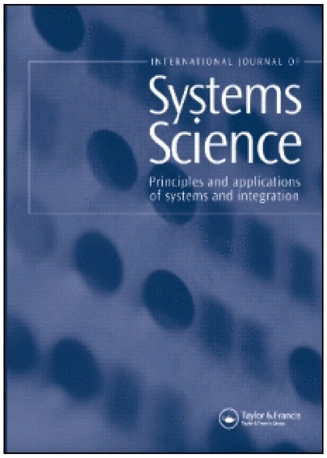


International Journal of Systems Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713697751

Optimal finite-precision controller realization of sampled-data systems

R. H. Istepanian; S. Chen; J. F. Whidborne

Online Publication Date: 01 April 2000
To cite this Article: Istepanian, R. H., Chen, S. and Whidborne, J. F. (2000) 'Optimal finite-precision controller realization of sampled-data systems', International Journal of Systems Science, 31:4, 429 - 438
To link to this article: DOI: 10.1080/002077200291019
URL: http://dx.doi.org/10.1080/002077200291019

International Journal of Systems Science, 2000, volume 31, number 4, pages 429–438

Optimal finite-precision controller realization of sampled-data systems

R. H. ISTEPANIAN†, S. CHEN‡, J. WU§ and J. F. WHIDBORNE¶

We investigate the sensitivity of closed-loop stability with respect to finite word length (FWL) effects in the implementation of digital controller coefficients. The optimal realization of digital controller structures with finite precision consideration is formulated as the solution of a constrained nonlinear optimization problem. A sophisticated optimization strategy involving the adaptive simulated annealing (ASA) optimizer is developed to provide an efficient computational method for searching the optimal FWL controller realization with maximum stability bound and minimum bit length requirement. A numerical simulation example is presented to illustrate the effectiveness of the proposed strategy.

1. Introduction

The recent advances in advanced control design methods means that there is a need for the efficient and accurate implementation of controllers with a higher order than traditional PID controllers. 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, particularly for mass market applications in the automotive and consumer electronics sectors. Thus, the consideration of FWL effects is an important issue in modern industrial digital control applications.

The FWL effects in digital signal processing have been extensively studied over the last two decades (Roberts and Mullis 1987). More recent studies have addressed the FWL effects and parameterization issues on digital controller realizations and relevant applications (Gevers and Li 1993, Madievski *et al.* 1995, Istepanian *et al.* 1996, 1998 b, Istepanian 1997). However, few studies to date address the closed-loop stability issues and the relevant effects of the finite-precision controller realizations (Moroney *et al.* 1980, Fialho and Georgiou 1994). For such systems, the formulation of a finite-precision controller structure with a too small word length may result in the loss of closed-loop system stability due to the well-known FWL effects. This is an interesting problem within the FWL parameterization framework that has not been studied extensively.

An earlier FWL stability measure, from which a minimum bit length that guarantees the closed-loop stability can be estimated for a digital controller realization, was proposed by Moroney *et al.* (1980). However, computing this measure explicitly seems numerically very difficult and is still an unsolved open problem. Recently, a tractable FWL stability measure, which is a lower bound of the stability measure given by Moroney *et al.* (1980), has been derived and design procedures that guarantee the stability of the resulting optimal FWL controller have been developed (Li 1998). An enhanced tractable FWL stability measure, which provides a better lower bound than the one given by Li (1998), has been introduced (Istepanian *et al.* 1998 a). Based on this enhanced lower bound of the stability measure, an efficient optimization procedure has been formulated to derive the optimal PID

Received 14 September 1998. Revised 12 April 1999.

† Department of Electrical and Computer Engineering, Ryerson Polytechnic University, 350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3; e-mail: ristepan@ee.ryerson.ca

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

§ National Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou, 310027, People's Republic of China.

¶ Department of Mechanical Engineering, King's College London, Strand, London WC2R 2LS, UK.

International Journal of Systems Science ISSN 0020–7721 print/ISSN 1464–5319 online © 2000 Taylor & Francis Ltd
http://www.tandf.co.uk/journals/tf/00207721.html

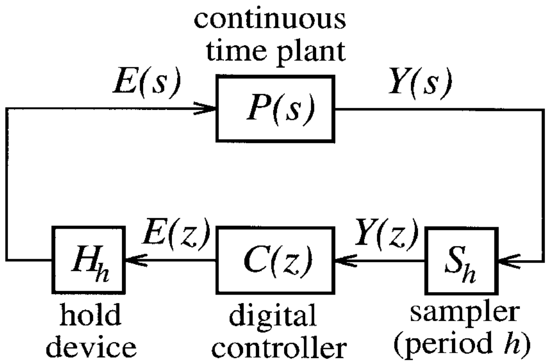


Figure 1. Sampled-data system with digital controller realization.

controller realization with maximum stability bound and minimum bit length (Chen *et al.* 2000).

In this paper, we extend the framework presented by Chen *et al.* (2000) and refine the theoretical concepts to generalize its applicability to any digital controller structures for sampled-data control systems. The problem is formulated as a constrained nonlinear optimization problem. As the cost function is non-smooth and non-convex, an efficient global optimization strategy based on the ASA (Ingber and Rosen 1992, Ingber 1996, Rosen 1997, Chen *et al.* 1998) is developed to search for the optimal FWL controller realization. The paper is organized as follows. In section 2, we present the problem formulation and establish notations and definitions of FWL stability measures. In section 3, we apply the results of section 2 to formulate the optimization framework for obtaining the optimal FWL controller realizations with maximum stability bounds and minimum bit length requirements. The detailed optimization procedure is presented in section 4, and a numerical example is given in section 5 to illustrate the effectiveness of the proposed approach. The paper ends with conclusions in section 6.

