Sensitivity Minimization for Controller Implementations: Fixed-Point Approach

Hsien-Ju Ko and Wen-Shyong Yu

Abstract— **In this paper, a novel approach is proposed to analyze and minimize fixed-point errors for digital controller implementations based on sensitivity measures in magnitude and phase-supplement-angle of eigenvalues. First, uncertainties of the controller parameters caused by roundoff and computational errors using fixed-point computations are expressed in function of register length. Then, a stability criterion of the closed-loop system based on fixed-point statistical model is derived by means of small gain theorem and Bellman-Grownwall Lemma. Thus, a measure that combines sensitivities of the magnitude and phase-supplement-angle of the closed-loop system eigenvalues with respect to controller parameters is constructed in the sense of mixed matrix-2/Frobenius norms. This measure is minimized by an optimal similarity transformation obtained from an analytically algebraic method. Based upon this transformation as well as the stability criterion, a least register length can be obtained. Finally, an example of the simplified model of a Vertical Take off and Landing (VTOL) aircraft is performed to illustrate the effectiveness of the proposed scheme.**

I. Introduction

Digital controller implementations have been widely used for many control engineering applications. However, the used computers are always with finite-word length (FWL). Thus, the well-designed control system may be degenerated or become unstable due to FWL effects. To date, there are two streams of research into studying the FWL effects in the literatures: one focuses on the design of digital filters affected by FWL effects[1]–[4]; the second approaches the design and implementation of the digital regulators or controllers in digital computers for controlling a physical system [5]–[9]. The infinite precision controllers implemented with state space realizations work with either fixed or floating point computations in digital computers. In general, processors using the mode of fixed-point arithmetic are generally with lower costs, lower complexity of implementations, easier for programming and higher operational speed, etc.. Wilkinson [1] first proposed an algebraic analysis for rounding errors in digital computers, either in fixed or floating point computations. Related research can further be found in [2]. There exist infinite realizations correspond to the same stabilizing controller in infinite precisions, and yield the same closed-loop system stability and performance. However, the closed-loop system with different controller realizations may have different output performance when these controllers are implemented in digital computers with FWL. Therefore, selecting a proper transformation may improve robustness of the closed-loop system under FWL consideration. In [8], the authors proposed the sensitivity in magnitude of eigenvalues to analyze the closed-loop system stability subject to FWL effects. However, if the system is with more than one eigenvalue and if it has eigenvalues that are complex with the same magnitudes but different phase angles, taking only the sensitivity of magnitude of the eigenvalue into account may lead to unsatisfied stability margin of the system. However, the method still lacked a systematic method to find the optimal transformation. In [9], the authors proposed an analytically algebraic method for solving an optimal transformation to achieve the minimal sensitivity subject to FWL effects. Since the eigenvalues may be changed to complex-valued ones due to FWL effects even if they are real originally, this method yet does not consider the phase of the sensitivity.

In this paper, a novel approach is proposed for analyzing the closed-loop stability for digital controller implementations subject to FWL effects based on the sensitivities of magnitude and supplement-angle of eigenvalues. First, uncertainties of the controller parameters caused by roundoff and computational errors using fixed-point computations are expressed in function of register length. Then, a stability criterion of the closed-loop system based on fixed-point statistical model is derived by means of small gain theorem and Bellman-Grownwall Lemma. Thus, a measure that combines sensitivities in magnitude and supplement-angle of eigenvalues with respect to controller parameters is minimized by an optimal similarity transformation obtained from an analytically algebraic method. By substituting the obtained optimal transformation into the stability criterion, a least word length less than or equal to the original one can be found. Finally, an example of the simplified model of a Vertical Take off and Landing (VTOL) aircraft is performed to illustrate the effectiveness of the proposed scheme.

II. Problem formulation and fixed-point arithmetic

For implementing the digital controller, we may consider a linear time-invariant hybrid feedback control system shown in Fig. 1, where the discrete state space models of the controlled plant

$$\left\{ \begin{array}{rcl} \mathbf{x}_p(k+1) & = & \mathbf{A}_p\mathbf{x}_p(k) + \mathbf{B}_p\mathbf{u}(k) \\ \mathbf{z}_p(k) & = & \mathbf{M}_p\mathbf{x}_p(k) \\ \mathbf{y}_p(k) & = & \mathbf{C}_p\mathbf{x}_p(k) \end{array} \right. \tag{1}$$

Hsien-Ju Ko is the Ph D. candidate of the Department of Electrical Engineering, Tatung University, 40 Chung-Shan North Rd. 3rd. Sec., Taipei, 104 Taiwan. robert@ctr3.ee.ttu.edu.tw

Wen-Shyong Yu is the Faculty of the Department of Electrical Engineering, Tatung University. wsyu@ctr1.ee.ttu.edu.tw

0-7803-8335-4/04/$17.00 ©2004 AACC

and the corresponding dynamical controller with the realizations $\{\mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c, \mathbf{D}_c\}$

$$\left\{ \begin{array}{rcl} \mathbf{x}_c(k+1) & = & \mathbf{A}_c\mathbf{x}_c(k) + \mathbf{B}_c\mathbf{z}_p(k) \\ \mathbf{u}(k) & = & \mathbf{C}_c\mathbf{x}_c(k) + \mathbf{D}_c\mathbf{z}_p(k) + \mathbf{r}_r(k) \end{array} \right. \tag{2}$$

where $\mathbf{x}_p(k)$, $\mathbf{u}(k)$, $\mathbf{y}_p(k)$, and $\mathbf{z}_p(k)$ are the system state, control input, system output and measurement output vectors of the controlled plant, respectively, and $\mathbf{x}_c(k)$ and $\mathbf{r}_r(k)$ are the state vector of the controller and reference input vector, respectively. It is assumed that the controlled plant is controllable and observable. Substituting (2) into (1) and considering FWL effects occurred in the digital controller, we obtain the following closed-loop control system:

$$\begin{array}{rcl} \mathbf{x}^*(k+1) & = & \mathbf{A}\mathbf{x}^*(k) + \mathbf{B}fi\{\mathbf{G}^*(\mathbf{M}\mathbf{x}^*(k) + \mathbf{I}_1\mathbf{n}_{AD})\} \\ & & +\mathbf{B}(\mathbf{r}(k) + \mathbf{n}_{DA}) \\ \mathbf{y}^*(k) & = & \mathbf{C}\mathbf{x}^*(k) + \mathbf{I}_0 fl\{\mathbf{G}^*(\mathbf{M}\mathbf{x}^*(k) + \mathbf{n}_{AD})\} \\ & & +\mathbf{I}_0(\mathbf{r}(k) + \mathbf{n}_{DA}) \end{array} \tag{3}$$

