Computing a FWL Stability Measure for Second Order Digital Systems

J. Wu

National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou 310027, P. R. China
Email: jwu@iipc.zju.edu.cn

S. Chen

Department of Electronics and Computer Science
University of Southampton, Highfield, Southampton SO17 1BJ, U.K.
Email: sqc@ecs.soton.ac.uk

J. Chu

National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou 310027, P. R. China
Email: chuj@iipc.zju.edu.cn

Abstract

The best measure quantifying FWL (finite word length) stability is the one that bases on the largest stable perturbation hypercube. But the computing of this FWL stability measure has not been solved. For second order digital systems, this paper develops an analytic computing method. Through solving 12 linear equations and 12 quadratic equations, the measure value can be obtained exactly.

1 Introduction

The recent advances in digital system design methods have led to a need for the efficient and accurate implementation of filters or controllers. Although the number of filter/controller implementations using floating-point processors is increasing due to their reduced price, for reasons of cost, simplicity, speed, memory space, and ease-of-programming, the use of fixed-point processors is more desirable for many industrial and consumer applications. The "robustness" of digital system stability under filter/controller parameter perturbations is critical issue in fixed-point implementations. A designed, stable digital system may become unstable when the "infinite-precision" filter/controller is implemented using a fixed-point processor due to finite-word-length (FWL) effects [1][2].

It is well known that a filter/control law can be accomplished with different realizations, and that these different realizations possess different degrees of stability robustness to FWL errors. The FWL stability measure υ addressed by Fialho and Georgiou [3] is the best measure quantifying the FWL stability character of a realization. Unfortunately, for a given realization nobody know how to calculate the value of υ. Since the computing of υ is very difficult, various tractable FWL stability measures are addressed to replace υ in some senses [4]–[9]. For second order digital systems, this paper develops an analytic method of computing υ explicitly. The remainder of this paper is organized in the following way. Section 2 introduces FWL stability and the measure υ. Section 3 presents and analyzes an algorithm of computing υ for second order digital systems. Numerical examples are given in section 4 to demonstrate the effectiveness of the proposed method, and the paper concludes at section 5.

2 FWL Stability and Its Measure υ

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

$$\begin{cases} \mathbf{x}_P(t+1) = \mathbf{A}_P\mathbf{x}_P(t) + \mathbf{B}_P\mathbf{u}(t) \\ \mathbf{z}(t) = \mathbf{C}_P\mathbf{x}_P(t) \end{cases} \quad (1)$$

where $\mathbf{A}_P \in \mathcal{R}^{m \times m}$, $\mathbf{B}_P \in \mathcal{R}^{m \times l}$ and $\mathbf{C}_P \in \mathcal{R}^{q \times m}$. The digital controller $C(z)$ is described by

$$\begin{cases} \mathbf{x}_C(t+1) = \mathbf{A}_C \mathbf{x}_C(t) + \mathbf{B}_C \mathbf{z}(t) \\ \mathbf{u}(t) = \mathbf{C}_C \mathbf{x}_C(t) + \mathbf{D}_C \mathbf{z}(t) \end{cases} \quad (2)$$

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 $\mathbf{D}_C \in \mathcal{R}^{l \times q}$. Denote the *realization* of $C(z)$ as

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix}. \quad (3)$$

Suppose that an initial realization of $C(z)$

$$\mathbf{X}_0 \triangleq \begin{bmatrix} \mathbf{D}_C^0 & \mathbf{C}_C^0 \\ \mathbf{B}_C^0 & \mathbf{A}_C^0 \end{bmatrix} \quad (4)$$

has been given by some controller synthesis method, and all the realizations of $C(z)$ form a set

$$\mathcal{S}_C \triangleq \left\{ \mathbf{X} : \mathbf{X} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \right\} \quad (5)$$

where the transformation $\mathbf{T} \in \mathcal{R}^{n \times n}$ is an arbitrary non-singular matrix, $\mathbf{0}$ and $\mathbf{I}$ denote the zero matrix and the identity matrix of appropriate dimension respectively. The stability of the closed-loop control system depends on the eigenvalues of the closed-loop transition matrix

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{X}) &= \begin{bmatrix} \mathbf{A}_P + \mathbf{B}_P \mathbf{D}_C \mathbf{C}_P & \mathbf{B}_P \mathbf{C}_C \\ \mathbf{B}_C \mathbf{C}_P & \mathbf{A}_C \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &\triangleq \mathbf{M}_0 + \mathbf{M}_1 \mathbf{X} \mathbf{M}_2. \end{aligned} \quad (6)$$

