Optimal Finite-Precision State-Estimate Feedback Controller Realizations of Discrete-Time Systems

Jun Wu, Sheng Chen, Gang Li, and Jian Chu

***Abstract*—This paper investigates the stability issue of a discrete-time control system, where a state-estimate feedback controller (SEFC), digitally implemented with a fixed-point format, is used. A tractable closed-loop stability related measure is derived with finite-word-length (FWL) implementation consideration of the controller. The optimal realizations of the SEFC are defined as those that maximize this measure and can be shown as the solutions of a nonlinear programming problem. A sophisticated optimization strategy is presented to provide an efficient method for solving this problem, and a numerical example is given to illustrate the design procedure.**

***Index Terms*—Finite word length, optimization, stability, state-estimate feedback controller.**

I. Introduction

The recent advances in digital control system design methods have led to a need for the efficient and accurate implementation of controllers with orders higher than that of the traditional PID controller. Although the number of 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 closed-loop stability under controller parameter perturbations is a critical issue in fixed-point implementations. It is well known that a designed, stable, closed-loop system may become unstable when the "infinite-precision" controller is implemented using a fixed-point processor due to finite-word-length (FWL) effects.

Many studies have investigated digital controller realizations with FWL considerations [1]–[5]. The first FWL stability measure was proposed in 1994 [3]. However, computing the value of this measure explicitly is still an unsolved open problem. Recently, two tractable FWL stability related measures have been derived, and the design procedures for searching for optimal FWL controller realizations have been developed [4]–[7]. In all of the above-mentioned works, controllers are output feedback controllers (OFC's). It is well known that there is another class of controllers, namely, state-estimate feedback controllers (SEFC's) [8]. The SEFC design is the product of a direct synthesis and design approach for linear control systems that combines modern state-space methods and observer theory. It also provides a unified formulation for single-input single-output and multi-input multi-output systems. The design of SEFC's is more transparent and simpler than the design of OFC's. Li and Gevers [9] studied the sensitivity and the roundoff noise gain of the closed-loop system transfer function with an FWL implemented SEFC. However, few studies to date investigate the effects of FWL implementation on the closed-loop stability for SEFC's.

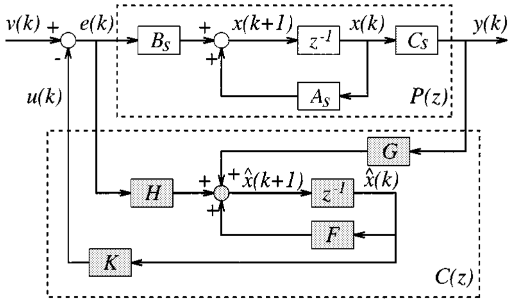

Fig. 1. Block diagram of the closed-loop system with state-estimate feedback controller.

This paper addresses the stability issues of FWL SEFC's. We derive a tractable measure that quantifies the "robustness" of the closed-loop stability under the controller parameter perturbations, and develop an optimization procedure to search for the optimal controller realizations that maximize the defined measure. The paper is organized as follows. Section II is devoted to formulating the problem to be dealt with and establishing necessary notations. A stability related measure that can be computed easily for a given SEFC realization is given in Section III. The optimal controller realization problem is also defined in this section. In Section IV, the optimization framework for obtaining the optimal FWL controller realization is presented. A numerical example is given in Section V to demonstrate the design procedure and the effectiveness of the proposed optimization method. Some concluding remarks are given in Section VI.

II. Notations and Problem Statement

Consider the discrete-time closed-loop system with an SEFC, as shown in Fig. 1. The discrete-time plant $P(z)$, which is assumed to be strictly proper, is represented as

$$\begin{cases} x(k+1) = A_s x(k) + B_s e(k) \\ y(k) = C_s\ x(k) \end{cases} \tag{1}$$

with $A_s \in \mathcal{R}^{n\times n}$, $B_s \in \mathcal{R}^{n\times p}$, and $C_s \in \mathcal{R}^{q\times n}$. The discrete-time SEFC $C(z)$ is given by

$$\begin{cases} \hat{x}(k+1) = F\hat{x}(k) + Gy(k) + He(k) \\ u(k) = K\hat{x}(k) \end{cases} \tag{2}$$

where $F \in \mathcal{R}^{n\times n}$, $H \in \mathcal{R}^{n\times p}$, the control gain $K \in \mathcal{R}^{p\times n}$, and the observer gain $G \in \mathcal{R}^{n\times q}$. The representation or realization for a given $C(z)$ is not unique. In fact, if (F_0, H_0, K_0, G_0) is a realization of $C(z)$, all the realizations of $C(z)$ form a realization set

$$S_C \triangleq \Big\{(F, H, K, G):\quad F = T^{-1}F_0T,\ H = T^{-1}H_0, \\ K = K_0T,\ G = T^{-1}G_0\Big\} \tag{3}$$

where T is any real-valued nonsingular matrix, called a similarity transformation. Define

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} \triangleq \begin{bmatrix} \mathrm{Vec}(F) \\ \mathrm{Vec}(H) \\ \mathrm{Vec}(K) \\ \mathrm{Vec}(G) \end{bmatrix}, \qquad \mathbf{w}_0 \triangleq \begin{bmatrix} \mathrm{Vec}(F_0) \\ \mathrm{Vec}(H_0) \\ \mathrm{Vec}(K_0) \\ \mathrm{Vec}(G_0) \end{bmatrix} \tag{4}$$

Manuscript received June 1, 1999; revised October 21, 1999. Recommended by Associate Editor, T. Chen. The work of S. Chen was supported in part by the UKEPSRC under Grant GR/M16894.