where $fi(\cdot)$ and the superscript * are denoted the fixed-point multiplicative operation and the roundoff/quantization operations on states and parameter matrices, respectively, and the A/D and D/A conversion errors are $\boldsymbol{\varepsilon}_{AD}$ and $\boldsymbol{\varepsilon}_{DA}$, respectively, and they are assumed to be zero mean, mutually independent uniform white noise, and the notations are defined as: $\mathbf{x}^*(k) = \begin{bmatrix} \mathbf{x}_p^*(k) \\ \mathbf{x}_c^*(k) \end{bmatrix}$, $\mathbf{y}^*(k) = \begin{bmatrix} \mathbf{y}_p^*(k) \\ \mathbf{u}^*(k) \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} \mathbf{A}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} \mathbf{B}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} \mathbf{C}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} \mathbf{D}_p \\ \mathbf{0} \end{bmatrix}$, $\mathbf{G}^* = \begin{bmatrix} \mathbf{D}_c^* & \mathbf{C}_c^* \\ \mathbf{B}_c^* & \mathbf{A}_c^* \end{bmatrix}$, $\mathbf{I}_0 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, $\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbf{M} = \begin{bmatrix} \mathbf{M}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$, $\mathbf{n}_{AD} = \begin{bmatrix} \boldsymbol{\varepsilon}_{AD} \\ 0 \end{bmatrix}$, $\mathbf{n}_{DA} = \begin{bmatrix} \boldsymbol{\varepsilon}_{DA} \\ 0 \end{bmatrix}$, $\mathbf{r}(k) = \begin{bmatrix} \mathbf{r}_r(k) \\ \mathbf{0} \end{bmatrix}$.

In fixed-point format, the word length of a real number can be divided into three parts: sign bit, bits of integer part and of fractional part, and they are denoted W_s, W_i, and W_f, respectively. Thus, total word length of fixed-point form is

$$W = 1 + W_i + W_f. \tag{4}$$

When a real number is represented by fixed-point form, and we consider that the overflow is not happened, the roundoff error can be defined by [1]

$$fi(a) = a + \epsilon_{rd} \tag{5}$$

where a is any real number (In (5), a can be seen as $a \cdot 1$), and ϵ_{rd} is denoted the roundoff error. This error is bounded by

$$|\epsilon_{rd}| < 2^{-W_f}. \tag{6}$$

When two real numbers a and b are multiplied, and since there is no overflow caused, we can assume that a and b are both in the range $(0, 1)$, and the fixed-point error representation can be described by

$$fi(ab) = ab + \hat{\epsilon}_{op} \tag{7}$$

where $\hat{\epsilon}_{op}$ is the operational error and is uniformly distributed in the range $(-2^{-W_f}, 2^{-W_f})$. If we normalize the small random quantity $\hat{\epsilon}_{op}$ to be with uniform probability distribution in the range $(-1, 1)$ denoted as ϵ_{op}, (7) can be rewrite as

$$fi(ab) = ab + \Delta\epsilon_{op} \tag{8}$$

where $\Delta = 2^{-W_f}$. Based on (8), the rounded inner product of two vectors $\mathbf{a}_p, \mathbf{b}_q \in \mathbb{R}^n$ can be given by

$$\sum_{i=1}^{n} fi(a_{pi}b_{qi}) = (a_{p1}b_{q1} + a_{p2}b_{q2} + \cdots + a_{pn}b_{qn}) + \Delta m_{(pq)} \tag{9}$$

where a_{pi}'s and b_{qi}'s are elements of $\mathbf{a}_p$ and $\mathbf{b}_q$, respectively, and $m_{(pq)} = m_{(pq)1} + m_{(pq)2} + \cdots + m_{(pq)n}$, and each $m_{(pq)i}$ is mutually independent and uniformly distributed on $(-1, 1)$. Thus, we have

$$\text{Var}(m_{(pq)i}) = \frac{1}{3} \tag{10}$$

where $\text{Var}(\cdot)$ denotes the variance of $(\cdot)$. Further, for any matrices $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{n \times q}$, we have

$$\begin{array}{rcl} fi(\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{Q}}) & = & \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{Q}} + \Delta \begin{bmatrix} m_{(11)} & m_{(12)} & \cdots & m_{(1q)} \\ m_{(21)} & m_{(22)} & \cdots & m_{(2q)} \\ \vdots & \vdots & \ddots & \vdots \\ m_{(n1)} & m_{(n2)} & \cdots & m_{(nq)} \end{bmatrix} \\ & \triangleq & \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{Q}} + \Delta\boldsymbol{\mathcal{M}} \end{array} \tag{11}$$

where $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n \times q}$ is with stochastic elements uniformly distributed on $(-1, 1)$ and they are mutually independent, and satisfies

$$\|\boldsymbol{\mathcal{M}}\|_2 = \sqrt{\lambda_{\max}(E[\boldsymbol{\mathcal{M}}^\top \boldsymbol{\mathcal{M}}])} = \sqrt{\frac{nq}{3}}. \tag{12}$$

III. Stability analysis for finite fixed-point controller implementation

According to (11), the closed-loop system with a finite fixed-point digital controller implementation, (3) can be represented as

$$\mathbf{x}(k+1)^* = \bar{\mathbf{A}}^*\mathbf{x}^*(k) + \Delta\mathbf{B}\boldsymbol{\mathcal{M}}_1 + \mathbf{B}\mathbf{G}^*\mathbf{I}_1\mathbf{n}_{DA} + \mathbf{B}\mathbf{r}(k) \tag{13}$$

$$\begin{array}{l} \mathbf{y}(k)^* = \mathbf{C}\mathbf{x}^*(k) + \mathbf{I}_0\mathbf{G}^*\mathbf{M}\mathbf{x}^*(k) + \Delta\mathbf{I}_0\boldsymbol{\mathcal{M}}_2 \\ \qquad + \mathbf{I}_0\mathbf{G}^*\mathbf{n}_{AD} + \mathbf{I}_0\mathbf{n}_{DA} + \mathbf{I}_0\mathbf{r}(k) \end{array} \tag{14}$$