2. FWL stability measures

In this section, we introduce the FWL stability measures derived by Moroney *et al.* (1980), Li (1998), Istepanian *et al.* (1998 a), and present some refinements, which will provide the basis for the derivation of the optimal FWL controller realization. Consider the sampled-data control system shown in figure 1, where $P(s)$ is the continuous-time linear time-invariant plant, $C(z)$ is the discrete-time linear shift-invariant controller, S_h is the sampler with sampling period h, and H_h is the hold device. The outputs of the sampler and hold device are given by

$$Y(z) = S_h Y(s): \qquad y(k) = y(t)|_{t=kh} \qquad (1)$$

and

$$E(s) = H_h E(z): \qquad e(t) = e(k), kh < t \le (k+1)h, \quad (2)$$

respectively. Assume that $P(s)$ is strictly proper. Let $(A_p, B_p, C_p, 0)$ be a state-space description of $P(s)$, that is, $P(s) = C_p(sI - A_p)^{-1}B_p$, where $A_p \in \mathcal{R}^{m\times m}$, $B_p \in \mathcal{R}^{m\times l}$ and $C_p \in \mathcal{R}^{q\times m}$. Let (A_c, B_c, C_c, D_c) be a state-space description of $C(z)$, that is, $C(z) = C_c(zI - A_c)^{-1}B_c + D_c$, where $A_c \in \mathcal{R}^{n\times n}$, $B_c \in \mathcal{R}^{n\times q}$, $C_c \in \mathcal{R}^{l\times n}$ and $D_c \in \mathcal{R}^{l\times q}$. We will refer to (A_c, B_c, C_c, D_c) as a realization of $C(z)$. The realizations of $C(z)$ are not unique. In fact, if (A_c, B_c, C_c, D_c) is a realization of $C(z)$, so is $(\mathcal{T}^{-1}A_c\mathcal{T}, \mathcal{T}^{-1}B_c, C_c\mathcal{T}, D_c)$ for any similarity transformation $\mathcal{T} \in \mathcal{R}^{n\times n}$.

Considering the behaviour of the sampled-data system at its sampling instants, we obtain a discrete-time feedback system:

$$\left.\begin{aligned} Y(z) &= S_h P(s) H_h E(z) \\ E(z) &= C(z) Y(z) \end{aligned}\right\}. \qquad (3)$$

The plant $P(z) = S_h P(s) H_h$ is the discretized $P(s)$, and $P(z) = C_z(zI - A_z)^{-1}B_z$ has a state-space description $(A_z, B_z, C_z, 0)$, where

$$A_z = e^{A_p h} \in \mathcal{R}^{m\times m},$$

$$B_z = \int_0^h e^{A_p\tau} B_p \, d\tau \in \mathcal{R}^{m\times l} \quad \text{and} \quad C_z = C_p \in \mathcal{R}^{q\times m}. \quad (4)$$

It can easily be seen that the corresponding state-space description $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ of the discrete-time closed-loop system is given by:

$$\bar{A} = \begin{bmatrix} A_z + B_z D_c C_z & B_z C_c \\ B_c C_z & A_c \end{bmatrix} = \begin{bmatrix} A_z & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B_z & 0 \\ 0 & I_n \end{bmatrix} \times \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix} \begin{bmatrix} C_z & 0 \\ 0 & I_n \end{bmatrix}$$

$$= M_0 + M_1 X M_2 = \bar{A}(X), \qquad (5)$$

$$\bar{B} = \begin{bmatrix} B_z \\ 0 \end{bmatrix}, \qquad \bar{C} = [C_z \quad 0], \quad \bar{D} = 0, \qquad (6)$$

where $M_0 \in \mathcal{R}^{(m+n)\times(m+n)}$, $M_1 \in \mathcal{R}^{(m+n)\times(l+n)}$ and $M_2 \in \mathcal{R}^{(q+n)\times(m+n)}$ are some fixed matrices that depend on $P(s)$ and h, I_n denotes the $n \times n$ identity matrix, and

$$X = \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & \cdots & p_{q+n} \\ p_{q+n+1} & p_{q+n+2} & \cdots & p_{2(q+n)} \\ \vdots & \vdots & \cdots & \vdots \\ p_{(l+n-1)(q+n)+1} & p_{(l+n-1)(q+n)+2} & \cdots & p_{(l+n)(q+n)} \end{bmatrix} \tag{7}$$

will be referred to as the controller matrix.

Suppose that $C(z)$ has been given to make the sampled-data system stable and the realization of $C(z)$ is X. Since the sampled-data system is stable if and only if the system (3) is stable (Chen and Francis 1991), it follows that the eigenvalues of $\bar{A}(X)$, denoted by $\{\lambda_i, 1 \le i \le m+n\}$, satisfy $|\lambda_i| < 1$, $\forall i \in \{1, \ldots, m+n\}$. When the realization (A_c, B_c, C_c, D_c) of $C(z)$ is implemented in finite-precision format, the controller matrix X is perturbed to $X + \Delta X$, where

$$\Delta X = \begin{bmatrix} \Delta p_1 & \Delta p_2 & \cdots & \Delta p_{q+n} \\ \Delta p_{q+n+1} & \Delta p_{q+n+2} & \cdots & \Delta p_{2(q+n)} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta p_{(l+n-1)(q+n)+1} & \Delta p_{(l+n-1)(q+n)+2} & \cdots & \Delta p_N \end{bmatrix} \tag{8}$$

and $N = (l+n)(q+n)$. Due to the FWL effects, each element of ΔX is bounded, that is,

$$\mu(\Delta X) \triangleq \max_{i \in \{1,\ldots,N\}} |\Delta p_i| \le \frac{\epsilon}{2}. \tag{9}$$

For a fixed-point processor of B_s bits

$$\epsilon = 2^{-(B_s - B_X)}, \tag{10}$$

where B_X is an integer and 2^{B_X} is a 'normalization' factor such that the absolute value of each element of $2^{-B_X}X$ is not larger than 1. With the perturbation ΔX, λ_i is moved to $\tilde{\lambda}_i$. The closed-loop system is unstable if and only if there exists $i \in \{1, \ldots, m+n\}$ such that $|\tilde{\lambda}_i| \ge 1$.

To see when the round-off error will cause the closed-loop system to become unstable, Moroney *et al.* (1980) defined an FWL stability measure as:

$$\mu_0(X) \triangleq \inf\{\mu(\Delta X) : \bar{A}(X) + M_1 \Delta X M_2 \text{ is unstable}\}. \tag{11}$$

How 'robust' a controller realization is to the FWL effects can also be viewed from a different angle. Let $B_s^{\min}$ be the smallest word length that can guarantee the closed-loop stability. It would be highly desirable to know $B_s^{\min}$ for a given controller realization. However, except in simulation, it is impractical to test the closed-loop system by reducing B_s until it becomes unstable. Based on $\mu_0(X)$, an estimate of $B_s^{\min}$ is given by

$$\hat{B}_{s0}^{\min} = \mathrm{Int}[-\log_2(\mu_0(X)] - 1 + B_X, \tag{12}$$

where $\mathrm{Int}[x]$ rounds x to the nearest integer and $\mathrm{Int}[x] \ge x$. From (9) to (12), we know that the sampled-data closed-loop system is stable when X is implemented with a fixed-point processor of at least $\hat{B}_{s0}^{\min}$ bits. The problem with this FWL stability measure is that computing explicitly the value of $\mu_0(X)$ is still an unsolved open problem. Thus, the stability measure $\mu_0(X)$ has very limited practical value.