Secondly, a discrete-time filter system can be viewed as the trivial case of closed-loop system (1)(2) with $P(z) = \mathbf{0}$, $m = 0$ and $C(z)$ represents the filter. Accordingly, the stability of the filter system still depends on $\overline{\mathbf{A}}(\mathbf{X})$ with $\mathbf{M}_0 = \mathbf{0}$, $\mathbf{M}_1 = \mathbf{I}$, $\mathbf{M}_2 = \mathbf{I}$ and $\mathbf{X} = \mathbf{A}_C$, *i.e.* $\overline{\mathbf{A}}(\mathbf{X}) = \mathbf{A}_C$.

All the different realizations $\mathbf{X}$ in $\mathcal{S}_C$ have exactly the same set of poles if they are implemented with infinite precision. Since the digital system has been designed to be stable, all the eigenvalues $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le k \le m + n$, are within the unit disk. When $\mathbf{X}$ is implemented in fixed-point format of finite word length, it is perturbed to $\mathbf{X} + \mathbf{\Delta}$. Each element of $\mathbf{\Delta}$ is bounded by $\pm \varepsilon$, that is,

$$\|\mathbf{\Delta}\|_{\mathrm{m}} \le \varepsilon \quad (7)$$

where $\|\mathbf{\Delta}\|_{\mathrm{m}}$ denotes the maximal absolute value of all elements in $\mathbf{\Delta}$, and ε is the maximum representation error of the digital processor. With the perturbation $\mathbf{\Delta}$, $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))$ is moved to $\lambda_k(\overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta}))$. If an eigenvalue of

$$\overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta}) = \overline{\mathbf{A}}(\mathbf{X}) + \mathbf{M}_1 \mathbf{\Delta} \mathbf{M}_2 \quad (8)$$

is outside the open unit disk, the digital system, designed to be stable, becomes unstable with the finite-precision implemented $\mathbf{X}$.

It is therefore critical, for a realization $\mathbf{X} \in \mathcal{S}_C$, to know how many degree of FWL error will cause closed-loop instability. This means that we would like to know the largest open "hypercube" in the perturbation space within which the digital system remains stable. The size of this perturbation hypercube quantifies the FWL stability characteristics of $\mathbf{X}$ and therefore Fialho and Georgiou [3] addressed the FWL stability measure

$$\upsilon(\mathbf{X}) \triangleq \inf\{\|\mathbf{\Delta}\|_{\mathrm{m}} : \overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta}) \text{ is unstable}\}. \quad (9)$$

From the definition of $\upsilon(\mathbf{X})$, it is easy to see

Theorem 1 $\overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta})$ **is stable if** $\|\mathbf{\Delta}\|_{\mathrm{m}} < \upsilon(\mathbf{X})$.

The above theorem implies that the larger $\upsilon(\mathbf{X})$ is, the larger FWL errors the realization $\mathbf{X}$ can tolerate.

3 Computing $\upsilon(\mathbf{X})$ for 2×2 $\overline{\mathbf{A}}(\mathbf{X})$

Although $\upsilon(\mathbf{X})$ is a pretty measure characterizing FWL stability, its computing is still an open problem. In other words, for a given $\mathbf{X}$, one does not know

$$\upsilon(\mathbf{X}) = \sup\{|\eta| : \|\mathbf{\Delta}\|_{\mathrm{m}} \le |\eta|, \overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta}) \text{ is stable}\}. \quad (10)$$

This paper will discuss computing $\upsilon(\mathbf{X})$ for 2×2 $\overline{\mathbf{A}}(\mathbf{X})$. It is supposed $\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} a_0 & c_0 \\ d_0 & b_0 \end{bmatrix} \in \mathcal{R}^{2 \times 2}$. Noticing that $\mathbf{\Delta}$ has the same dimension as $\mathbf{X}$ and that $\mathbf{X} = \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix}$ lying in $\mathcal{R}^{2 \times 1}$, $\mathcal{R}^{1 \times 2}$ or $\mathcal{R}$ hardly makes sense in filter/controller analysis and synthesis, we assume that $\mathbf{\Delta} = \begin{bmatrix} \delta_1 & \delta_3 \\ \delta_4 & \delta_2 \end{bmatrix} \in \mathcal{R}^{2 \times 2}$, $\begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} b_P & 0 \\ 0 & 1 \end{bmatrix} \in \mathcal{R}^{2 \times 2}$ and $\begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} c_P & 0 \\ 0 & 1 \end{bmatrix} \in \mathcal{R}^{2 \times 2}$ with $b_P \ne 0$ and $c_P \ne 0$. Briefly, in the remainder of this paper,