J. Wu and J. Chu are with the National Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, 310027, China.

S. Chen is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: sqc@ecs.soton.ac.uk).

G. Li is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Publisher Item Identifier S 0018-9286(00)04222-7.

where $\mathrm{Vec}(\cdot)$ denotes the column stacking operator and $N = n^2 + n(2p+q)$. Let $(\overline{A}, \overline{B}, \overline{C}, \overline{D})$ denote the state-space description of the closed-loop system. It is easy to see that

$$\begin{aligned}\overline{A}(\mathbf{w}) &= \begin{bmatrix} A_s & -B_s K \\ GC_s & F - HK \end{bmatrix} \\ &= \begin{bmatrix} I_n & 0_n \\ 0_n & T^{-1} \end{bmatrix} \overline{A}(\mathbf{w}_0) \begin{bmatrix} I_n & 0_n \\ 0_n & T \end{bmatrix}. \end{aligned} \tag{5}$$

Although different controller realizations yield different $\overline{A}$, the closed-loop poles or the eigenvalues of $\overline{A}$, denoted as $\{\lambda_i\}$, remain the same; i.e., $\lambda_i = \lambda_i(\overline{A}(\mathbf{w})) = \lambda_i(\overline{A}(\mathbf{w}_0))$, $i \in \{1, 2, \cdots, 2n\}$. In the controller design, (F, H, K, G) will have been chosen to make the closed-loop system stable and, therefore, $|\lambda_i| < 1, \forall i$.

When a controller realization $\mathbf{w}$ is implemented in fixed-point format, it is perturbed into

$$\mathbf{w} + \Delta\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} + \begin{bmatrix} \Delta w_1 \\ \vdots \\ \Delta w_N \end{bmatrix} \tag{6}$$

due to the FWL effects. Each element of $\Delta\mathbf{w}$ is bounded:

$$\mu(\Delta\mathbf{w}) \triangleq \max_{i\in\{1,\cdots,N\}} |\Delta w_i| \le \frac{\epsilon}{2}. \tag{7}$$

For a fixed-point processor of B_s bits

$$\epsilon = 2^{-(B_s - B_w)} \tag{8}$$

where B_w is an integer and 2^{B_w} is a "normalization" factor to make the absolute value of each element of $2^{-B_w}\mathbf{w}$ no larger than 1. With the perturbation $\Delta\mathbf{w}$, $\lambda_i(\overline{A}(\mathbf{w}))$ is moved to $\lambda_i(\overline{A}(\mathbf{w}+\Delta\mathbf{w}))$, which may be outside the unit circle. Thus, the closed-loop system designed to be stable may become unstable with an FWL-implemented controller realization $\mathbf{w}$.

It is, therefore, critical to know when the FWL error will cause the closed-loop system to become unstable. This means to compute the following stability measure [3]

$$\mu_0(\mathbf{w}) \triangleq \inf\left\{\mu(\Delta\mathbf{w})\colon \overline{A}(\mathbf{w}+\Delta\mathbf{w}) \text{ is unstable}\right\}. \tag{9}$$

The larger $\mu_0(\mathbf{w})$ is, the bigger FWL error the closed-loop stability can tolerate. Let $B_s^{\min}$ be the smallest word length that, when used to implement $\mathbf{w}$, can guarantee the closed-loop stability. Except in simulation, $B_s^{\min}$ is unknown. An estimate of $B_s^{\min}$ can be provided by

$$\hat{B}_{s0}^{\min} = \mathrm{Int}[-\log_2(\mu_0(\mathbf{w}))] - 1 + B_w \tag{10}$$

where $\mathrm{Int}[x]$ rounds x to the nearest integer with $\mathrm{Int}[x] \ge x$. From (7)–(10), it can be seen that the closed-loop system is stable when $\mathbf{w}$ is implemented with a fixed-point processor of at least $\hat{B}_{s0}^{\min}$ bits. Moreover, as the stability measure $\mu_0(\mathbf{w})$ is a function of the controller realization $\mathbf{w}$, we can search for an "optimal" realization that maximizes $\mu_0(\mathbf{w})$:

$$\mathbf{w}_{\mathrm{opt}} = \arg\max_{\mathbf{w}\in S_C} \mu_0(\mathbf{w}). \tag{11}$$

The difficulty with this approach is that computing explicitly the value of $\mu_0(\mathbf{w})$ is still an unsolved open problem. Thus, the stability measure $\mu_0(\mathbf{w})$ and the optimization procedure (11) have very limited practical value. An alternative measure that can not only quantify the FWL effects on stability robustness but can also be computed easily must be sought.