To overcome the difficulty in the computation of $\mu_0(X)$, Istepanian *et al.* (1998 a) introduced an FWL stability measure as:

$$\mu_1(X) \triangleq \min_{i \in \{1,\ldots,m+n\}} \frac{1 - |\lambda_i|}{\sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j} \Big|_X \right|}. \tag{13}$$

We have the following proposition.

Proposition 1: *$\bar{A}(X + \Delta X)$ is stable if $\mu(\Delta X) < \mu_1(X)$.*

Proof: When ΔX is small, using a first-order approximation we assume

$$\Delta \lambda_i = \tilde{\lambda}_i - \lambda_i = \sum_{j=1}^{N} \frac{\partial \lambda_i}{\partial p_j}\Big|_X \Delta p_j, \quad 1 \le i \le m+n, \tag{14}$$

where $\tilde{\lambda}_i$ are the eigenvalues of $\bar{A}(X + \Delta X)$. It follows that

$$|\Delta \lambda_i| \le \sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right| |\Delta p_j| \le \mu(\Delta X) \sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right|. \tag{15}$$

Thus, for $1 \le i \le m+n$, if

$$\mu(\Delta X) < \frac{1 - |\lambda_i|}{\sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right|}, \tag{16}$$

we have

$$|\tilde{\lambda}_i| \le |\lambda_i| + |\Delta \lambda_i| \le |\lambda_i| + \mu(\Delta X) \sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right| < |\lambda_i| + \frac{1 - |\lambda_i|}{\sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right|} \sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\Big|_X \right| = 1, \tag{17}$$

which means that $\bar{A}(X + \Delta X)$ is stable. □

The assumption that the controller coefficient perturbations are very small is generally valid. For example, standard fixed-point processors have 16 bits. Assuming $B_X = 4$, then the controller parameter errors are bounded by 2^{-13}. Notice that $\mu_1(X)$ is a lower bound of $\mu_0(X)$. The proof is straightforward. Define the set

$$\mathcal{P}_X \triangleq \{\Delta X : \bar{A}(X+\Delta X) \text{ is unstable}\}. \tag{18}$$

For any $\Delta X \in \mathcal{P}_X$, we must have $\mu_1(X) \le \mu(\Delta X)$; otherwise, according to Proposition 1, $\bar{A}(X+\Delta X)$ is stable, which is a contradiction. By the definition of $\mu_0(X)$, it follows that $\mu_1(X) \le \mu_0(X)$.

Unlike $\mu_0(X)$, $\mu_1(X)$ is a tractable stability measure as it can be computed easily using the following lemma. The proof of this lemma is given by Istepanian *et al.* (1998 a).

Lemma 1: *Let $\bar{A}(X) = M_0 + M_1 X M_2$ be diagonalizable and have $\{\lambda_i, i = 1,\dots,m+n\}$ as its eigenvalues, and $\mathbf{x}_i$ be a right eigenvector of $\bar{A}(X)$ corresponding to the eigenvalue λ_i. Denote $M_x = [\mathbf{x}_1 \cdots \mathbf{x}_{m+n}]$ and $M_y = [\mathbf{y}_1 \cdots \mathbf{y}_{m+n}] = M_x^{-\mathrm{H}}$, where $\mathbf{y}_i$ is called the reciprocal left eigenvector corresponding to λ_i, and $^{\mathrm{H}}$ denotes the transpose and conjugate operation. Then $\forall i \in \{1,\dots,m+n\}$*

$$\frac{\partial \lambda_i}{\partial X} = \begin{bmatrix} \dfrac{\partial \lambda_i}{\partial p_1} & \dfrac{\partial \lambda_i}{\partial p_2} & \cdots & \dfrac{\partial \lambda_i}{\partial p_{q+n}} \\ \dfrac{\partial \lambda_i}{\partial p_{q+n+1}} & \dfrac{\partial \lambda_i}{\partial p_{q+n+2}} & \cdots & \dfrac{\partial \lambda_i}{\partial p_{2(q+n)}} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial \lambda_i}{\partial p_{(l+n-1)(q+n)+1}} & \dfrac{\partial \lambda_i}{\partial p_{(l+n-1)(q+n)+2}} & \cdots & \dfrac{\partial \lambda_i}{\partial p_N} \end{bmatrix}$$

$$= M_1^{\mathrm{T}} \mathbf{y}_i^* \mathbf{x}_i^{\mathrm{T}} M_2^{\mathrm{T}}, \tag{19}$$

where $^{\mathrm{T}}$ denotes the transpose operation, and * the conjugate operation.

When a designed infinite-precision stable controller X is implemented with a fixed-point processor of B_s bits, the norm of the controller perturbation $\mu(\Delta X)$ and the lower-bound stability measure $\mu_1(X)$ can be evaluated. According to Proposition 1, if $\mu_1(X) > \mu(\Delta X)$, the closed-loop stability is maintained. Furthermore, from (9) and (10), it is easily seen that the closed-loop system is stable if $\mu_1(X) > (2^{-(B_s - B_X)})/2$. Define

$$\hat{B}_{s1}^{\min} = \mathrm{Int}\,[-\log_2(\mu_1(X)] - 1 + B_X. \tag{20}$$

It can be used as an estimate of $B_s^{\min}$. Obviously, as $\hat{B}_{s1}^{\min} \ge \hat{B}_{s0}^{\min} \ge B_s^{\min}$, $\hat{B}_{s1}^{\min}$ is a more conservative estimate of $B_s^{\min}$ than $\hat{B}_{s0}^{\min}$. However, $\hat{B}_{s0}^{\min}$ is impractical to obtain.