$$\begin{aligned} & \overline{\mathbf{A}}(\mathbf{X} + \mathbf{\Delta}) \\ &= \begin{bmatrix} a_0 & c_0 \\ d_0 & b_0 \end{bmatrix} + \begin{bmatrix} b_P & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 & \delta_3 \\ \delta_4 & \delta_2 \end{bmatrix} \begin{bmatrix} c_P & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} a_0 + b_P c_P \delta_1 & c_0 + b_P \delta_3 \\ d_0 + c_P \delta_4 & b_0 + \delta_2 \end{bmatrix} \in \mathcal{R}^{2 \times 2}. \quad (11) \end{aligned}$$

Let

$$\begin{aligned} a &\triangleq a_0 + b_P c_P \delta_1, & (12) \\ b &\triangleq b_0 + \delta_2, & (13) \\ c &\triangleq c_0 + b_P \delta_3, & (14) \\ d &\triangleq d_0 + c_P \delta_4. & (15) \end{aligned}$$

It can be seen that, for $\eta \in \mathcal{R}$, the set

$$\mathcal{P}_\eta \triangleq \left\{ \begin{bmatrix} a & b & c & d \end{bmatrix}^T : \begin{array}{l} -|\eta| \le \delta_1 \le |\eta| \\ -|\eta| \le \delta_2 \le |\eta| \\ -|\eta| \le \delta_3 \le |\eta| \\ -|\eta| \le \delta_4 \le |\eta| \end{array} \right\} \quad (16)$$

forms a hyper-cuboid in $\mathcal{R}^4$. Clearly, each edge of $\mathcal{P}_\eta$ is parallel to the a, b, c or d axis, and $\mathcal{P}_\eta$ has 16 vertices expressed as $\begin{bmatrix} a_0 \pm b_P c_P \eta, b_0 \pm \eta, c_0 \pm b_P \eta, d_0 \pm c_P \eta \end{bmatrix}^T$. It is well known that $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is stable if and only if all the roots of the polynomial $\lambda^2 - (a+b)\lambda + ab - cd$ lie in the open unit disk. For a second order polynomial, Jury [10] gave a necessary and sufficient condition to check whether all its roots lie in the open unit disk as

Lemma 1 all the roots of $\alpha_2 \lambda^2 + \alpha_1 \lambda + \alpha_0$ ($\alpha_2 \neq 0$) lie in the open unit disk if and only if $|\alpha_0| < |\alpha_2|$ and $|\alpha_0 + \alpha_2| > |\alpha_1|$.

Then the following lemma is from Lemma 1 directly.

Lemma 2 $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is stable if and only if these inequalities hold

$$ab - cd - 1 < 0, \quad (17)$$
$$(a+1)(b+1) - cd > 0, \quad (18)$$
$$(a-1)(b-1) - cd > 0. \quad (19)$$

Based on Lemma 2, we obtain an useful result as

Theorem 2 $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is stable for any $\begin{bmatrix} a & b & c & d \end{bmatrix}^T \in \mathcal{P}_\eta$ if and only if that $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is stable at all the 16 vertices of $\mathcal{P}_\eta$.

Proof: Necessity is obvious.