III. A Tractable Stability Related Measure

Roughly speaking, how easily the FWL error $\Delta\mathbf{w}$ can cause a stable control system to become unstable is determined by how close $\lambda_i(\overline{A}(\mathbf{w}))$ are to the unit circle and how sensitive they are to the controller parameter perturbations. Let us consider the following stability related measure:

$$\mu_1(\mathbf{w}) \triangleq \min_{i\in\{1,\cdots,2n\}} \frac{1 - |\lambda_i(\overline{A}(\mathbf{w}))|}{\displaystyle\sum_{j=1}^{N} \left| \frac{\partial\lambda_i}{\partial w_j}\bigg|_{\mathbf{w}} \right|} \tag{12}$$

where $1 - |\lambda_i(\overline{A}(\mathbf{w}))|$ is called the stability margin of the ith eigenvalue. Defining

$$\mathcal{P}(\mathbf{w}) \triangleq \left\{ \Delta\mathbf{w}\colon |\lambda_i(\overline{A}(\mathbf{w}+\Delta\mathbf{w}))| - |\lambda_i(\overline{A}(\mathbf{w}))| \le \mu(\Delta\mathbf{w}) \sum_{j=1}^{N} \left| \frac{\partial\lambda_i}{\partial w_j}\bigg|_{\mathbf{w}} \right|, \quad \forall i \right\} \tag{13}$$

we have the following proposition, the proof of which is straightforward.

Proposition 1: $\overline{A}(\mathbf{w} + \Delta\mathbf{w})$ is stable if $\Delta\mathbf{w} \in \mathcal{P}(\mathbf{w})$ and $\mu(\Delta\mathbf{w}) < \mu_1(\mathbf{w})$.

Remarks: The requirement for $\Delta\mathbf{w} \in \mathcal{P}(\mathbf{w})$ is not over-restricted. In practice, we will only be interested in those $\Delta\mathbf{w}$ that lie in the bounded region: $\mathcal{Q}(\mathbf{w}) \triangleq \{\Delta\mathbf{w}\colon \mu(\Delta\mathbf{w}) < \mu_0(\mathbf{w})\}$, i.e., those $\Delta\mathbf{w}$ that will not cause the closed-loop instability. Since $\partial\lambda_l/\partial w_j$ is continuous

$$\begin{aligned} &\lambda_l\big(\overline{A}(\mathbf{w}+\Delta\mathbf{w})\big) \\ &\quad = \lambda_l\big(\overline{A}(\mathbf{w})\big) + \sum_{j=1}^{N} \int_{\mathcal{C}} \frac{\partial\lambda_l}{\partial w_j}\, dw_j \\ &\quad = \lambda_l\big(\overline{A}(\mathbf{w})\big) + \sum_{j=1}^{N} \left(\mathrm{Re}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{a}_j} \right] + i\,\mathrm{Im}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{b}_j} \right] \right) \\ &\qquad \cdot \Delta w_j \end{aligned} \tag{14}$$

where $\mathcal{C}$ is the oriented segment from $\mathbf{w}$ to $\mathbf{w} + \Delta\mathbf{w}$, $\mathbf{a}_j$ and $\mathbf{b}_j$ are some points on $\mathcal{C}$. Hence

$$\begin{aligned} &\big|\lambda_l\big(\overline{A}(\mathbf{w}+\Delta\mathbf{w})\big)\big| - \big|\lambda_l\big(\overline{A}(\mathbf{w})\big)\big| \\ &\quad \le \left| \sum_{j=1}^{N} \left(\mathrm{Re}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{a}_j} \right] + i\,\mathrm{Im}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{b}_j} \right] \right) \Delta w_j \right|. \end{aligned} \tag{15}$$

Now, let us compare

$$\left| \sum_{j=1}^{N} \left(\mathrm{Re}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{a}_j} \right] + i\,\mathrm{Im}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{b}_j} \right] \right) \Delta w_j \right| \quad \text{with} \quad \mu(\Delta\mathbf{w}) \sum_{j=1}^{N} \left| \frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{w}} \right|. \tag{16}$$

Note that all of the N real-valued items $|\partial\lambda_l/\partial w_j|_{\mathbf{w}}|$ are in alignment; while the N complex-valued items

$$\left(\mathrm{Re}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{a}_j} \right] + i\,\mathrm{Im}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{b}_j} \right] \right) \Delta w_j$$

are generally out of alignment. Moreover, $|\Delta w_j| \le \mu(\Delta\mathbf{w})$, $\mathrm{Re}[\partial\lambda_l/\partial w_j]$ and $\mathrm{Im}[\partial\lambda_l/\partial w_j]$ are differentiable. Thus, a rather large positive κ exists such that $\forall\,\Delta\mathbf{w} \in \{\Delta\mathbf{w}\colon \mu(\Delta\mathbf{w}) \le \kappa\}$

$$\left| \sum_{j=1}^{N} \left(\mathrm{Re}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{a}_j} \right] + i\,\mathrm{Im}\left[\frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{b}_j} \right] \right) \Delta w_j \right| \le \mu(\Delta\mathbf{w}) \sum_{j=1}^{N} \left| \frac{\partial\lambda_l}{\partial w_j}\bigg|_{\mathbf{w}} \right|. \tag{17}$$

The above analysis shows that $\mathcal{P}(\mathbf{w})$ exists and at least a large part of $\mathcal{Q}(\mathbf{w})$ is covered by $\mathcal{P}(\mathbf{w})$.