Another tractable FWL stability measure, introduced by Li (1998), is defined as

$$\mu_2(X) \triangleq \min_{i\in\{1,\dots,m+n\}} \frac{1-|\lambda_i|}{\sqrt{N \sum_{j=1}^{N} \left| \dfrac{\partial \lambda_i}{\partial p_j}\right|_X \right|^2}}. \tag{21}$$

It is also a lower bound of $\mu_0(X)$. Similarly, an estimate $\hat{B}_{s2}^{\min}$ of $B_s^{\min}$ can be computed based on $\mu_2(X)$. Since

$$\left(\sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\right|_X \right| \right)^2 \le N \sum_{j=1}^{N} \left| \frac{\partial \lambda_i}{\partial p_j}\right|_X \right|^2, \tag{22}$$

where the equality only holds under a very restricted condition, we have $\mu_2(X) \le \mu_1(X) \le \mu_0(X)$. It follows that $\hat{B}_{s2}^{\min} \ge \hat{B}_{s1}^{\min}$. Thus, $\mu_1(X)$, which is closer to $\mu_0(X)$, is a better FWL stability measure and can provide a better estimate of $B_s^{\min}$.

3. Optimal FWL controller realization

Since $\mu_1(X)$ is a better tractable FWL stability measure than $\mu_2(X)$, we will use it as the basis for the derivation of the optimal FWL controller realization problem. It is known that there are different realizations X for a given $C(z)$, and the stability measure $\mu_1(X)$ is a function of the realization. It is of practical importance to find a realization that maximizes $\mu_1(X)$. Such a realization is optimal in the sense that it has maximum closed-loop stability robustness to FWL effects. The digital controller implemented with an optimal realization can guarantee the stability of the closed-loop system with a minimum hardware requirement in terms of word length.

To start the optimal design procedure, it is assumed that an initial realization of $C(z)$,

$$X_0 = \begin{bmatrix} D_c^0 & C_c^0 \\ B_c^0 & A_c^0 \end{bmatrix}, \tag{23}$$

is available. Any realization of $C(z)$ can be expressed as:

$$X_{\mathcal{T}} \triangleq \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} X_0 \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T} \end{bmatrix}, \tag{24}$$

where $\mathcal{T} \in \mathcal{R}^{n\times n}$ and $\det(\mathcal{T}) \neq 0$. From (5), the closed-loop transition matrix is

$$\bar{A}(X_{\mathcal{T}}) = \begin{bmatrix} A_z & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B_z & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} X_0 \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T} \end{bmatrix} \times \begin{bmatrix} C_z & 0 \\ 0 & I_n \end{bmatrix}$$
$$= \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} \begin{bmatrix} A_z & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T} \end{bmatrix} + \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} \begin{bmatrix} B_z & 0 \\ 0 & I_n \end{bmatrix} X_0 \begin{bmatrix} C_z & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T} \end{bmatrix}$$
$$= \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} \bar{A}(X_0) \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T} \end{bmatrix}. \quad (25)$$

Notice that $\bar{A}(X_{\mathcal{T}})$ has the same eigenvalues as $\bar{A}(X_0)$. Let λ_i^0 be the ith eigenvalue of $\bar{A}(X_0)$, and $\mathbf{x}_i^0$ and $\mathbf{y}_i^0$ be the corresponding right and reciprocal left eigenvectors, respectively. It is easily seen from (25) that the ith right and reciprocal left eigenvectors of $\bar{A}(X_{\mathcal{T}})$ are

$$\begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{-1} \end{bmatrix} \mathbf{x}_i^0 \in \mathcal{C}^{m+n} \quad \text{and} \quad \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} \mathbf{y}_i^0 \in \mathcal{C}^{m+n}, \quad (26)$$

respectively. Applying Lemma 1, we have

$$\left.\frac{\partial \lambda_i}{\partial X}\right|_{X=X_{\mathcal{T}}} = \begin{bmatrix} B_z^{\mathrm{T}} & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} (\mathbf{y}_i^0)^* (\mathbf{x}_i^0)^{\mathrm{T}} \times \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{T}^{-\mathrm{T}} \end{bmatrix} \begin{bmatrix} C_z^{\mathrm{T}} & 0 \\ 0 & I_n \end{bmatrix}$$
$$= \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} B_z^{\mathrm{T}} & 0 \\ 0 & I_n \end{bmatrix} (\mathbf{y}_i^0)^* (\mathbf{x}_i^0)^{\mathrm{T}} \begin{bmatrix} C_z^{\mathrm{T}} & 0 \\ 0 & I_n \end{bmatrix} \times \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T}^{-\mathrm{T}} \end{bmatrix}$$
$$= \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} \left.\frac{\partial \lambda_i}{\partial X}\right|_{X=X_0} \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T}^{-\mathrm{T}} \end{bmatrix}. \quad (27)$$

We can describe the optimal FWL realization problem of digital controllers by the following maximization problem:

$$\varphi \triangleq \max_{X_{\mathcal{T}}} \mu_1(X_{\mathcal{T}}) = \max_{X_{\mathcal{T}}} \min_{1 \le i \le m+n} \frac{1 - |\lambda_i^0|}{\sum_{j=1}^{N} \left| \left.\frac{\partial \lambda_i}{\partial p_j}\right|_{X=X_{\mathcal{T}}} \right|}. \quad (28)$$

For the complex-valued matrix $M \in \mathcal{C}^{(n+l)\times(n+q)}$ with elements $M_{i,j}$, denote

$$\|M\|_s \triangleq \sum_{i=1}^{n+l} \sum_{j=1}^{n+q} |M_{i,j}|. \quad (29)$$

The optimization problem (28) is equivalent to the minimization problem

$$\nu = \frac{1}{\varphi} \triangleq \min_{X_{\mathcal{T}}} \max_{1 \le i \le m+n} \frac{\left\| \left.\frac{\partial \lambda_i}{\partial X}\right|_{X=X_{\mathcal{T}}} \right\|_s}{1 - |\lambda_i^0|}$$
$$= \min_{\substack{\mathcal{T} \in \mathcal{R}^{n\times n} \\ \det(\mathcal{T}) \neq 0}} \max_{1 \le i \le m+n} \left\| \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} \Phi_i \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T}^{-\mathrm{T}} \end{bmatrix} \right\|_s, \quad (30)$$

where

$$\Phi_i = \frac{\left.\frac{\partial \lambda_i}{\partial X}\right|_{X=X_0}}{1 - |\lambda_i^0|}, \quad 1 \le i \le m+n \quad (31)$$

are the fixed eigenvalue sensitivity matrices depending only on the initial realization. Define the cost function

$$f(\mathcal{T}) = \max_{1 \le i \le m+n} \left\| \begin{bmatrix} I_l & 0 \\ 0 & \mathcal{T}^{\mathrm{T}} \end{bmatrix} \Phi_i \begin{bmatrix} I_q & 0 \\ 0 & \mathcal{T}^{-\mathrm{T}} \end{bmatrix} \right\|_s. \quad (32)$$