where $\mathbf{A} + \mathbf{B}\mathbf{G}^*\mathbf{M} \equiv \bar{\mathbf{A}}^*$, and $\boldsymbol{\mathcal{M}}_1$ and $\boldsymbol{\mathcal{M}}_2$ are with stochastic elements uniformly distributed on $(-1, 1)$ and they are mutually independent. For obtaining specific quantization error bound to derive the stability criterion in terms of word length W in (4), we have the following: Let Ω denote all controller realizations and $\boldsymbol{\Phi}$=$\{\mathbf{A}_c$, $\mathbf{B}_c$, $\mathbf{C}_c$, $\mathbf{D}_c$, $\mathbf{E}_c\} \in \Omega$ some realization. Suppose that ϕ_ℓ is the ℓth nontrivial element of the parameter matrices of some realization. For quantization errors, we have

$$\hat{\phi}_\ell = \left\{ \begin{array}{ll} sgn(\phi_\ell)(\phi_\ell + 2^{-W_f}), & \text{if } \phi_\ell \text{ is not an integer.} \\ sgn(\phi_\ell)\phi_\ell, & \text{if } \phi_\ell \text{ is an integer.} \end{array} \right. \tag{15}$$

Since $\|\hat{\boldsymbol{\Phi}}_i - \boldsymbol{\Phi}_i\|_2 \geq \|\boldsymbol{\Phi}_i^* - \boldsymbol{\Phi}\|_{\mathbf{2}}$, where $\boldsymbol{\Phi}_{\mathbf{i}}$ is some parameter matrix in $\boldsymbol{\Phi}$, a stability criterion can be developed by the following.

Theorem 1:: If the stability criterion

$$r_s + \zeta(\varepsilon_{q_1} + \varepsilon_{c_1}) < 1 \tag{16}$$

is satisfied, then

$$\|\mathbf{x}^*(k)\|_2 \leq \sup_{0\leq i<\infty} \frac{\zeta(\|\mathbf{B}\|_2\|\mathbf{r}(i)\|_2 + n_x)}{1-(r_s+\zeta(\varepsilon_{q_1}+\varepsilon_{c_1}))} \tag{17}$$

$$\|\mathbf{y}^*(k)\|_2 \leq \sup_{0\leq i<\infty} \Big(\frac{\zeta(\|\mathbf{C}\|_2 + \|\mathbf{I}_0\mathbf{G}\mathbf{M}\|_2 + \varepsilon_{q_2})}{1-(r_s+\zeta(\varepsilon_{q_1}+\varepsilon_{c_1}))} \\ \times(\|\mathbf{B}\|_2 + n_x) + \|\mathbf{I}_0\|_2\Big)\|\mathbf{r}(i)\|_2 + \varepsilon_{c_2} + n_y \tag{18}$$

where r_s denotes the the *spectral radius*, (i.e., $r_s = \max_i|\lambda_i(\bar{\mathbf{A}})|$) $\zeta = \max_k(\|\bar{\mathbf{A}}^k\|/r_s^k)$, $n_x = \|\mathbf{B}\hat{\mathbf{G}}\mathbf{I}_1\mathbf{n}_{DA}\|_2$, $\varepsilon_{q_1} = \|\hat{\bar{\mathbf{A}}} - \bar{\mathbf{A}}\|_2 = 2^{-W_f}\|\mathbf{B}\ sgn(\mathbf{G})\mathbf{M}\|_2 \overset{\triangle}{=} 2^{-W_f}\varepsilon'_{q_1}$, $\varepsilon_{c_1} = \|\Delta\mathbf{B}\boldsymbol{\mathcal{M}}_1\|_2 = 2^{-W_f}\|\mathbf{B}\boldsymbol{\mathcal{M}}_1\|_2 \overset{\triangle}{=} 2^{-W_f}\varepsilon'_{c_1}$, $\varepsilon_{q_2} = \|\mathbf{I}_0\hat{\mathbf{G}}\mathbf{M} - \mathbf{I}_0\mathbf{G}\mathbf{M}\|_2$, $\varepsilon_{c_1} = \|\Delta\mathbf{I}_0\boldsymbol{\mathcal{M}}_2\|_2$, and $n_y = \|\mathbf{I}_0\hat{\mathbf{G}}\mathbf{n}_{AD}\|_2 + \|\mathbf{I}_0\mathbf{n}_{DA}\|_2$

Proof: The proof can be obtained by using small-gain theory and Bellman-Grownwall Lemma. For fitting with the conference length, the detailed procedures are neglected here. ■

Based on (15) and the criterion (16), an estimated bit-number that the stability of the closed-loop system is guaranteed can be obtained as follows:

$$W_{(est)} = 1 + int\left[\log_2\left(\frac{\max_{\ell}|\phi_\ell|\zeta(\varepsilon'_{q_1}+\varepsilon'_{c_1})}{1-r_s}\right)\right] \tag{19}$$

where $int[\cdot]$ denotes the smallest integer equal to or greater than $\cdot$.

Remark 1::

1) In this paper, all controller parameters are assumed to be no overflow happened. Thus, the stability criterion focuses on fractional word length.
2) The stability criterion is only a sufficient condition and may be conservative in some controller design cases.

IV. Sensitivities of the Magnitude and Supplement-Angle of Eigenvalues

In general, a closed-loop system has more than one eigenvalue. If some of the eigenvalues are changed to complex-valued ones due to FWL effects even if they are real originally, we should consider the behavior of all eigenvalues within the unit circle subject to FWL effects. For more clarity, suppose that one of the eigenvalues locates at a point o_2 as shown in Fig. 2. Obviously, it may drift to an unknown position subject to FWL effects and may affect the stability margin of the closed-loop system. Therefore, the closed-loop system has larger stability margin if this eigenvalue is drifted to point d than that it is drifted to point c. On the contrary, if $|\delta\theta| = \pi$, we have the largest stability margin of the system.

Let λ_k, $k = 1, 2, \ldots, n$, be the kth eigenvalue of the closed-loop system matrix $\bar{\mathbf{A}}$. Then, we have

$$\lambda_k = |\lambda_k|e^{j\theta_k} = |\lambda_k|\big(\cos(\theta_k) + j\sin(\theta_k)\big) \tag{20}$$

where $|\lambda_k|$ and θ_k are the magnitude and phase angle in radians of the kth eigenvalue, respectively. The infinite-precision closed-loop system matrix $\bar{\mathbf{A}} = \mathbf{A} + \mathbf{B}\mathbf{G}\mathbf{M}$ can be rewritten as:

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_p + \mathbf{B}_p\mathbf{D}_c\mathbf{M}_p & \mathbf{B}_p\mathbf{C}_c \\ \mathbf{B}_c\mathbf{M}_p & \mathbf{A}_c \end{bmatrix} \tag{21}$$

It is seen that the closed-loop system is stable *if and only if* all the eigenvalues of $\bar{\mathbf{A}}$ distribute within the unit circle.