Generally speaking, there is no rigorous relationship between $\mu_0(\mathbf{w})$ and $\mu_1(\mathbf{w})$, but $\mu_1(\mathbf{w})$ is connected with a lower bound of $\mu_0(\mathbf{w})$ in some manners. Define

$$\rho(\mathcal{P}(\mathbf{w})) \triangleq \inf_{\Delta\mathbf{w}\notin\mathcal{P}(\mathbf{w})} \mu(\Delta\mathbf{w}). \tag{18}$$

Proposition 2: $\mu_1(\mathbf{w}) \leq \mu_0(\mathbf{w})$ if $\rho(\mathcal{P}(\mathbf{w})) > \mu_0(\mathbf{w})$.

Proof: From the definition of $\rho(\mathcal{P}(\mathbf{w}))$ and the condition $\rho(\mathcal{P}(\mathbf{w})) > \mu_0(\mathbf{w})$, a $\Delta\mathbf{w} \in \mathcal{P}(\mathbf{w})$ exists such that $\mu_0(\mathbf{w}) = \mu(\Delta\mathbf{w})$ and $\overline{A}(\mathbf{w}+\Delta\mathbf{w})$ is unstable. It follows from Proposition 1 that $\mu_0(\mathbf{w}) \geq \mu_1(\mathbf{w})$.

From Proposition 2, it can be seen that $\mu_1(\mathbf{w})$ can be considered as a lower bound of $\mu_0(\mathbf{w})$, provided that $\mu_0(\mathbf{w})$ is small enough. The assumption of small $\mu_0(\mathbf{w})$ is not too restricted, as it does not make much sense to study the FWL effects on the closed-loop stability for those situations where the closed-loop systems have a very large stability robustness. Most digital control systems do have a small stability robustness, especially when fast sampling is applied.

To compute $\mu_1(\mathbf{w})$, one needs $\{\partial\lambda_i/\partial w_j\}$, which can be calculated with the following theorem. A proof of this theorem can be found in [4].

Theorem 1: Let $A = M_0 + M_1 X M_2 \in \mathcal{R}^{m\times m}$ be diagonalisable where $X \in \mathcal{R}^{l\times r}$, and M_0, M_1, and M_2 are independent of X with proper dimensions. Denote $\{\lambda_i\} = \{\lambda_i(A)\}$ as its eigenvalues. Let $\mathbf{x}_i$ be a right eigenvector of A corresponding to the eigenvalue λ_i. Denote $M_x = [\mathbf{x}_1\ \ \mathbf{x}_2\ \ ,\cdots,\ \ \mathbf{x}_m]$ and $M_y = [\mathbf{y}_1\ \ \mathbf{y}_2\ \ ,\cdots,\ \ \mathbf{y}_m] = M_x^{-\mathcal{H}}$, where $\mathcal{H}$ denotes the transpose and conjugate operation and $\mathbf{y}_i$ is called the reciprocal left eigenvector corresponding to λ_i. Then

$$\frac{\partial\lambda_i}{\partial X} = \begin{bmatrix} \frac{\partial\lambda_i}{\partial x_{11}} & ,\cdots, & \frac{\partial\lambda_i}{\partial x_{1r}} \\ \vdots & ,\cdots, & \vdots \\ \frac{\partial\lambda_i}{\partial x_{l1}} & ,\cdots, & \frac{\partial\lambda_i}{\partial x_{lr}} \end{bmatrix} = M_1^{\mathcal{T}} \mathbf{y}_i^* \mathbf{x}_i^{\mathcal{T}} M_2^{\mathcal{T}} \tag{19}$$

where the superscript $\mathcal{T}$ denotes the transpose operator and $*$ the conjugate operation.

Without confusion, we use $\mathbf{x}_i$ and $\mathbf{y}_i$ to denote the right and reciprocal left eigenvectors related to $\lambda_i(\overline{A}(\mathbf{w}))$, respectively. $\overline{A}(\mathbf{w})$ can be arranged in the following equivalent forms:

$$\overline{A}(\mathbf{w}) = \begin{bmatrix} A_s & -B_sK \\ GC_s & -HK \end{bmatrix} + \begin{bmatrix} 0_n \\ I_n \end{bmatrix} F [\,0_n \quad I_n\,] \tag{20}$$

$$\overline{A}(\mathbf{w}) = \begin{bmatrix} A_s & -B_sK \\ GC_s & F \end{bmatrix} + \begin{bmatrix} 0_n \\ I_n \end{bmatrix} H [\,0_n \quad -K\,] \tag{21}$$

$$\overline{A}(\mathbf{w}) = \begin{bmatrix} A_s & 0_n \\ GC_s & F \end{bmatrix} + \begin{bmatrix} -B_s \\ -H \end{bmatrix} K [\,0_n \quad I_n\,] \tag{22}$$

$$\overline{A}(\mathbf{w}) = \begin{bmatrix} A_s & -B_sK \\ 0_n & F-HK \end{bmatrix} + \begin{bmatrix} 0_n \\ I_n \end{bmatrix} G [\,C_s \quad 0_n\,]. \tag{23}$$