The optimal FWL controller realization problem is presented as

$$\nu = \min_{\substack{\mathcal{T} \in \mathcal{R}^{n\times n} \\ \det(\mathcal{T}) \neq 0}} f(\mathcal{T}). \quad (33)$$

The above problem is a constrained nonlinear optimization problem. Because the cost function (32) is non-smooth and non-convex, 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 (Kowalik and Osborne 1968, Beveridge and Schechter 1970, Dixon 1972), in general can only find a local minimum. We adopt a global optimization strategy based on the ASA (Ingber and Rosen 1992, Ingber 1996, Rosen 1997, Chen *et al.* 1998) to search for a global optimal solution.

4. The ASA optimization procedure

The ASA is an efficient scheme for solving the following general optimization problem:

$$\min_{\mathbf{w} \in \mathcal{W}} J(\mathbf{w}), \quad (34)$$

where $\mathbf{w} = [w_1 \cdots w_{n_d}]^{\mathrm{T}}$ is the n_d-dimensional parameter vector to be optimized,

$$\mathcal{W} \triangleq \{\mathbf{w} \in \mathcal{R}^{n_d} : L_i \le w_i \le U_i, 1 \le i \le n_d\} \quad (35)$$

is the feasible set of $\mathbf{w}$, L_i and U_i are the lower and upper bounds of w_i, respectively. The cost function $J(\mathbf{w})$ can be multimodal and non-smooth.

4.1. *Search guiding mechanisms*

The ASA belongs to a class of so-called guided random search methods. It evolves a solution $\mathbf{w}$ in the state space $\mathcal{W}$ with the search mechanisms that imitate the random behaviour of molecules during the annealing process. The seemingly random search is guided by certain underlying probability distributions. Specifically, the ASA algorithm is described by three functions (Rosen 1997).

(1) Generating probability density function:

$$G(w_i^{\text{old}}, w_i^{\text{new}}, V_{i,\text{gen}}; 1 \le i \le n_d). \qquad (36)$$

This determines how a new state $\mathbf{w}^{\text{new}}$ is created, and from what neighbourhood and probability distributions it is generated, given the current state $\mathbf{w}^{\text{old}}$. The generating 'temperatures' $V_{i,\text{gen}}$ describe the widths or scales of the generating distribution along each dimension w_i of the state space.

Often a cost function has different sensitivities along different dimensions of the state space. Ideally, the generating distribution used to search a steeper and more sensitive dimension should have a narrower width than that of the distribution used in searching a dimension less sensitive to change. The ASA adopts a so-called re-annealing scheme to re-scale $V_{i,\text{gen}}$ periodically, so that they optimally adapt to the current status of the cost function. This is an important mechanism, which not only speeds up the search process but also makes the optimization process robust to different problems.

(2) Acceptance function:

$$P_{\text{accept}}(J(\mathbf{w}^{\text{old}}), J(\mathbf{w}^{\text{new}}), V_{\text{accept}}). \qquad (37)$$

This gives the probability of $\mathbf{w}^{\text{new}}$ being accepted. The acceptance temperature determines the frequency of accepting new states of poorer quality.

The probability of acceptance is very high at very high temperature V_{accept}, and it becomes smaller as V_{accept} is reduced. At every acceptance temperature, there is a finite probability of accepting the new state. This produces occasionally an uphill move, enables the algorithm to escape from local minima, and allows a more effective search of the state space to find a global minimum. The ASA also periodically adapts V_{accept} to best suit the status of the cost function. This helps to improve convergence speed and robustness.

(3) Reducing temperatures or annealing schedule:

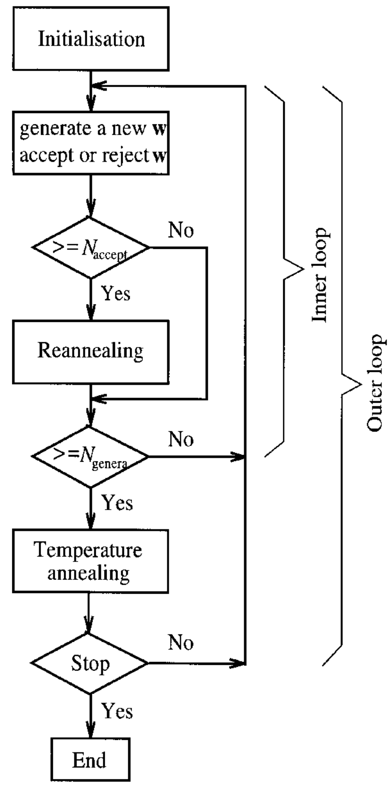


Figure 2. Flow chart of the adaptive simulated annealing.

$$\left.\begin{aligned} V_{\text{accept}}(k_a) &\rightarrow V_{\text{accept}}(k_a+1) \\ V_{i,\text{gen}}(k_i) &\rightarrow V_{i,\text{gen}}(k_i+1), 1 \le i \le n_d \end{aligned}\right\}, \qquad (38)$$

where k_a and k_i are some annealing time indexes. The reduction of temperatures should be sufficiently gradual in order to ensure that the algorithm finds a global minimum.

This mechanism is based on the observations of the physical annealing process. When the metal is cooled from a high temperature, if the cooling is sufficiently slow, the atoms line themselves up and form a crystal, which is the state of minimum energy in the system. The slow convergence of many standard simulated annealing algorithms is rooted in this slow annealing process. The ASA, however, can employ a very fast annealing schedule, as it has the self-adaptation ability to re-scale temperatures.

4.2. *Algorithm implementation*

An implementation of the ASA is illustrated in figure 2. How the ASA realizes the above three functions will become clear in the following detailed description.