Lemma 1: :[8] Suppose that the closed-loop realization

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{\mathbf{A}}_{(11)} & \bar{\mathbf{A}}_{(12)} \\ \bar{\mathbf{A}}_{(21)} & \bar{\mathbf{A}}_{(22)} \end{bmatrix} \tag{22}$$

is diagonalizable. Let

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{x}_{k(1)} \\ \mathbf{x}_{k(2)} \end{bmatrix} \text{ and } \mathbf{y}_k = \begin{bmatrix} \mathbf{y}_{k(1)} \\ \mathbf{y}_{k(2)} \end{bmatrix} \tag{23}$$

be the right and left eigenvectors of $\bar{\mathbf{A}}$ corresponding to the kth eigenvalue λ_k, respectively. Then, we have

$$\frac{\partial|\lambda_k|}{\partial\bar{\mathbf{A}}} = \begin{bmatrix} \frac{Re(\lambda_k\mathbf{y}_{k(1)}\mathbf{x}_{k(1)}^{\mathcal{H}})}{|\lambda_k|} & \frac{Re(\lambda_k\mathbf{y}_{k(1)}\mathbf{x}_{k(2)}^{\mathcal{H}})}{|\lambda_k|} \\ \frac{Re(\lambda_k\mathbf{y}_{k(2)}\mathbf{x}_{k(1)}^{\mathcal{H}})}{|\lambda_k|} & \frac{Re(\lambda_k\mathbf{y}_{k(2)}\mathbf{x}_{k(2)}^{\mathcal{H}})}{|\lambda_k|} \end{bmatrix} \tag{24}$$

where $Re(\cdot)$ denotes the real part of $\cdot$ and the superscript $^{\mathcal{H}}$ the complex conjugate transpose operator. ■

Let

$$\mathbf{X} = [\mathbf{X}_{(1)}^\top \ \ \mathbf{X}_{(2)}^\top]^\top = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n] \tag{25}$$

$$\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_{(1)}^\top \ \ \tilde{\mathbf{Y}}_{(2)}^\top]^\top = [(\mathbf{Y}_{(1)}\mathbf{M}_y)^\top \ \ (\mathbf{Y}_{(2)}\mathbf{M}_y)^\top]^\top \tag{26}$$

where $\mathbf{x}_k$ and $\mathbf{y}_k$ are the right and left eigenvectors of some realization of $\bar{\mathbf{A}}$ corresponding to the kth eigenvalue, respectively, and $\mathbf{M}_y = diag\{\frac{\lambda_1}{|\lambda_1|}, \frac{\lambda_2}{|\lambda_2|}, \cdots, \frac{\lambda_n}{|\lambda_n|}\}$. Also from (25) and (26), we have $|\lambda_k|\cos(\theta_k) = Re(\mathbf{y}_k^{\mathcal{H}}\bar{\mathbf{A}}\mathbf{x}_k)$ and $|\lambda_k|\sin(\theta_k) = Im(\mathbf{y}_k^{\mathcal{H}}\bar{\mathbf{A}}\mathbf{x}_k)$, where $Im(\cdot)$ denotes the imaginary part of $\cdot$. By taking the tangent of supplement angle, $\theta'_k = \pi - \theta_k$, as shown in Fig. 2, we have

$$\tan(\theta'_k) = \frac{\frac{1}{j}(\mathbf{y}_k^\top\bar{\mathbf{A}}\mathbf{x}_k^* - \mathbf{y}_k^{\mathcal{H}}\bar{\mathbf{A}}\mathbf{x}_k)}{(\mathbf{y}_k^{\mathcal{H}}\bar{\mathbf{A}}\mathbf{x}_k + \mathbf{y}_k^\top\bar{\mathbf{A}}\mathbf{x}_k^*)} \overset{\triangle}{=} \frac{f_k(\bar{\mathbf{A}})}{g_k(\bar{\mathbf{A}})}. \tag{27}$$

It is seen that some stable eigenvalues of the closed-loop system will be perturbed subject to FWL effects, which results in the element ϕ_ℓ being perturbed toward $\phi_\ell + \delta\phi_\ell$, $\forall\ell$, where the uncertainty $\delta\phi_\ell$ is bounded. Since a small parameter uncertainty $\delta\phi_\ell = \tilde{\phi}_\ell - \phi_\ell$ will shift the eigenvalue λ_k to $\tilde{\lambda}_k$, taking the first-order approximation will lead to

$$\delta|\lambda_k| \overset{\triangle}{=} |\tilde{\lambda}_k| - |\lambda_k| = \sum_{\ell=1}^{N}\frac{\partial|\lambda_k|}{\partial\phi_\ell}\delta\phi_\ell \tag{28}$$

$$\delta\theta'_k \overset{\triangle}{=} \tilde{\theta}'_k - \theta'_k = \sum_{\ell=1}^{N}\frac{\partial\theta'_k}{\partial\phi_\ell}\delta\phi_\ell \tag{29}$$

respectively. Since we indeed do not know which eigenvalue shifts near to the unit circle or moves across the unit circle, we thus consider the sensitivity for all eigenvalues subject to elements of the system matrices as follows:

$$\Gamma \triangleq \sum_{\ell=1}^{N} \left(\sum_{k=1}^{n} \left| \frac{\partial |\lambda_k|}{\partial \phi_\ell} \right| \right)^2 \tag{30}$$

and

$$\Theta' \triangleq \sum_{\ell=1}^{N} \left(\sum_{k=1}^{n} \left| \frac{\partial \theta'_k}{\partial \phi_\ell} \right| \right)^2 . \tag{31}$$