Sufficiency can be shown in the following manner. In the ab plane grained with curves $ab = \omega$ ($-\infty < \omega < \infty$), observe a closed rectangle block K with 2 edges parallel to the a axis and the other 2 edges parallel to the b axis. It is seen easily that wherever K is placed in the ab plane, $\max_{[a,b]^T \in K} ab$ is achieved at one vertex of K while $\min_{[a,b]^T \in K} ab$ is achieved at another vertex of K. Fig. 1 displays 3 rectangle blocks at different places in the ab plane. For K_1, $\max_{[a,b]^T \in K_1} ab$ is achieved at the upper-right vertex while $\min_{[a,b]^T \in K_1} ab$ is achieved at the lower-left vertex. For K_2, $\max_{[a,b]^T \in K_2} ab$ is achieved at the lower-left vertex while $\min_{[a,b]^T \in K_2} ab$ is achieved at the upper-left vertex. For K_3, $\max_{[a,b]^T \in K_3} ab$ is achieved at the lower-left vertex while $\min_{[a,b]^T \in K_3} ab$ is achieved at the lower-right vertex. This observation make it understood, for any $\begin{bmatrix} a & b & c & d \end{bmatrix}^T \in \mathcal{P}_\eta$, that the value ab is bounded from both upper and lower sides respectively by two of the values $(a_0 + b_P c_P \eta)(b_0 + \eta)$, $(a_0 + b_P c_P \eta)(b_0 - \eta)$, $(a_0 - b_P c_P \eta)(b_0 + \eta)$ and $(a_0 - b_P c_P \eta)(b_0 - \eta)$, as well as that the value cd is bounded from both upper and lower sides respectively by two of the values $(c_0 + b_P \eta)(d_0 + c_P \eta)$, $(c_0 + b_P \eta)(d_0 - c_P \eta)$, $(c_0 - b_P \eta)(d_0 + c_P \eta)$ and $(c_0 - b_P \eta)(d_0 - c_P \eta)$. Without loss of the generality, here we suppose that

$$ab \le (a_0 + b_P c_P \eta)(b_0 + \eta), \quad (20)$$
$$(c_0 - b_P \eta)(d_0 - c_P \eta) \le cd. \quad (21)$$

Then

$$ab - cd$$
$$\le (a_0 + b_P c_P \eta)(b_0 + \eta) - (c_0 - b_P \eta)(d_0 - c_P \eta). \, (22)$$

Noticing that

$$\begin{bmatrix} a_0 + b_P c_P \eta, b_0 + \eta, c_0 - b_P \eta, d_0 - c_P \eta \end{bmatrix}^T$$

is a vertex of $\mathcal{P}_\eta$ at which

$$(a_0 + b_P c_P \eta)(b_0 + \eta) - (c_0 - b_P \eta)(d_0 - c_P \eta) - 1 < 0 \quad (23)$$

holds, we arrive at $ab - cd - 1 < 0$. Similarly, one can prove $(a+1)(b+1) - cd > 0$ and $(a-1)(b-1) - cd > 0$. Hence, $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is stable for any $\begin{bmatrix} a & b & c & d \end{bmatrix}^T \in \mathcal{P}_\eta$.

Let

$$f_1(\begin{bmatrix} a & b & c & d \end{bmatrix}^T) = ab - cd - 1, \quad (24)$$
$$f_2(\begin{bmatrix} a & b & c & d \end{bmatrix}^T) = (a+1)(b+1) - cd, (25)$$
$$f_3(\begin{bmatrix} a & b & c & d \end{bmatrix}^T) = (a-1)(b-1) - cd. (26)$$

Theorem 2 offers a way of finding the largest open hyper-cuboid in which $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is stable. That is, beginning by $\eta = 0$, to enlarge $\mathcal{P}_\eta$ until a vertex of $\mathcal{P}_\eta$ touches one of the curved hyper-surfaces $f_i(\begin{bmatrix} a & b & c & d \end{bmatrix}^T) = 0$, $i \in \{1, 2, 3\}$. With the change of η, the 16 vertices of $\mathcal{P}_\eta$ move along 8 curves in $\mathcal{R}^4$ (each pair of vertices symmetrical about $\begin{bmatrix} a & b & c & d \end{bmatrix}^T$ are contained in the same curve). The 8 curves include

$$\mathbf{p}_1(\eta) = \begin{bmatrix} a_0 + b_P c_P \eta \\ b_0 + \eta \\ c_0 + b_P \eta \\ d_0 + c_P \eta \end{bmatrix}, \mathbf{p}_2(\eta) = \begin{bmatrix} a_0 - b_P c_P \eta \\ b_0 + \eta \\ c_0 + b_P \eta \\ d_0 + c_P \eta \end{bmatrix},$$

$$\mathbf{p}_3(\eta) = \begin{bmatrix} a_0 + b_P c_P \eta \\ b_0 - \eta \\ c_0 + b_P \eta \\ d_0 + c_P \eta \end{bmatrix}, \mathbf{p}_4(\eta) = \begin{bmatrix} a_0 + b_P c_P \eta \\ b_0 + \eta \\ c_0 - b_P \eta \\ d_0 + c_P \eta \end{bmatrix},$$