(i) In the initialization, an initial $\mathbf{w} \in \mathcal{W}$ is randomly generated, the initial temperature of the acceptance probability function, $V_{\text{accept}}(0)$, is set to $J(\mathbf{w})$, and the initial temperatures of the parameter generating probability functions, $V_{i,\text{gen}}(0)$, $1 \le i \le n_d$, are set to 1.0. A user-defined annealing control parameter c is given, and the annealing times, k_i for $1 \le i \le n_d$ and k_a, are all set to 0.0.

(ii) The algorithm generates a new point in the parameter space with:

$$w_i^{\text{new}} = w_i^{\text{old}} + q_i(U_i - L_i) \qquad \text{for } 1 \le i \le n_d \text{ and } \mathbf{w}^{\text{new}} \in \mathcal{W}, \quad (39)$$

where q_i is calculated as

$$q_i = \text{sgn}\,(v_i - \tfrac{1}{2}) V_{i,\text{gen}}(k_i) \left(\left(1 + \frac{1}{V_{i,\text{gen}(K_i)}}\right)^{|2v_i - 1|} - 1 \right), \quad (40)$$

and v_i is a uniformly distributed random variable in $[0, 1]$. Notice that if a generated $\mathbf{w}^{\text{new}} \notin \mathcal{W}$, it is simply discarded and a new point is tried again until $\mathbf{w}^{\text{new}} \in \mathcal{W}$. The value of the cost function $J(\mathbf{w}^{\text{new}})$ is then evaluated and the acceptance probability function of $\mathbf{w}^{\text{new}}$ is given by

$$P_{\text{accept}} = \frac{1}{1 + \exp\left(\left(J(\mathbf{w}^{\text{new}}) - J(\mathbf{w}^{\text{old}})\right)/V_{\text{accept}}(k_a)\right)}. \quad (41)$$

A uniform random variable P_{unif} is generated in $[0,\ 1]$. If $P_{\text{unif}} \le P_{\text{accept}}$, $\mathbf{w}^{\text{new}}$ is accepted; otherwise it is rejected.

(iii) After every N_{accept} acceptance points, re-annealing takes place by first calculating the sensitivities

$$s_i = \left| \frac{J(\mathbf{w}^{\text{best}} + \mathbf{e}_i\delta) - J(\mathbf{w}^{\text{best}})}{\delta} \right|, \quad 1 \le i \le n_d, \quad (42)$$

where $\mathbf{w}^{\text{best}}$ is the best point found so far, δ is a small step size, the n_d-dimensional vector $\mathbf{e}_i$ has unit ith element and the rest of elements of $\mathbf{e}_i$ are all zeros. Let $s_{\max} = \max\{s_i, 1 \le i \le n_d\}$. Each parameter generating temperature $V_{i,\text{gen}}$ is scaled by a factor $s_{\max}/s_i$ and the annealing time k_i is reset

$$\left.\begin{aligned} V_{i,\text{gen}}(k_i) &= \frac{s_{\max}}{s_i} V_{i,\text{gen}}(k_i), \\ k_i &= \left(-\frac{1}{c} \log\left(\frac{V_{i,\text{gen}}(k_i)}{V_{i,\text{gen}}(0)} \right) \right)^{n_d} \end{aligned}\right\}. \quad (43)$$

Similarly, $V_{\text{accept}}(0)$ is reset to the value of the last accepted cost function, $V_{\text{accept}}(k_a)$ is reset to $J(\mathbf{w}^{\text{best}})$ and the annealing time k_a is rescaled accordingly

$$k_a = \left(-\frac{1}{c} \log\left(\frac{V_{\text{accept}}(k_a)}{V_{\text{accept}}(0)} \right) \right)^{n_d}. \quad (44)$$

(iv) After every N_{genera} generated points, annealing takes place with

$$\left.\begin{aligned} k_i &= k_i + 1 \\ V_{i,\text{gen}}(k_i) &= V_{i,\text{gen}}(0) \exp\left(-c k_i^{\frac{1}{n_d}}\right) \end{aligned}\right\} 1 \le i \le n_d \quad (45)$$

and

$$\left.\begin{aligned} k_a &= k_a + 1 \\ V_{\text{accept}}(k_a) &= V_{\text{accept}}(0) \exp\left(-c k_a^{\frac{1}{n_d}}\right) \end{aligned}\right\}; \quad (46)$$

Otherwise, go to Step (ii).

(v) The algorithm is terminated if the parameters have remained unchanged for a few successive re-annealings or a preset maximum number of cost function evaluations has been reached; otherwise, go to Step (ii).

The ASA contains two loops. The inner loop ensures that the parameter space is searched sufficiently at a given temperature, which is necessary to guarantee that the algorithm finds a global optimum (Aarts and Korst 1989). The ASA uses only the value of the cost function in the optimization process and is very simple to programme. It is statistically guaranteed to converge to a global optimum (Geman and Geman 1984).

4.3. *Algorithm parameter tuning*

Most of the ASA algorithm parameters are automatically set and 'tuned', and the user only needs to assign an annealing rate control parameter c and set two values N_{accept} and N_{genera}. Obviously, the optimal values of N_{accept} and N_{genera} are problem dependent, but our experience suggests that an adequate choice for N_{accept} is in the range of tens to hundreds and an appropriate value for N_{genera} is in the range of hundreds to thousands. The annealing control parameter c can be determined from the chosen initial temperature, final temperature and predetermined number of annealing steps (Ingber and Rosen 1992, Ingber 1996). We have found out that a choice of c in the range 1.0 to 30.0 is often adequate. It should be emphasized that, as the ASA has excellent self-adaptation ability, the performance of the algorithm is not critically influenced by the specific chosen values of c, N_{accept} and N_{genera}.