Let $|S|$ be denoted as:

$$|S| \triangleq \begin{bmatrix} |s_{11}| & \cdots & |s_{1n}| \\ \vdots & \ddots & \vdots \\ |s_{m1}| & \cdots & |s_{mn}| \end{bmatrix} \tag{32}$$

where $|\cdot|$ denotes the absolute value of $\cdot$ and s_j is the ij-component of a matrix $S \in \mathbb{C}^{m\times n}$. Therefore from (30) and (31), we have by simple algebraic manipulations

$$\Gamma{=}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{A}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{B}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{C}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{D}_c}\right|\right\|_F^2 \tag{33}$$

$$\Theta'{=}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\theta'_k|}{\partial \mathbf{A}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\theta'_k|}{\partial \mathbf{B}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\theta'_k|}{\partial \mathbf{C}_c}\right|\right\|_F^2{+}\left\|\sum_{k=1}^{n}\left|\frac{\partial|\theta'_k|}{\partial \mathbf{D}_c}\right|\right\|_F^2 \tag{34}$$

for some controller realization, where the subscription $_F$ denotes the Frobenius norm. Now comparing (21) with (22)

for some realization, we have $\bar{\mathbf{A}}_{(11)} = \mathbf{A}_p + \mathbf{B}_p\mathbf{D}_c\mathbf{M}_p$, $\bar{\mathbf{A}}_{(12)} = \mathbf{B}_p\mathbf{C}_c$, $\bar{\mathbf{A}}_{(21)} = \mathbf{B}_c\mathbf{M}_p$, and $\bar{\mathbf{A}}_{(22)} = \mathbf{A}_c$. Since λ_k is the kth eigenvalue of $\bar{\mathbf{A}}$, it follows from Lemma 1 that

$$\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{A}_c}\right| = \sum_{k=1}^{n}\left| Re\Big(\frac{\lambda_k \mathbf{y}_{k(2)}\mathbf{x}_{k(2)}^{\mathcal{H}}}{|\lambda_k|}\Big)\right| \tag{35}$$

$$\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{B}_c}\right| = \sum_{k=1}^{n}\left| Re\Big(\frac{\lambda_k \mathbf{y}_{k(2)}\mathbf{x}_{k(1)}^{\mathcal{H}}}{|\lambda_k|}\Big)\mathbf{M}_p^\top\right| \tag{36}$$

$$\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{C}_c}\right| = \sum_{k=1}^{n}\left| \mathbf{B}_p^\top Re\Big(\frac{\lambda_k \mathbf{y}_{k(1)}\mathbf{x}_{k(2)}^{\mathcal{H}}}{|\lambda_k|}\Big)\right| \tag{37}$$

$$\sum_{k=1}^{n}\left|\frac{\partial|\lambda_k|}{\partial \mathbf{D}_c}\right| = \sum_{k=1}^{n}\left| \mathbf{B}_p^\top Re\Big(\frac{\lambda_k \mathbf{y}_{k(1)}\mathbf{x}_{k(1)}^{\mathcal{H}}}{|\lambda_k|}\Big)\mathbf{M}_p^\top\right| \tag{38}$$

Based on the results (35)–(38), by the fact of the inequalities

$$\left\|\sum_{k=1}^{n}\left|Re(\frac{\lambda_k \mathbf{y}_{k(i)}\mathbf{x}_{k(j)}^{\mathcal{H}}}{|\lambda_k|})\right|\right\|_F \leq \left\| |\tilde{\mathbf{Y}}_{(i)}| |\mathbf{X}_{(j)}^{\mathcal{H}}| \right\|_F ,\ i,j=1,2, \tag{39}$$

the properties of $\| \, |\cdot| \, \|_F = \|\cdot\|_F$ and $\|\cdot\|_F \leq \sqrt{\nu}\|\cdot\|_2$ where ν is the row number of $\cdot$, the fact of $\|\tilde{\mathbf{Y}}_{(i)}\| = \|\mathbf{Y}_{(i)}\|$, for $i=1,2$ (since $\tilde{\mathbf{Y}}_{(i)} = \mathbf{Y}_{(i)}\mathbf{M}_y$ in (26) and $\mathbf{M}_y$ is orthogonal), and the submultiplicative property of the Frobenius norm, we have

$$\begin{aligned}\Gamma &\leq \nu^2\left\|\mathbf{Y}_{(2)}\right\|_2^2\left\|\mathbf{X}_{(2)}\right\|_2^2 + c_1^2\left\|\mathbf{Y}_{(2)}\right\|_F^2 + c_2^2\left\|\mathbf{X}_{(2)}\right\|_F^2 + c_3^2 \\ &\triangleq \ \bar{\Gamma}\end{aligned} \tag{40}$$

where $c_1 = \left\|\mathbf{X}_{(1)}^{\mathcal{H}}\mathbf{M}_p^\top\right\|_F$, $c_2 = \left\|\mathbf{B}_p^\top\mathbf{Y}_{(1)}\right\|_F$, $c_3 = \left\|\left|\mathbf{B}_p^\top\tilde{\mathbf{Y}}_{(1)}\right|\left|\mathbf{X}_{(1)}^{\mathcal{H}}\mathbf{M}_p^\top\right|\right\|_F$, and ν is the row number of $\mathbf{X}_{(2)}$. Note that the constants c_1, c_2, and c_3 are closed-loop structure independence.

Further from (27), the sensitivity for supplement-angle of the kth eigenvalue subject to elements of the system matrices is given by

$$\frac{\partial\theta'_k}{\partial\bar{\mathbf{A}}} = c_{k1}Re(\mathbf{y}_k\mathbf{x}_k^{\mathcal{H}}) - c_{k2}Im(\mathbf{y}_k\mathbf{x}_k^{\mathcal{H}}) \tag{41}$$

where $c_{k1} = \frac{Im(\lambda_k)}{|\lambda_k|^2}$ and $c_{k2} = \frac{Re(\lambda_k)}{|\lambda_k|^2}$. Thus, we have

$$\sum_{k=1}^{n}\left|\frac{\partial\theta'_k}{\partial\mathbf{A}_c}\right| = \sum_{k=1}^{n}\left|c_{k1}Re(\mathbf{y}_{k(2)}\mathbf{x}_{k(2)}^{\mathcal{H}}) - c_{k2}Im(\mathbf{y}_{k(2)}\mathbf{x}_{k(2)}^{\mathcal{H}})\right| \tag{42}$$

$$\sum_{k=1}^{n}\left|\frac{\partial\theta'_k}{\partial\mathbf{B}_c}\right| = \sum_{k=1}^{n}\left|\Big(c_{k1}Re(\mathbf{y}_{k(2)}\mathbf{x}_{k(1)}^{\mathcal{H}}) - c_{k2}Im(\mathbf{y}_{k(2)}\mathbf{x}_{k(1)}^{\mathcal{H}})\Big)\mathbf{M}_p^\top\right| \tag{43}$$