$$\mathbf{p}_5(\eta) = \begin{bmatrix} a_0 + b_P c_P \eta \\ b_0 + \eta \\ c_0 + b_P \eta \\ d_0 - c_P \eta \end{bmatrix}, \mathbf{p}_6(\eta) = \begin{bmatrix} a_0 - b_P c_P \eta \\ b_0 - \eta \\ c_0 + b_P \eta \\ d_0 + c_P \eta \end{bmatrix},$$

$$\mathbf{p}_7(\eta) = \begin{bmatrix} a_0 - b_P c_P \eta \\ b_0 + \eta \\ c_0 - b_P \eta \\ d_0 + c_P \eta \end{bmatrix}, \mathbf{p}_8(\eta) = \begin{bmatrix} a_0 - b_P c_P \eta \\ b_0 + \eta \\ c_0 + b_P \eta \\ d_0 - c_P \eta \end{bmatrix},$$

$\eta \in \mathcal{R}$.

For computing $\upsilon(\mathbf{X})$, it is needed to know the values of η at which $\mathbf{p}_j(\eta)$, $j \in \{1, \cdots, 8\}$ intersects $f_i([\begin{matrix} a & b & c & d \end{matrix}]^T) = 0$, $i \in \{1,2,3\}$. This actually requires to solve the equation $f_i(\mathbf{p}_j(\eta)) = 0$. Appendix lists totally 24 equations $f_i(\mathbf{p}_j(\eta)) = 0$, $i \in \{1,2,3\}$, $j \in \{1,\cdots,8\}$ consisting of 12 linear equations and 12 quadratic equations. By solving these 24 equations, we obtain the set

$$\mathcal{Q} \triangleq \{\eta : \exists i \in \{1,2,3\}, j \in \{1,\cdots,8\}, f_i(\mathbf{p}_j(\eta)) = 0\}. \tag{27}$$

Comment: For some values of a, b, c, d, b_P and c_P, it possibly happens that any linear $f_i(\mathbf{p}_j(\eta)) = 0$ has no solution or any quadratic $f_i(\mathbf{p}_j(\eta)) = 0$ has no real-value solution. As an example, when $a_0 = b_0 = c_0 = d_0 = 0$ and $b_P = c_P = 1$, $f_1(\mathbf{p}_1(\eta)) = 0$ expressed as $-1 = 0$ has no solution and $f_1(\mathbf{p}_4(\eta)) = 0$ expressed as $2\eta^2 + 1 = 0$ has no real-value solution. This means neither $\mathbf{p}_1(\eta)$ or $\mathbf{p}_4(\eta)$ intersects $f_1([\begin{matrix} a & b & c & d \end{matrix}]^T) = 0$. Of course, $\mathcal{Q}$ is not empty, otherwise $\upsilon(\mathbf{X})$ is infinitely large that never be true.

By $\mathcal{Q}$, one can compute $\upsilon(\mathbf{X})$ according to

Theorem 3 $\upsilon(\mathbf{X}) = \min_{\eta \in \mathcal{Q}} |\eta|$.

Proof: Without the loss of generality, suppose that $\eta_1 = \min_{\eta \in \mathcal{Q}} |\eta|$ and $f_1(\mathbf{p}_1(\eta_1)) = 0$. Firstly,

$$\begin{bmatrix} a_0 + b_P c_P \eta_1 & c_0 + b_P \eta_1 \\ d_0 + c_P \eta_1 & b_0 + \eta_1 \end{bmatrix}$$
$$= \begin{bmatrix} a_0 & c_0 \\ d_0 & b_0 \end{bmatrix} + \begin{bmatrix} b_P & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_1 & \eta_1 \\ \eta_1 & \eta_1 \end{bmatrix} \begin{bmatrix} c_P & 0 \\ 0 & 1 \end{bmatrix} \tag{28}$$

