of sampled-data systems subject to word-length constraint. *IEEE Trans. Automatic Control*, **39** (12): 2476 – 2481.

Fialho, I.J. and Georgiou, T.T., 1999. Optimal Finite Wordlength Digital Controller Realization. Proc. American Control Conf., San Diego, p.4326 – 4327.

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

Istepanian, R.S.H., Whidborne, J.F. and Bauer, P., 2000. Stability Analysis of Block Floating Point Digital Controllers. Proc. UKACC Int. Conf. Control, Cambridge, CD-ROM, 6 pages.

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

Keel, L.H. and Bhattacharryya, S.P., 1997. Robust, fragile, or optimal? *IEEE Trans. Automatic Control*, **42**(8): 1098 – 1105.

Li, G., 1998. On the structure of digital controllers with finite word length consideration. *IEEE Trans. Automatic Control*, **43**(5): 689 – 693.

Whidborne, J.F., Wu, J. and Istepanian, R.S.H., 2000. Finite word length stability issues in an l_1 framework. *Int. J. Control*, **73**(2): 166 – 176.

Whidborne, J.F., Istepanian, R.S.H. and Wu, J., 2001. Reduction of controller fragility by pole sensitivity minimization. *IEEE Trans. Automatic Control*, **46** (2): 320 – 325.

Whidborne, J.F. and Gu, D., 2001. Optimal finite-precision controller and filter realizations using floating-point arithmetic. *King's College London Mechanical Engineering Department Report EM*/2001/07, London.

Wu, J., Chen, S., Li, G., Istepanian, R.S.H. and Chu, J., 2001. An improved closed-loop stability related measure for finite-precision digital controller realizations. *IEEE Trans. Automatic Control*, **46**(7): 1162 – 1166.