$$\sum_{k=1}^{n}\left|\frac{\partial\theta'_k}{\partial\mathbf{C}_c}\right| = \sum_{k=1}^{n}\left|\mathbf{B}_p^\top\Big(c_{k1}Re(\mathbf{y}_{k(1)}\mathbf{x}_{k(2)}^{\mathcal{H}}) - c_{k2}Im(\mathbf{y}_{k(1)}\mathbf{x}_{k(2)}^{\mathcal{H}})\Big)\right| \tag{44}$$

$$\sum_{k=1}^{n}\left|\frac{\partial\theta'_k}{\partial\mathbf{D}_c}\right| = \sum_{k=1}^{n}\left|\mathbf{B}_p^\top\Big(c_{k1}Re(\mathbf{y}_{k(1)}\mathbf{x}_{k(1)}^{\mathcal{H}}) - c_{k2}Im(\mathbf{y}_{k(1)}\mathbf{x}_{k(1)}^{\mathcal{H}})\Big)\mathbf{M}_p^\top\right|. \tag{45}$$

Define

$$\begin{aligned}\hat{\mathbf{Y}}_{(1)} &= \begin{bmatrix} c_{12}\mathbf{y}_{1(1)} & c_{22}\mathbf{y}_{2(1)} & \cdots & c_{n2}\mathbf{y}_{n(1)} \end{bmatrix} = \mathbf{Y}_{(1)}\mathbf{M}_{\Re} \\ \check{\mathbf{Y}}_{(1)} &= \begin{bmatrix} c_{11}\mathbf{y}_{1(1)} & c_{21}\mathbf{y}_{2(1)} & \cdots & c_{n1}\mathbf{y}_{n(1)} \end{bmatrix} = \mathbf{Y}_{(1)}\mathbf{M}_{\Im} \\ \hat{\mathbf{Y}}_{(2)} &= \begin{bmatrix} c_{12}\mathbf{y}_{1(2)} & c_{22}\mathbf{y}_{2(2)} & \cdots & c_{n2}\mathbf{y}_{n(2)} \end{bmatrix} = \mathbf{Y}_{(2)}\mathbf{M}_{\Re} \\ \check{\mathbf{Y}}_{(2)} &= \begin{bmatrix} c_{11}\mathbf{y}_{1(2)} & c_{21}\mathbf{y}_{2(2)} & \cdots & c_{n1}\mathbf{y}_{n(2)} \end{bmatrix} = \mathbf{Y}_{(2)}\mathbf{M}_{\Im}\end{aligned}$$

where $\mathbf{M}_{\Re} = diag\{\frac{Re(\lambda_1)}{|\lambda_1|^2}, \frac{Re(\lambda_2)}{|\lambda_2|^2}, \cdots, \frac{Re(\lambda_n)}{|\lambda_n|^2}\}$ and $\mathbf{M}_{\Im} = diag\{\frac{Im(\lambda_1)}{|\lambda_1|^2}, \frac{Im(\lambda_2)}{|\lambda_2|^2}, \cdots, \frac{Im(\lambda_n)}{|\lambda_n|^2}\}$. Then, we may obtain the upper bound of Θ' shown as follows

$$\bar{\Theta}'{=}\nu^2c_4^2\left\|\mathbf{X}_{(2)}\right\|_2^2\left\|\mathbf{Y}_{(2)}\right\|_2^2{+}c_1^2c_4^2\left\|\mathbf{Y}_{(2)}\right\|_F^2{+}c_5^2\left\|\mathbf{X}_{(2)}\right\|_F^2{+}c_6^2 \tag{46}$$

where $c_4{=}\left\|\left|\mathbf{M}_{\Re}\right|{+}\left|\mathbf{M}_{\Im}\right|\right\|_F$, $c_5{=}\left\|\left|\mathbf{B}_p^\top\check{\mathbf{Y}}_{(1)}\right|{+}\left|\mathbf{B}_p^\top\hat{\mathbf{Y}}_{(1)}\right|\right\|_F$, and $c_6{=}\left\|\Big(\left|\mathbf{B}_p^\top\check{\mathbf{Y}}_{(1)}\right| + \left|\mathbf{B}_p^\top\hat{\mathbf{Y}}_{(1)}\right|\Big)\left|\mathbf{X}_{(1)}^{\mathcal{H}}\mathbf{M}_p^\top\right|\right\|_F$, and they are closed-loop structure independence.

V. Optimal transformation for controller implementation

The problem of finding optimal transformation in the sense of the sensitivities of magnitude and supplement-angle of all eigenvalues can be stated from (40) and (46) in terms of the following minimization problem:

$$\Upsilon \triangleq \min_{\mathbf{T}_c}\big(\bar{\Gamma}\bar{\Theta}'\big) \tag{47}$$

Thus, it is expected that there exists an optimal similarity transformation, $\mathbf{T}_{c(opt)}$, in closed form, such that (47) is achieved, where

$$\mathbf{A}_{c(opt)}{=}\mathbf{T}_{c(opt)}^{-1}\mathbf{A}_c\mathbf{T}_{c(opt)}, \mathbf{B}_{c(opt)}{=}\mathbf{T}_{c(opt)}^{-1}\mathbf{B}_c, \\ \mathbf{C}_{c(opt)}{=}\mathbf{C}_c\mathbf{T}_{c(opt)}, \mathbf{D}_{c(opt)}{=}\mathbf{D}_c \tag{48}$$

for which

$$\mathbf{T}_{(opt)} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{c(opt)} \end{bmatrix}. \tag{49}$$

In the following, we provide methods for finding optimal transformations from the minima of the $\bar{\Gamma}$ and $\bar{\Theta}'$, respectively.

Theorem 2: : Let the similarity transformations for minimizing the $\bar{\Gamma}$ and $\bar{\Theta}'$ be expressed as $\mathbf{T}_{\bar{\Gamma}}$ and $\mathbf{T}_{\bar{\Theta}'}$, respectively. Then by Hermitian transform, we have

$$\text{(a) } \mathbf{T}_{\bar{\Gamma}} = \sqrt{\frac{c_2\sqrt{\nu}}{c_1\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\|_F}}\Big(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}}\Big)^{1/2}\mathbf{F}_1 \tag{50}$$

$$\text{(b) } \mathbf{T}_{\bar{\Theta}'} = \sqrt{\frac{c_5\sqrt{\nu}}{c_1c_4\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\|_F}}\Big(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}}\Big)^{1/2}\mathbf{F}_2 \tag{51}$$

to achieve the following minima

$$\min_{\Phi\in\Omega}(\bar{\Gamma}){=}\nu^2\left\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\right\|_2^2{+}2c_1c_2\sqrt{\nu}\left\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\right\|_F{+}c_3^2.$$

$$\min_{\Phi\in\Omega}(\bar{\Theta}'){=}\nu^2c_4^2\left\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\right\|_2^2{+}2c_1c_4c_5\sqrt{\nu}\left\|\mathbf{X}_{(2)}^{\mathcal{H}}\mathbf{Y}_{(2)}\right\|_F{+}c_6^2.$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are any arbitrary real orthogonal matrices.