is unstable and hence $\upsilon(\mathbf{X}) \leq \eta_1$. Secondly, if $\upsilon(\mathbf{X}) < \eta_1$, there exists η_2, δ_1, δ_2, δ_3 and δ_4 such that $|\eta_2| < \eta_1$ and $\begin{bmatrix} a_0 + b_P c_P \delta_1 & c_0 + b_P \delta_3 \\ d_0 + c_P \delta_4 & b_0 + \delta_2 \end{bmatrix} \in \mathcal{P}_{\eta_2}$ is unstable. From Theorem 2, this is equivalent to that $\begin{bmatrix} a & c \\ d & b \end{bmatrix}$ is unstable at one vertex of $\mathcal{P}_{\eta_2}$. In other words, one of the 3 inequalities (17)–(19) is not satisfied at one vertex $\mathbf{p}_j(\eta_2)$ of $\mathcal{P}_{\eta_2}$. We assume $f_2(\mathbf{p}_2(\eta_2)) \geq 0$ without the loss of generality. Noting that $f_2(\mathbf{p}_2(\eta))$ is a continuous function and that $\begin{bmatrix} a_0 & c_0 \\ d_0 & b_0 \end{bmatrix}$ is stable, *i.e.* $f_2(\mathbf{p}_2(0)) < 0$, there is η_3 such that $|\eta_3| \leq |\eta_2|$ and $f_2(\mathbf{p}_2(\eta_3)) = 0$ from the Intermediate Value Theorem [11]. Further $f_2(\mathbf{p}_2(\eta_3)) = 0$ leads to $\eta_3 \in \mathcal{Q}$ and $|\eta_3| \leq |\eta_2|$ leads to $|\eta_3| < \eta_1$. The results $\eta_3 \in \mathcal{Q}$ and $|\eta_3| < \eta_1$ clearly contradicts the fact $\eta_1 = \min_{\eta \in \mathcal{Q}} |\eta|$. Therefore, $\upsilon(\mathbf{X}) = \eta_1$.

Now, the analytic algorithm of computing $\upsilon(\mathbf{X})$ for 2×2 $\overline{\mathbf{A}}(\mathbf{X})$ is summarized as

Step 1 Construct 24 equations $f_i(\mathbf{p}_j(\eta)) = 0$, $i \in \{1,2,3\}$, $j \in \{1,\cdots,8\}$ according to Appendix.

Step 2 Solve the 24 equations, obtain the set $\mathcal{Q}$.

Step 3 $\upsilon(\mathbf{X}) = \min_{\eta \in \mathcal{Q}} |\eta|$.

4   Numerical Example

An initial realization $\mathbf{X}_{ini}$ of a second order digital filter is given by

$$\begin{aligned} \mathbf{A}_{Cini} &= \begin{bmatrix} -1.0061e-2 & 9.8327e-1 \\ -9.9386e-1 & 1.6731e0 \end{bmatrix}, \\ \mathbf{B}_{Cini} &= \begin{bmatrix} -1.1380e-2 \\ 9.9980e-1 \end{bmatrix}, \\ \mathbf{C}_{Cini} &= [\begin{matrix} -2.2420e-3 & 2.4713e-1 \end{matrix}], \\ \mathbf{D}_{Cini} &= 6.7500e-2. \end{aligned}$$

Then

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{X}_{ini}) &= \begin{bmatrix} a_0 & c_0 \\ d_0 & b_0 \end{bmatrix} \\ &= \begin{bmatrix} -1.0061e-2 & 9.8327e-1 \\ -9.9386e-1 & 1.6731e0 \end{bmatrix}, \\ \begin{bmatrix} b_P & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\ \begin{bmatrix} c_P & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

Our algorithm produces the solution $\upsilon(\mathbf{X}_{ini}) = 1.0814e-2$. Li [4] addressed a FWL measure $\tilde{\upsilon}$ approximating υ. In order to compare $\tilde{\upsilon}$ and υ, we compute $\tilde{\upsilon}(\mathbf{X}_{ini}) = 5.2463e-3$ by the method given in [4]. Obviously, $\tilde{\upsilon}(\mathbf{X}_{ini})$ is more conservative than $\upsilon(\mathbf{X}_{ini})$. For the digital filter, an approach given in [1] can be used to maximize $\tilde{\upsilon}$ within $\mathcal{S}_C$ and obtain an optimal realization $\mathbf{X}_{opt}$ whose

$$\begin{aligned} \mathbf{A}_{Copt} &= \begin{bmatrix} 8.3152e-1 & 5.1863e-1 \\ -5.1863e-1 & 8.3152e-1 \end{bmatrix}, \\ \mathbf{B}_{Copt} &= \begin{bmatrix} -9.4809e-1 \\ 1.6686e0 \end{bmatrix}, \\ \mathbf{C}_{Copt} &= [\begin{matrix} 1.1853e-1 & 2.1544e-1 \end{matrix}], \\ \mathbf{D}_{Copt} &= 6.7500e-2. \end{aligned}$$

With $\mathbf{X}_{opt}$, we have $\tilde{\upsilon}(\mathbf{X}_{opt}) = 1.0000e-2$ and $\upsilon(\mathbf{X}_{opt}) = 1.4509e-2$. $\tilde{\upsilon}(\mathbf{X}_{opt})$ is also more conservative than $\upsilon(\mathbf{X}_{opt})$.