4.4. *Dealing with constraint and choice of initial realization*

The optimization (33) is constrained. Define $\Omega \triangleq \{\mathcal{T} \in \mathcal{R}^{n\times n} : \det(\mathcal{T}) = 0\}$. As Ω is only a manifold in $\mathcal{R}^{n\times n}$, starting from $\mathcal{T}_0 \notin \Omega$, it is rare for an iterative sequence $\{\mathcal{T}_i\}$ to move into Ω. Thus, in the iterative procedure, the constraint $\det(\mathcal{T}) \neq 0$ could be ignored. This would reduce the optimization problem (33) to an unconstrained one:

$$\tilde{\nu} = \min_{\mathcal{T}\in\mathcal{R}^{n\times n}} f(\mathcal{T}). \qquad (47)$$

The possible pitfall of violating the constraint however remains, which may result in an invalid solution. To guarantee $\det(\mathcal{T}) \neq 0$, we notice that $\mathcal{T}^{-1}$ is required in the computation of $f(\mathcal{T})$. We calculate the inverse of $\mathcal{T}$ using the singular value (SV) decomposition. If an SV of $\mathcal{T}$ is too small, $\mathcal{T}$ is near singular and we add a small perturbation ηI_n to $\mathcal{T}$ such that $\mathcal{T} + \eta I_n \notin \Omega$. This small perturbation will guarantee $\{\mathcal{T}_i\}$ to be well conditioned but will not affect the convergence of the iterative procedure.

Basically, any available state-space realization of $C(z)$ can be used as the initial realization X_0. A convenient choice of X_0 is the controllable canonical realization. It is well known that the canonical realization can become ill-conditioned at fast sampling conditions. In such situations, non-canonical realizations can be used as the initial realization. For example, X_0 can be chosen to be the realization obtained by direct discretizing a state-space description of the continuous-time controller $C(s)$. Notice that the choice of initial realization will not affect the closed-loop eigenvalues, but the eigenvalue sensitivity matrices Φ_i depend on the chosen initial realization. Thus, for different X_0 the shape of the cost function $f(\mathcal{T})$ will change, giving rise to a different degree of difficulty to optimize. It is therefore important to use an efficient and preferably global optimization method.

5. Illustrative numerical example

This section presents a numerical example to illustrate how the optimization approach presented in the previous sections can be used effectively to design the optimal FWL realization of digital controller structures. This example was cited by Chen and Francis (1991) and used by Li (1998) for studying the corresponding FWL stability measure $\mu_2(X)$. The continuous-time plant model was given by

$$P(s) = \frac{1.6188s^2 - 0.1575s - 43.9425}{(s^4 + 0.1736s^3 + 27.9001s^2 + 0.0186s)(s+1)}, \qquad (48)$$

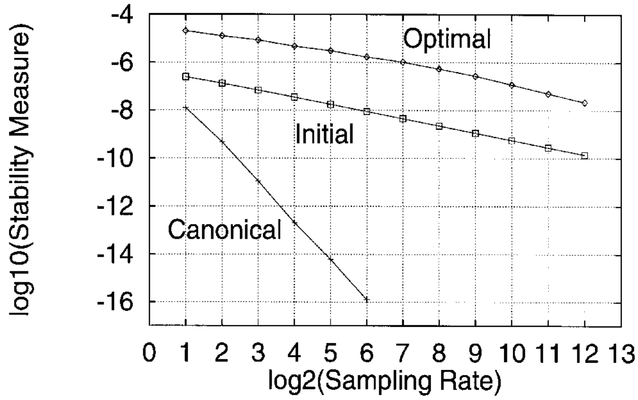


Figure 3. FWL stability measure μ_1 as a function of sampling rate for different controller realizations.

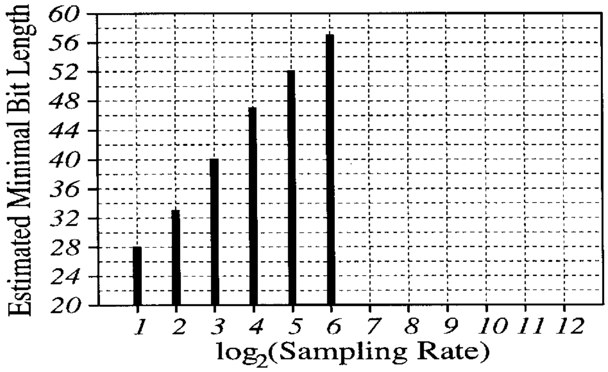
and the continuous-time stabilizing controller designed using the $\mathcal{H}_\infty$ method was

$$C(s) = \frac{\begin{array}{r}0.046s^6 + 1.5862s^5 + 3.09s^4 + 44.3s^3 \\ + 42.7785s^2 + 0.02867s + 1.58\times 10^{-4}\end{array}}{\begin{array}{r}s^6 + 3.766s^5 + 34.9509s^4 + 106.2s^3 \\ + 179.2s^2 + 166.43s + 0.0033\end{array}}. \qquad (49)$$

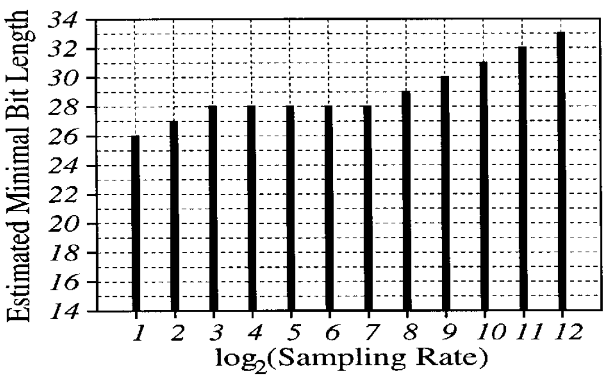
The range of the sampling rate tested in the simulation was 2^1 to 2^{12}, covering the slow to fast sampling conditions.

Given a sampling rate, the discrete-time plant model $P(z)$ and the digital controller $C(z)$ were obtained using the discretizing routines in MATLAB. It was found that the controllable canonical realization X_{can} of $C(z)$ became very ill-conditioned, resulting in an unstable closed-loop system, when the sampling rate was larger than 2^6. The initial realization X_0 was therefore chosen to be the non-canonical form as the result of a direct discretizing of the state-space model of $C(s)$. When X_0 was provided, the eigenvalues $\{\lambda_i^0\}$ of the ideal closed-loop system without FWL effects and the eigenvalue sensitivity matrices $\{\Phi_i\}$ were computed. The ASA was then used to search for an optimal transform matrix $\mathcal{T}_{opt}$ by solving for the minimization problem (33). This produced a corresponding optimal realization X_{opt} that maximizes the stability measure $\mu_1(X)$. For this example, the controller order was $n = 6$ and the optimization space had a dimension of $n \times n = 36$. This was by no means a small task. The ASA algorithm performed very efficiently.