Proof: See [9].

■

It is seen that, the optimal similarity transformation for achieving the minimum of (47) is a combination of $\mathbf{T}_{\bar{\Gamma}}$ and $\mathbf{T}_{\bar{\Theta}'}$ since the only difference between them is a constant term. Hence, it is reasonable to take the optimal similarity transformation for (47) as the following form:

$$\mathbf{T}_{c(opt)} = k_c\Big(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}}\Big)^{1/2}\mathbf{F}_3 \tag{52}$$

where k_c can be obtained by solving numerically the following equation

$$h(k_c) = 0 \tag{53}$$

where $h(k_c)$ is an eighth-order polynomial of k_c and given by

$$h(k_c){=}\tfrac{d(\bar{\Gamma}\bar{\Theta})}{dk_c}{=}2c_1^4c_4^2d_2^4k_c^8{+}(2\nu^2c_1^2c_4^2d_1^2d_2^2{+}c_1^2c_6^2d_2^2{+}c_1^2c_3^2c_4^2d_2^2)k_c^6 \\ -(\nu^2c_5^2d_1^2d_3^2{+}\nu^2c_2^2c_4^2d_1^2d_3^2{+}c_2^2c_6^2d_3^2{+}c_3^2c_5^2d_3^2)k_c^2{-}2c_2^2c_5^2d_3^4 \tag{54}$$

where $d_1 = \left\|(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}})^{1/2}\mathbf{Y}_{(2)}\right\|_2\left\|\mathbf{X}_{(2)}^{\mathcal{H}}(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}})^{-1/2}\right\|_2$, $d_2 = \left\|(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}})^{1/2}\mathbf{Y}_{(2)}\right\|_F$, and$d_3 = \left\|\mathbf{X}_{(2)}^{\mathcal{H}}(\mathbf{X}_{(2)}\mathbf{X}_{(2)}^{\mathcal{H}})^{-1/2}\right\|_F$.

Remark 2:: Since the coefficients of $h(k_c)$ with odd degrees are zeros, the coefficients of the ones with even degrees are positive except the fourth-degree one, and $h(0) < 0$, it has only one positive root which leads to the existence and uniqueness of the optimal transformation $\mathbf{T}_{c(opt)}$.

VI. Numerical example

In this example, a continuous linearized Vertical Take Off and Land (VTOL) aircraft controlled plant discretized with sampling period $T_s = 0.05sec.$ studied in [10] is give by

$$\mathbf{A}_p = \begin{bmatrix} 0.9982 & 0.0013 & 0.0004 & -0.0229 \\ 0.0023 & 0.9507 & -0.0048 & -0.1962 \\ 0.0049 & 0.0176 & 0.9670 & 0.0679 \\ 0.0001 & 0.0004 & 0.0492 & 1.0017 \end{bmatrix}, \mathbf{B}_p = \begin{bmatrix} 0.0221 & 0.0086 \\ 0.1733 & -0.3705 \\ -0.2697 & 0.2173 \\ -0.0068 & 0.0055 \end{bmatrix},$$

and $\mathbf{M}_p = [0\ 1\ 0\ 0]$. In this design example, we assign the poles to lie within the unit circle and set them as 0.9875, -0.6686, $0.8957 \pm j0.3126$, and $0.9892 \pm j0.0286$. Hence, the initial parameters implementation of the dynamical controller can be given by

$$\mathbf{A}_c = \begin{bmatrix} 0.9562 & 0.5568 \\ -0.1852 & -0.7852 \end{bmatrix}, \mathbf{B}_c = \begin{bmatrix} -0.0199 \\ 0.0200 \end{bmatrix},$$

$$\mathbf{C}_c = \begin{bmatrix} -5.3425 & 0.1729 \\ -20.9894 & 7.5324 \end{bmatrix}, \mathbf{D}_c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In what follows, we then compute the values of the *spectral radius* $r_s = 0.9896$, of the bounds $\|\varepsilon_{q_1'}\|_2 = 2.4495$ and $\|\varepsilon_{c_1'}\|_2 = 1.4142$, and of $\zeta = 22.4762$, respectively. Hence, by using (19), the stability guaranteed bit-number estimation is given by

$$W_{(est)} = 20. \tag{55}$$

For optimal controller realization design, we may get the following values: $\nu = 2$, $c_1 = 1.5349$, $c_2 = 0.6892$, $c_3 = 0.4845$, $c_4 = 3.0046$, $c_5 = 0.7604$, $d_1 = 1.2501$, $d_2 = 1.6912$, and $d_3 = 1.4142$. By using (53) and after substituting these values into (54), the following equation is obtained as

$$819.8206k_c^8 + 777.4798k_c^6 - 61.4861kc^2 - 2.1971 = 0, \tag{56}$$

and the solutions are given by $k_c = \pm j0.9212, \pm 0.5147, \pm j0.5730,$ and $\pm j0.1906$. For controller implementations, only this pair ± 0.5147 can be used, and since we have let k_c be positive, the optimal similarity transformation (52) (let $\mathbf{F}_3 = \mathbf{I}$) for controller implementations is given by

$$\mathbf{T}_{copt} = \begin{bmatrix} 0.0407 & -0.0365 \\ -0.0365 & 0.1406 \end{bmatrix}. \tag{57}$$

As well, the corresponding optimal controller realization can be given by

$$\mathbf{A}_{copt} = \begin{bmatrix} 0.7711 & 0.5288 \\ 0.3499 & -0.6001 \end{bmatrix}, \mathbf{B}_{copt} = \begin{bmatrix} -0.4702 \\ 0.0203 \end{bmatrix},$$

$$\mathbf{C}_{copt} = \begin{bmatrix} -0.2240 & 0.2191 \\ -1.1298 & 1.8242 \end{bmatrix}, \mathbf{D}_{copt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

By using this optimal controller implementation, the *spectral radius* r_s is indeed invariable. The quantization error and computational error bounds are given by $\|\varepsilon_{q_1'}\|_2 = 2.0055$ and $\|\varepsilon_{c_1}'\|_2 = 1.4142$, and value of ζ is 2.0071. Hence, the stability guaranteed bit-number estimation is obtained as

$$W_{(est)} = 12. \tag{58}$$

The step responses of the states and by using 3-bit controller implementations are shown in Fig. 4.