5   Conclusions

An interesting fact is that the stability of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ for any $[\begin{matrix} a & b & c & d \end{matrix}]^T \in \mathcal{P}_\eta$ is equivalent to the stability

of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ at $\mathcal{P}_\eta$'s vertices. Based on this observation, an analytic method has been proposed to compute the FWL stability measure υ for second order digital systems. As the "vertex result" does not hold in the systems of order higher than two, the proposed method can not be extended to those systems.

Appendix

$e_1 \triangleq a_0 b_0 - c_0 d_0 - 1$
$e_2 \triangleq a_0 b_0 - c_0 d_0 + a_0 + b_0 + 1$
$e_3 \triangleq a_0 b_0 - c_0 d_0 - a_0 - b_0 + 1$

$f_1(\mathbf{p}_1(\eta)) = g_{11}\eta + e_1 = 0$
$g_{11} = a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P$

$f_1(\mathbf{p}_2(\eta)) = h_{12}\eta^2 + g_{12}\eta + e_1 = 0$
$h_{12} = -2b_P c_P$
$g_{12} = a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P$

$f_1(\mathbf{p}_3(\eta)) = h_{13}\eta^2 + g_{13}\eta + e_1 = 0$
$h_{13} = -2b_P c_P$
$g_{13} = -a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P$

$f_1(\mathbf{p}_4(\eta)) = h_{14}\eta^2 + g_{14}\eta + e_1 = 0$
$h_{14} = 2b_P c_P$
$g_{14} = a_0 + b_0 b_P c_P - c_0 c_P + d_0 b_P$

$f_1(\mathbf{p}_5(\eta)) = h_{15}\eta^2 + g_{15}\eta + e_1 = 0$
$h_{15} = 2b_P c_P$
$g_{15} = a_0 + b_0 b_P c_P + c_0 c_P - d_0 b_P$

$f_1(\mathbf{p}_6(\eta)) = g_{16}\eta + e_1 = 0$
$g_{16} = -a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P$

$f_1(\mathbf{p}_7(\eta)) = g_{17}\eta + e_1 = 0$
$g_{17} = a_0 - b_0 b_P c_P - c_0 c_P + d_0 b_P$

$f_1(\mathbf{p}_8(\eta)) = g_{18}\eta + e_1 = 0$
$g_{18} = a_0 - b_0 b_P c_P + c_0 c_P - d_0 b_P$

$f_2(\mathbf{p}_1(\eta)) = g_{21}\eta + e_2 = 0$
$g_{21} = a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P + b_P c_P + 1$

$f_2(\mathbf{p}_2(\eta)) = h_{22}\eta^2 + g_{22}\eta + e_2 = 0$
$h_{22} = -2b_P c_P$

$g_{22} = a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P - b_P c_P + 1$

$f_2(\mathbf{p}_3(\eta)) = h_{23}\eta^2 + g_{23}\eta + e_2 = 0$
$h_{23} = -2b_P c_P$
$g_{23} = -a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P + b_P c_P - 1$

$f_2(\mathbf{p}_4(\eta)) = h_{24}\eta^2 + g_{24}\eta + e_2 = 0$
$h_{24} = 2b_P c_P$
$g_{24} = a_0 + b_0 b_P c_P - c_0 c_P + d_0 b_P + b_P c_P + 1$

$f_2(\mathbf{p}_5(\eta)) = h_{25}\eta^2 + g_{25}\eta + e_2 = 0$
$h_{25} = 2b_P c_P$
$g_{25} = a_0 + b_0 b_P c_P + c_0 c_P - d_0 b_P + b_P c_P + 1$

$f_2(\mathbf{p}_6(\eta)) = g_{26}\eta + e_2 = 0$
$g_{26} = -a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P - b_P c_P - 1$

$f_2(\mathbf{p}_7(\eta)) = g_{27}\eta + e_2 = 0$
$g_{27} = a_0 - b_0 b_P c_P - c_0 c_P + d_0 b_P - b_P c_P + 1$

$f_2(\mathbf{p}_8(\eta)) = g_{28}\eta + e_2 = 0$
$g_{28} = a_0 - b_0 b_P c_P + c_0 c_P - d_0 b_P - b_P c_P + 1$