Applying Theorem 1, we obtain

$$\frac{\partial\lambda_i}{\partial F} = [\,0_n \quad I_n\,] \mathbf{y}_i^* \mathbf{x}_i^{\mathcal{T}} \begin{bmatrix} 0_n \\ I_n \end{bmatrix} \tag{24}$$

$$\frac{\partial\lambda_i}{\partial H} = [\,0_n \quad I_n\,] \mathbf{y}_i^* \mathbf{x}_i^{\mathcal{T}} \begin{bmatrix} 0_n \\ -K^{\mathcal{T}} \end{bmatrix} \tag{25}$$

$$\frac{\partial\lambda_i}{\partial K} = [\,-B_s^{\mathcal{T}} \quad -H^{\mathcal{T}}\,] \mathbf{y}_i^* \mathbf{x}_i^{\mathcal{T}} \begin{bmatrix} 0_n \\ I_n \end{bmatrix} \tag{26}$$

$$\frac{\partial\lambda_i}{\partial G} = [\,0_n \quad I_n\,] \mathbf{y}_i^* \mathbf{x}_i^{\mathcal{T}} \begin{bmatrix} C_s^{\mathcal{T}} \\ 0_n \end{bmatrix}. \tag{27}$$

For a complex-valued matrix $M \in \mathcal{C}^{l\times r}$ with elements m_{ij}, define a norm of M as

$$\|M\|_S \triangleq \sum_{i=1}^{l} \sum_{j=1}^{r} |m_{ij}|. \tag{28}$$

Then

$$\begin{aligned} \sum_{j=1}^{N} \left| \frac{\partial\lambda_i}{\partial w_j} \right| &= \left\| \frac{\partial\lambda_i}{\partial \mathbf{w}} \right\|_S \\ &= \left\| \frac{\partial\lambda_i}{\partial F} \right\|_S + \left\| \frac{\partial\lambda_i}{\partial H} \right\|_S + \left\| \frac{\partial\lambda_i}{\partial K} \right\|_S + \left\| \frac{\partial\lambda_i}{\partial G} \right\|_S . \end{aligned} \tag{29}$$

For a given controller realization $\mathbf{w}$, the smallest word length $B_s^{\min}$ can be estimated with $\mu_1(\mathbf{w})$ using the following:

$$\hat{B}_{s1}^{\min} = \mathrm{Int}[-\log_2(\mu_1(\mathbf{w}))] - 1 + B_w. \tag{30}$$

More importantly, as $\mu_1(\mathbf{w})$ is tractable, one can estimate the optimal controller realizations defined in (11) with

$$\hat{\mathbf{w}}_{\mathrm{opt}} = \arg \max_{\mathbf{w}\in S_C} \mu_1(\mathbf{w}) \tag{31}$$

which will be discussed in the next section.

IV. Optimization Procedure

With the computationally tractable stability related measure $\mu_1(\mathbf{w})$, we now present a practical optimization procedure to search for an optimal controller realization $\hat{\mathbf{w}}_{\mathrm{opt}}$ defined in (31).[1] Assume that an initial controller realization $(F_0,\ H_0,\ K_0,\ G_0)$ has been provided. For example, the observer gain G_0 is obtained using some observer design method with given observer poles, the state-feedback gain K_0 is obtained using some state-feedback design method with given closed-loop poles, $F_0 = A_s - G_0C_s$ and $H_0 = B_s$. Let $\{\lambda_{0i},\ i = 1,\cdots,2n\}$ be the eigenvalues of $\overline{A}(\mathbf{w}_0)$, and $\mathbf{x}_{0i}$ and $\mathbf{y}_{0i}$ be the right and reciprocal left eigenvectors corresponding to λ_{0i}, respectively. Partition $\mathbf{x}_{0i}$ and $\mathbf{y}_{0i}$ into

$$\mathbf{x}_{0i} = \begin{bmatrix} \mathbf{x}_{0i}(1) \\ \mathbf{x}_{0i}(2) \end{bmatrix}, \qquad \mathbf{x}_{0i}(1),\ \mathbf{x}_{0i}(2) \in \mathcal{C}^n \tag{32}$$

and

$$\mathbf{y}_{0i} = \begin{bmatrix} \mathbf{y}_{0i}(1) \\ \mathbf{y}_{0i}(2) \end{bmatrix}, \qquad \mathbf{y}_{0i}(1),\ \mathbf{y}_{0i}(2) \in \mathcal{C}^n \tag{33}$$

respectively. Let $\mathbf{w}$ be a controller realization transformed from $\mathbf{w}_0$ with T. It is easy to see from (5) that