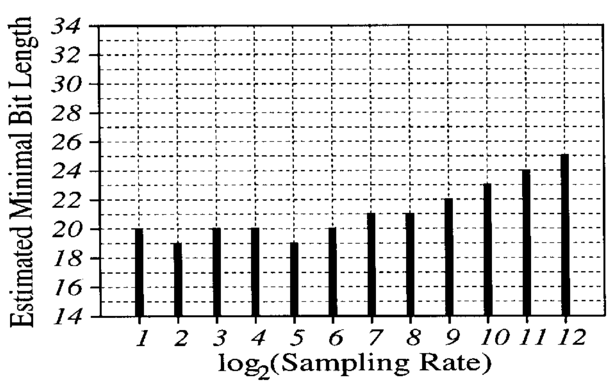
Figure 3 depicts the values of the FWL stability measure μ_1 for the three different realizations, X_{can}, X_0 and X_{opt}, under various sampling conditions. Figure 4 shows the corresponding estimated minimum bit length $\hat{B}_{s1}^{min}$ that can guarantee the closed-loop stability for these controller realizations. The results clearly show that



(a) Canonical realization X_{can}



(b) Initial (non-canonical) realization X_0



(c) Optimal realization X_{opt}

Figure 4. Estimated minimum bit length $\hat{B}_{s1}^{\min}$ as a function of sampling rate for different controller realizations.

the optimal controller realization has a much larger closed-loop stability margin than the non-optimal initial realization and requires a much smaller word length in fixed-point implementation. Specifically, for this example, the optimization achieved an improvement by two orders of magnitude on the stability measure and an average 8-bit reduction in the required minimum bit length over the range of sampling rates. It can also be seen that, for the controllable canonical realization, at the sampling rate of 2^6, the stability measure had already dropped to 10^{-16}, which indicated that the closed-loop system with this controller realization was very close to unstable.

6. Conclusions

This paper has presented an optimization procedure for obtaining the optimal realization of finite-precision digital controller structures. The procedure possesses the maximum closed-loop stability robustness subject to FWL implemented controller coefficients and requires the minimum bit length in fixed-point implementation. The approach can be regarded as an extension of an earlier procedure, derived for optimal finite-precision PID controller structures, to the general form of sampled-data controller structures. The study has shown that the optimal FWL controller realization problem can be formulated as a constrained nonlinear optimization problem. An efficient global optimization strategy based on the ASA algorithm has been developed to solve this non-smooth and non-convex optimization problem. The theoretical results have been verified using a numerical control system.

Acknowledgments

The third author, J. Wu, would like to thank the National Natural Science Foundation of China under Grant 69504010 and Cao Guangbiao Foundation of Zhejiang University for their financial support.

References

Aarts, E. H. L., and Korst, J. H. M., 1989, *Simulated Annealing and Boltzmann Machines* (Chichester: Wiley).

Beveridge, G. S. C., and Schechter, R. S., 1970, *Optimization: Theory and Practice* (McGraw-Hill).

Chen, T., and Francis, B. A., 1991, Input–output stability of sampled-data systems. *IEEE Transactions on Automatic Control*, **36,** 50–58.

Chen, S., Luk, B. L., and Liu, Y., 1998, Application of adaptive simulated annealing to blind channel identification with HOC fitting. *Electronics Letters*, **34**, 234–235.

Chen, S., Wu, J., Istepanian, R. H., and Chu, J., 2000, Optimizing stability bounds of finite-precision PID controller structures. *IEEE Transactions on Automatic Control*, to appear.

Dixon, L. C. W., 1972, *Nonlinear Optimisation* (London: The English Universities Press Ltd).

Fialho, I. J., and Georgiou, T. T., 1994, On stability and performance of sampled-data systems subject to wordlength constraint. *IEEE Transactions on Automatic Control*, **39,** 2476–2481.

Geman, S., and Geman, D., 1984, Stochastic relaxation, Gibbs distribution and the Bayesian restoration in images. *IEEE Transactions on Pattern Analysis and Machine Intelligent*, **6,** 721–741.

GEVERS, M., and LI, G., 1993, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects* (London: Springer Verlag).

INGBER, L., 1996, Adaptive simulated annealing (ASA): lessons learned. *Journal of Control and Cybernetics*, **25,** 33–54.

INGBER, L., and ROSEN, B. E., 1992, Genetic algorithms and very fast simulated reannealing: a comparison. *Mathematical and Computer Modelling*, **16,** 87–100.

ISTEPANIAN, R. H., 1997, Implementation issues for discrete PID algorithms using shift and delta operators parameterizations. In *Proceedings of the 4th IFAC Workshop Algorithms and Architectures for Real-Time Control*, Vilamoura, Portugal, pp. 117–122.

ISTEPANIAN, R. H., PRATT, I., GOODALL, R., and JONES, S., 1996, Effect of fixed point parameterization on the performance of active suspension control systems. In *Proceedings of the 13th IFAC World Congress*, San Francisco, USA, pp. 291–295.

ISTEPANIAN, R. H., LI, G., WU, J., and CHU, J., 1998a, Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds. *IEE Proceedings for Control Theory and Applications*, **145,** 472–478.

ISTEPANIAN, R. H., WU, J., WHIDBORNE, J. F., YAN, J., and SALCUDEAN, S. E., 1998b, Finite-word-length stability issues of tele-operation motion-scaling control system. In *Proceedings of UKACC Control '98*, Swansea, UK, pp. 1676–1681.

KOWALIK, J., and OSBORNE, M. R., 1968, *Methods for Unconstrained Optimization Problems* (New York: Elsevier).

LI, G., 1998, On the structure for digital controllers with finite word length consideration. *IEEE Transactions on Automatic Control*, **43,** 689–693.

MADIEVSKI, A. G., ANDERSON, B. D. O., and GEVERS, M., 1995, Optimum realizations of sampled-data controllers for FWL sensitivity minimization. *Automatica*, **31,** 367–379.

MORONEY, P., WILLSKY, A. S., and HOUPT, P. K., 1980, The digital implementation of control compensators: the coefficient wordlength issue. *IEEE Transactions on Automatic Control*, **25,** 621–630.

ROBERTS, R. A., and MULLIS, C. T., 1987, *Digital Signal Processing* (Addison-Wesley).

ROSEN, B. E., 1997, Rotationally parameterized very fast simulated reannealing. Submitted to *IEEE Transactions on Neural Networks*.