$f_3(\mathbf{p}_1(\eta)) = g_{31}\eta + e_3 = 0$
$g_{31} = a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P - b_P c_P - 1$

$f_3(\mathbf{p}_2(\eta)) = h_{32}\eta^2 + g_{32}\eta + e_3 = 0$
$h_{32} = -2b_P c_P$
$g_{32} = a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P + b_P c_P - 1$

$f_3(\mathbf{p}_3(\eta)) = h_{33}\eta^2 + g_{33}\eta + e_3 = 0$
$h_{33} = -2b_P c_P$
$g_{33} = -a_0 + b_0 b_P c_P - c_0 c_P - d_0 b_P - b_P c_P + 1$

$f_3(\mathbf{p}_4(\eta)) = h_{34}\eta^2 + g_{34}\eta + e_3 = 0$
$h_{34} = 2b_P c_P$
$g_{34} = a_0 + b_0 b_P c_P - c_0 c_P + d_0 b_P - b_P c_P - 1$

$f_3(\mathbf{p}_5(\eta)) = h_{35}\eta^2 + g_{35}\eta + e_3 = 0$
$h_{35} = 2b_P c_P$
$g_{35} = a_0 + b_0 b_P c_P + c_0 c_P - d_0 b_P - b_P c_P - 1$

$f_3(\mathbf{p}_6(\eta)) = g_{36}\eta + e_3 = 0$
$g_{36} = -a_0 - b_0 b_P c_P - c_0 c_P - d_0 b_P + b_P c_P + 1$

$f_3(\mathbf{p}_7(\eta)) = g_{37}\eta + e_3 = 0$

$$g_{37} = a_0 - b_0 b_P c_P - c_0 c_P + d_0 b_P + b_P c_P - 1$$

$$f_3(\mathbf{p}_8(\eta)) = g_{38}\eta + e_3 = 0$$
$$g_{38} = a_0 - b_0 b_P c_P + c_0 c_P - d_0 b_P + b_P c_P - 1$$

Acknowledgements

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] P. Mantey, "Eigenvalue sensitivity and state-variable selection," *IEEE Trans. Automatic Control*, Vol.13, No.3, pp.263–269, 1968.

[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] I.J. Fialho, and T.T. Georgiou, "Computational algorithms for sparse optimal digital controller realizations," In R.S.H. Istepanian and J.F. Whidborne, eds., *Digital Controller Implementation and Fragility: A Modern Perspective*, London: Springer Verlag. pp.105–121. 2001.

[9] W. Yu, and H. Ko, "Improved eigenvalue sensitivity for finite-precision digital controller realisations via orthogonal Hermitian transform," *IEE Proc. Control Theory Applications*, Vol.150, No.4, pp.365–375, 2003.

[10] E. Jury, "Additions to 'Notes on the stability criterion for linear discrete systems'," *IRE Trans. Automatic Control*, Vol.6, No.3, pp.342–343, 1961.

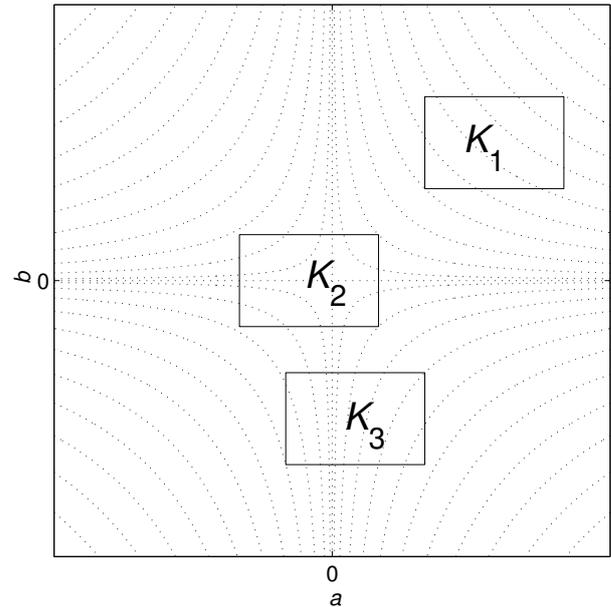[11] R.A. Hunt, *Calculus of a Single Variable.* New York: HarperCollins College Publishers, 1994.

Figure 1: Rectangle blocks in the ab plane grained with $ab = \omega$.