$$\mathbf{x}_i = \begin{bmatrix} I_n & 0_n \\ 0_n & T^{-1} \end{bmatrix} \mathbf{x}_{0i} = \begin{bmatrix} \mathbf{x}_{0i}(1) \\ T^{-1}\mathbf{x}_{0i}(2) \end{bmatrix} \tag{34}$$

is a right eigenvector of $\overline{A}(\mathbf{w})$ corresponding to the same eigenvalue, and

$$\mathbf{y}_i = \begin{bmatrix} I_n & 0_n \\ 0_n & T^{\mathcal{T}} \end{bmatrix} \mathbf{y}_{0i} = \begin{bmatrix} \mathbf{y}_{0i}(1) \\ T^{\mathcal{T}}\mathbf{y}_{0i}(2) \end{bmatrix} \tag{35}$$

is the corresponding reciprocal left eigenvector. Substituting (34) and (35) into (24)–(27) yields

$$\frac{\partial\lambda_i}{\partial F} = T^{\mathcal{T}} \mathbf{y}_{0i}^*(2) \mathbf{x}_{0i}^{\mathcal{T}}(2) T^{-\mathcal{T}} \tag{36}$$

$$\frac{\partial\lambda_i}{\partial H} = -\,T^{\mathcal{T}} \mathbf{y}_{0i}^*(2) \mathbf{x}_{0i}^{\mathcal{T}}(2) K_0^{\mathcal{T}} \tag{37}$$

$$\frac{\partial\lambda_i}{\partial K} = -\left(B_s^{\mathcal{T}} \mathbf{y}_{0i}^*(1) + H_0^{\mathcal{T}} \mathbf{y}_{0i}^*(2) \right) \mathbf{x}_{0i}^{\mathcal{T}}(2) T^{-\mathcal{T}} \tag{38}$$

[1] In the sequel, by an *optimal realization* we mean a solution to (31) rather than (11). The latter, as mentioned before, is not tractable.

$$\frac{\partial\lambda_i}{\partial G} = T^{\mathcal{T}}\mathbf{y}_{0i}^*(2)\mathbf{x}_{0i}^{\mathcal{T}}(1)C_s^{\mathcal{T}}. \tag{39}$$

We can define (31) in an alternative way:

$$\begin{aligned}\nu &= \frac{1}{\max\limits_{\mathbf{w}\in S_C}\mu_1(\mathbf{w})}\\ &= \min_{\substack{T\in\mathcal{R}^{n\times n}\\ \det(T)\neq 0}}\ \max_{i\in\{1,\,\cdots,\,2n\}}\ \frac{\sum\limits_{j=1}^{N}\left|\dfrac{\partial\lambda_i}{\partial w_j}\right|_{\mathbf{w}}}{1-|\lambda_{0i}|}\\ &= \min_{\substack{T\in\mathcal{R}^{n\times n}\\ \det(T)\neq 0}}\ \max_{i\in\{1,\,\cdots,\,2n\}}\\ &\quad\cdot\frac{\left\|\dfrac{\partial\lambda_i}{\partial F}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial H}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial K}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial G}\right\|_S}{1-|\lambda_{0i}|}\end{aligned} \tag{40}$$

which means that finding an optimal realization of the SEFC is equivalent to obtaining a similarity transformation that is a solution to the following nonlinear optimization problem:

$$T_{\text{opt}} = \arg\min_{\substack{T\in\mathcal{R}^{n\times n}\\ \det(T)\neq 0}} f(T) \tag{41}$$

with the cost function

$$f(T) \triangleq \max_{i\in\{1,\,\cdots,\,2n\}} \frac{\left\|\dfrac{\partial\lambda_i}{\partial F}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial H}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial K}\right\|_S+\left\|\dfrac{\partial\lambda_i}{\partial G}\right\|_S}{1-|\lambda_{0i}|}. \tag{42}$$

To find a T_{opt}, we will adopt an iterative optimization procedure to generate a sequence $\{T_0, T_1, T_2, \cdots\}$, which converges to T_{opt}.

The optimization (41) is constrained. Define $\Omega \triangleq \{T \in \mathcal{R}^{n\times n}: \det(T) = 0\}$. As Ω is only a manifold in $\mathcal{R}^{n\times n}$, starting from a $T_0 \notin \Omega$, it is rare for an iterative sequence $\{T_i\}$ to move into Ω. Thus, in the iterative procedure, the constraint $\det(T) \neq 0$ can practically be ignored, leading to an unconstrained optimization problem:

$$\tilde{\nu} = \min_{T\in\mathcal{R}^{n\times n}} f(T). \tag{43}$$

The possible pitfall of violating the constraint can readily be avoided by the following measure. As the inverse of T is required in the computation of $f(T)$, it is obtained using the singular value (SV) decomposition. If an SV of T is too small, T is almost singular and a small perturbation ηI_n is added to T so that $T + \eta I_n \notin \Omega$. This small perturbation, which is rarely needed, will not affect the convergence of the iterative procedure.