For comparing the differences of control performances between the controller implementations by means of $\mathbf{T}_{\bar{\Gamma}}$ studied in the previous work [9], the step responses of the states by using 3-bit controller implementations is shown in Fig. 3. In addition, the sensitivity measures with respect to original given and the optimal controllers can be seen in Table I.

VII. Conclusions

In this paper, we have proposed an efficient algorithm for analyzing the stability robustness of the closed-loop system for digital controller implementations using finite fixed-point arithmetic. Based on small gain theorem and Bellman Grownwall Lemma, a sufficient stability criterion for the closed-loop system is derived subject to finite fixed-point computation, and from which a least bit number is determined. The main contribution of this paper is that an analytically algebraic method for solving the optimal similarity transformation based on the mixed measure constituted by sensitivities of the magnitude and phase-supplement-angle of the eigenvalues. Numerical simulations show that the obtained least bit number used for digital controller implementations by the proposed algorithm is smaller than that of the original one which is not optimal.

Acknowledgment

This work is supported by Tatung University, Taipei, Taiwan, R.O.C. under project B9208-E01-019.

References

[1] J.H. Wilkinson, *Rounding Errors in Algebraic Processes*, Englewood Cliffs, NJ: Prentice Hall, 1963.

[2] B.-S. Chen and C.-T. Kuo, "Stability analysis of digital filters under finite word length effects," *IEE Proceedings*, vol. 136, Pt. G, no. 4, pp. 167-172, 1989.

[3] L.M. Smith and M.E. Henderson, Jr., "Roundoff noise reduction in cascade realizations of FIR digital filters," *IEEE Trans. Signal Processing*, vol. 48, pp. 1196-1200, Apr. 2000.

[4] T.H. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W.-S. Lu, "Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability grammians," *IEEE Trans. Circuits Syst. I*, vol. 49, no. 9, pp. 1279-1289, 2002.

[5] D. Williamson and K. Kadiman, "Optimal finite wordlength linear quadratic regulation," *IEEE Trans. Automat. Contr.*, vol. AC-34, no. 12, pp. 1218-1228, 1989.

[6] K. Liu, R.E. Skelton, and K. Grigoriadis, "Optimal controllers for finite wordlength implementation," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1294-1304, 1992.

[7] J. Wu, S. Chen, G. Li, R.H. Istepanian, and J. Chu, "Shift and delta operator realizations for digital controllers under finite word length considerations", *IEE Proc.-Control Theory Appl.*, vol. 147, no. 6, pp. 664-672, 2000.

[8] J. Wu, S. Chen, G. Li, R.H. Istepanian, and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations", *IEEE Trans. Automat. Contr.*, vol. 46, no. 7, pp. 1162-1166, 2001.

[9] Wen-Shyong Yu and Hsien-Ju Ko, "Improved Eigenvalue Sensitivity for Finite-Precision Digital Controller Realizations via Orthogonal Hermitian Transform," *IEE Proc.-Control Theory Appl.*, 2003, (to appear).

[10] R.E. Skelton, T. Iwasaki, and K. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*, Taylor and Francis, 1998.

TABLE I

Comparisons of the measures of the similarity transformation matrices for the Example.

Controller Structure	$\bar{\Gamma}$	$\bar{\Theta}'$	$\Upsilon = \bar{\Gamma}\bar{\Theta}'$	Γ	Θ'	$W_{(est)}$	$W_{(act.)}$
*	1.1629e+03	1.0497e+04	1.2207e+07	173.8927	692.1792	20	10
$\mathbf{T}_{\bar{\Gamma}}$	11.5457	82.7509	955.4153	4.0281	1.8599	13	4
$\mathbf{T}_{\bar{\Theta}'}$	14.3043	73.6056	1.0529e+03	3.4091	2.0636	13	4
$\mathbf{T}_{copt}$	11.8568	77.3095	916.6446	3.6881	1.7926	12	3

* Original state space system.

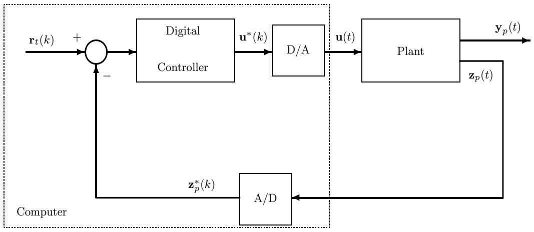

Fig. 1. Schematic diagram of a computer-controlled system.

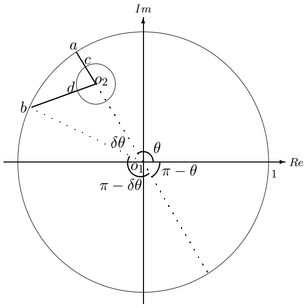

Fig. 2. The diagram of the perturbed magnitude/supplement-angle of the eigenvalue within the unite circle.

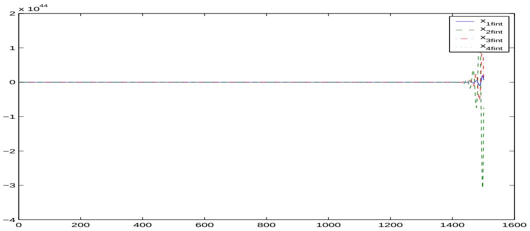

Fig. 3. The responses of the states with 3-bit word length using transformation $\mathbf{T}_{\bar{\Gamma}}$

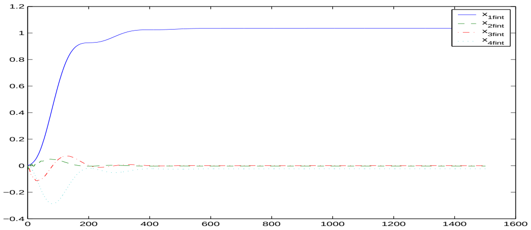

Fig. 4. The responses of the states with 3-bit word length using optimal similarity transformation $\mathbf{T}_{c(opt)}$.