Because the cost function $f(T)$ is nonsmooth and nonconvex, optimization must be based on a direct search without the aid of cost function derivatives. The conventional optimization methods for this kind of problem, such as Rosenbrock and Simplex algorithms [10], [11], generally can only find a local minimum. Although the choice of initial realization will not affect the closed-loop eigenvalues, the eigenvalue sensitivities $\partial\lambda_i/\partial\mathbf{w}$, $\forall i$ depend on the chosen initial realization. Thus, for different $\mathbf{w}_0$, the shape of the cost function $f(T)$ will change, giving rise to a different degree of difficulty in the optimization procedure. It is therefore important to use an efficient, and preferably global, optimization method. We adopt a global optimization strategy based on the adaptive simulated annealing (ASA) [12], [13] to search for a true global optimum $\hat{\mathbf{w}}_{\text{opt}}$.

V. Illustrative Example

This section presents a numerical example to illustrate the design procedure and how the proposed optimization approach can be used effectively to search for the optimal FWL realization of SEFC's. This example was taken from [9]. The discrete-time plant $P(z)$ was given by

$$\begin{aligned}A_s &= \begin{bmatrix} 2.758\,200e+0 & -2.534\,177e+0 & 7.755\,853e-1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\\ B_s &= \begin{bmatrix}1\\0\\0\end{bmatrix}\\ C_s &= [2.200\,000e-3 \quad 4.400\,000e-3 \quad 2.200\,000e-3].\end{aligned}$$

The initial realization of the controller $C(z)$ was chosen to be

$$\begin{aligned}F_0 &= \begin{bmatrix} 2.497\,941e+0 & -3.054\,695e+0 & 5.153\,264e-1\\ 7.776\,040e-1 & -4.447\,920e-1 & -2.223\,960e-1\\ -1.801\,490e-1 & 6.397\,019e-1 & -1.801\,490e-1\end{bmatrix}\\ H_0 &= \begin{bmatrix}1\\0\\0\end{bmatrix}\\ K_0 &= [4.761\,000e-1 \quad -8.183\,439e-1 \quad 3.505\,623e-1]\\ G_0 &= \begin{bmatrix}1.182\,995e+2\\ 1.010\,891e+2\\ 8.188\,593e+1\end{bmatrix}.\end{aligned}$$

The corresponding transition matrix $\overline{A}(\mathbf{w}_0)$ was formed, from which the poles and eigenvectors $\{\lambda_{0j}, \mathbf{x}_{0j}, \mathbf{y}_{0j}, j = 1, \cdots, 6\}$ of the ideal closed-loop system were computed.

The ASA algorithm was used to search for an T_{opt} by solving the optimization problem (41), and it produced the following solution:

$$T_{\text{opt}} = \begin{bmatrix} -2.492\,226e+2 & -8.436\,334e+1 & 2.500\,780e+2\\ -1.712\,397e+2 & -6.278\,793e+1 & 2.126\,909e+2\\ -9.225\,780e+1 & -3.503\,457e+1 & 1.704\,995e+2\end{bmatrix}.$$

This gave rise to the optimal FWL controller realization $\hat{\mathbf{w}}_{\text{opt}}$:

$$\begin{aligned}F_{\text{opt}} &= \begin{bmatrix} 7.273\,562e-1 & -1.063\,087e-1 & 1.577\,498e-1\\ -1.906\,839e-1 & 6.591\,385e-1 & 2.312\,892e-1\\ 7.272\,022e-2 & -3.150\,339e-2 & 4.865\,053e-1\end{bmatrix}\\ H_{\text{opt}} &= \begin{bmatrix}-5.926\,328e-2\\ 1.743\,758e-1\\ 3.763\,549e-3\end{bmatrix}\\ K_{\text{opt}} &= [-1.086\,402e+1 \quad -1.065\,065e+0 \quad 4.778\,530e+0]\\ G_{\text{opt}} &= \begin{bmatrix}-1.558\,451e-3\\ 6.013\,747e-2\\ 4.917\,847e-1\end{bmatrix}.\end{aligned}$$

For the initial and optimal controller realizations, we exploit the true minimal bit lengths $B_s^{\min}$ using the following computer simulation method. Let initial bit length be enough big, e.g., $B_s = 100$. Rounding (F, H, K, G) to B_s bits, we obtain the B_s-bits representation (F_r, H_r, K_r, G_r) and then check the stability of the closed-loop system composed of (A_s, B_s, C_s) and (F_r, H_r, K_r, G_r), i.e., observe whether the closed-loop poles are in the open unit disk. Reduce B_s by 1 and repeat rounding and checking until there appears to be closed-loop instability at B_u bits. Then, $B_s^{\min} = B_u + 1$. Table I compares the values of the stability related measure, estimated minimum bit lengths, and true

TABLE I
COMPARISON OF STABILITY RELATED MEASURES, ESTIMATED MINIMUM BIT LENGTHS AND TRUE MINIMUM BIT LENGTHS FOR THE INITIAL AND OPTIMAL CONTROLLER REALIZATIONS

realization	μ_1	$\hat{B}_{s1}^{\min}$	$B_s^{\min}$
initial $\mathbf{w}_0$	1.995885e-5	22	15
optimal $\hat{\mathbf{w}}_{\text{opt}}$	6.019238e-4	14	7

minimum bit lengths for the initial and optimal controller realizations. It can be seen that, for this example, the optimization achieved an improvement by a factor of 30 on the closed-loop stability related measure and an 8-bit reduction in the required minimum bit length.

VI. CONCLUSIONS

In this paper, we have presented an approach to address the stability issues of the closed-loop discrete-time system where a state-estimate feedback controller is implemented with a fixed-point processor. An FWL closed-loop stability related measure has been derived, which is computationally tractable. As this measure is a function of the controller realization; the optimal realization problem of state-estimate feedback controllers is to find a realization that maximizes this measure. It has been shown that this optimal realization problem can be interpreted as a nonlinear programming problem. An efficient global optimization strategy based on the ASA algorithm has been adopted to solve this nonsmooth and nonconvex optimization problem.

REFERENCES

[1] P. Moroney, A. S. Willsky, and P. K. Houpt, "The digital implementation of control compensators: The coefficient wordlength issue," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 621–630, Aug. 1980.

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

[3] I. J. Fialho and T. T. Georgiou, "On stability and performance of sampled data systems subject to word length constraint," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2476–2481, Dec. 1994.

[4] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 689–693, 1998.

[5] R. H. Istepanian, G. Li, J. Wu, and J. Chu, "Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds," *Proc. Inst. Elect. Eng. Contr. Th. Applicat.*, vol. 145, no. 5, pp. 472–478, 1998.

[6] S. Chen, J. Wu, R. H. Istepanian, and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 2149–2153, Nov. 1999.

[7] R. H. Istepanian, J. Wu, J. F. Whidborne, J. Yan, and S. E. Salcudean, "Finite-word-length stability issues of teleoperation motion-scaling control system," in *Proc. UKACC Contr.'98*, Swansea, UK, Sept. 1–4, 1998, pp. 1676–1681.

[8] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[9] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1487–1498, 1990.

[10] G. S. G. Beveridge and R. S. Schechter, *Optimization: Theory and Practice*. New York: McGraw-Hill, 1970.

[11] L. C. W. Dixon, *Nonlinear Optimization*. London: English Universities Press, 1972.

[12] L. Ingber, "Simulated annealing: Practice versus theory," *Math. Comput. Model.*, vol. 18, no. 11, pp. 29–57, 1993.

[13] S. Chen and B. L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Process.*, vol. 79, no. 1, pp. 117–128, 1999.

Practical Stability and Stabilization

Luc Moreau and Dirk Aeyels

***Abstract*—We present a practical stability result for dynamical systems depending on a small parameter. This result is applied to a practical stability analysis of fast time-varying systems studied in averaging theory, and of highly oscillatory systems studied by Sussmann and Liu. Furthermore, the problem of practically stabilizing control affine systems with drift is discussed.**

***Index Terms*—Approximation methods, Lie algebras, stability, time-varying systems.**

I. INTRODUCTION

In the present note, dynamical systems that depend on a small parameter are studied from the viewpoint of continuity of solutions.

Consider a system that depends on a small parameter $\varepsilon > 0$

$$\dot{x} = f^{\varepsilon}(t, x) \tag{1}$$

and a system

$$\dot{x} = g(t, x) \tag{2}$$

with the assumption that trajectories of (1) converge—uniformly on compact time intervals—to trajectories of (2) as $\varepsilon \downarrow 0$.

A particular example is given by fast time-varying systems studied in averaging theory

$$\dot{x} = f\left(\frac{t}{\varepsilon}, x\right). \tag{3}$$

It is well known that, under appropriate technical conditions, there exists an associated averaged system

$$\dot{x} = f_{\text{av}}(x) \tag{4}$$

such that trajectories of (3) converge—uniformly on compact time intervals—to trajectories of (4) as $\varepsilon \downarrow 0$.

Teel *et al.* [1] have proven that, under appropriate technical conditions, if the origin of the averaged system (4) is a globally asymptotically stable equilibrium point, then the fast time-varying system (3) is practically stable. Their proof is based on advanced Lyapunov techniques.

In the present note, it is recognized that this practical stability result is of a topological nature, that it is a consequence of the convergence property of solutions: we prove the general result that, under appropriate technical conditions, if the origin of system (2) is a globally uniformly asymptotically stable equilibrium point, then system (1) is practically stable. This approach provides an alternative proof for the practical stability result [1] mentioned above, and extends it to a larger class of systems: it is not only applicable to fast time-varying systems as in averaging theory, but also, for example, to highly oscillatory systems studied by Sussmann and Liu [2]. This latter application is useful for control purposes. Indeed, it leads to a practical stabilization algorithm for a class of control affine systems with drift.

An outline of this note is as follows. Section II introduces some notations and hypotheses. Section III introduces a notion of practical

Manuscript received April 8, 1999; revised November 29, 1999. Recommended by Associate Editor, W. Lin. The work of L. Moreau was supported by BOF Grant 011D0696 of the Ghent University.

The authors are with the SYSTeMS Group, Ghent University, Technologiepark 9, 9052 Zwijnaarde, Belgium (e-mail: lmoreau@ensmain.rug.ac.be).

Publisher Item Identifier S 0018-9286(00)06078-5.