

Optimal controller and filter realizations using finite-precision, floating-point arithmetic

JAMES F. WHIDBORNE*[†], DA-WEI GU[‡], JUN WU[§] and SHENG CHEN[¶]

[†]Department of Aerospace Sciences, Cranfield University, Bedfordshire MK43 0AL, UK

[‡]Department of Engineering, University of Leicester, Leicester LE1 7RH, UK

[§]National Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou, 310027, PR China

[¶]Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

(Received 13 January 2003; revised 14 February 2005; accepted 8 April 2005)

The problem of reducing the fragility of digital controllers and filters implemented using finite-precision, floating-point arithmetic is considered. Floating-point arithmetic parameter uncertainty is multiplicative, unlike parameter uncertainty resulting from fixed-point arithmetic. Based on first-order eigenvalue sensitivity analysis, an upper bound on the eigenvalue perturbations is derived. Consequently, open-loop and closed-loop eigenvalue sensitivity measures are proposed. These measures are dependent upon the filter/controller realization. Problems of obtaining the optimal realization with respect to both the open-loop and the closed-loop eigenvalue sensitivity measures are posed. The problem for the open-loop case is completely solved. Solutions for the closed-loop case are obtained using non-linear programming. The problems are illustrated with a numerical example.

Keywords: Finite-precision arithmetic; Finite word length; Digital controller; Digital filter; Eigenvalue sensitivity; Floating-point arithmetic; Controller fragility

1. Introduction

The finite word-length used for number representation in digital computers means that controllers and filters implemented with digital hardware are subjected to errors. The errors in the arithmetic result from two sources (Mullis and Roberts 1976). The first is quantization errors resulting from the quantization of the signals and roundoff of the results of multiplication and addition. The second is coefficient errors resulting from the rounding of the coefficients of the filter/controller. This paper is concerned with the second of these. There are other finite word-length effects that need to be considered in implementing digital filters/controllers, notably the effects of overflow and (for floating point arithmetic) underflow and limit cycles resulting from the quantization. These are not considered in this paper.

In the past, digital controllers were often implemented using fixed point arithmetic; however, the reducing cost and increasing speed of computer hardware means that there is an increasing tendency for implementations to use floating-point arithmetic. It is well known (Wilkinson 1963) that quantization and rounding effects with floating point arithmetic is of a different nature to that of fixed point. Fixed-point quantization error results in additive noise independent of the signal, but with floating-point arithmetic, the quantization error is correlated with the signal that is being quantized. Similarly, coefficient rounding in fixed-point arithmetic results in additive perturbations on the coefficients, whereas with floating-point arithmetic the perturbations are multiplicative. Thus, the analysis and optimization of finite-precision filter and controller implementations needs to take the arithmetic into account.

The quantization error effect on digital filters resulting from the finite precision using floating-point arithmetic

*Corresponding author. Email: j.f.whidborne@cranfield.ac.uk

has been fairly extensively studied over the last four decades (for example, Sandberg 1967, Liu and Kaneko 1969, Kan and Aggarwal 1971, Kaneko and Liu 1971, Liu 1971, Zeng and Neuvo 1991, Smith *et al.* 1992, Rao 1996, Bomar *et al.* 1997, Tsai 1997, Ko and Bitmead 2004), see Kontro *et al.* (1992) for a review. The effect of coefficient rounding in floating-point arithmetic seems first to have been considered by Kaneko and Liu (1971) (see also Liu 1971), who analysed the sensitivity of the filter poles and the sensitivity of the frequency response to multiplicative perturbations on the coefficients for several filter structures. Liu (1971) also performs an analysis of the sensitivity of the filter frequency response, as do both Ku and Ng (1975) and Kalliojärvi and Astola (1994).

The finite-precision effects on closed-loop control systems have been extensively studied for fixed-point implementations; see Istepanian and Whidborne (2001) for a review. There has been far less work looking explicitly at the finite-precision effects for floating-point digital controller implementations. The quantization errors have been analysed by Rink and Chong (1979a, b), and by Vanwingerden and De Koning (1984) for optimal controllers. Miller *et al.* (1988) have also analysed the quantization errors, but notably also include the inter-sample behaviour. A method to design optimal controllers that minimize the quantization errors has been developed by de Oliveira and Skelton (2001). The effect on the robust stability caused by coefficient rounding has been analysed by Molchanov and Bauer (1995), but an additive perturbation is assumed for the floating point implementation. Closed-loop stability subject to perturbations on the floating-point coefficients has been analysed by Faris *et al.* (1998) using modern robust techniques. The sensitivity of the time responses has been analysed by Farrell and Michel (1989) for both fixed and floating-point arithmetic.

It is known that some controller/filter realizations are very sensitive to small errors in the parameters and these small errors can even lead to instability. These parameter errors may result from the finite-precision of the computing device. Such controller realizations can be described as *fragile* (Keel and Bhattacharyya 1997). However, a dynamical system has an infinite number of equivalent realizations. If a digital linear system is implemented in the state space form, $C(zI - A)^{-1}B + D$, then $CT(zI - T^{-1}AT)^{-1}T^{-1}B + D$ is an equivalent realization for any non-singular matrix T . It so happens that the effect of the finite precision is partially dependent upon the realization. Thus, in order to ensure a non-fragile implementation, it is of interest to know the realization or matrix T which minimizes the effect on the system of the finite precision.

One approach to obtaining non-fragile realizations is to minimize the sensitivity of the system eigenvalues. This approach has been extensively investigated for fixed-point realizations. It was first considered for the open-loop (filter) case by Mantey (1968) and subsequently by Gevers and Li (1993), who solved the problem for state-space realizations based on a norm for the open-loop eigenvalue sensitivities. The case of the closed-loop system eigenvalue sensitivity for state-space controller realizations was first considered by Li (1998) and has subsequently been thoroughly investigated (Istepanian *et al.* 1998, 2000, Chen *et al.* 1999, Wu *et al.* 1999, 2000, Whidborne *et al.* 2001).

In this paper, a simple eigenvalue sensitivity measure is considered for both filter and controller realizations. The filter problem is completely solved whilst solutions to the controller problem may be obtained using non-linear programming. The main results of this paper were originally presented by Whidborne and Gu (2002). Other eigenvalue sensitivity minimization indices for floating-point implementations have recently been proposed by Wu *et al.* (2003, 2004). An alternative eigenvalue sensitivity index has been proposed for floating point arithmetic by Ko and Yu (2004) and conditions for the existence of a minimizing realization established. However, additive perturbations on the coefficients are assumed, and this index is actually an upper bound on an index proposed by Whidborne *et al.* (2001).

In the next section, floating-point arithmetic is discussed and the rounding operation is shown to result in multiplicative perturbations on the filter/controller coefficients. Based on this perturbation model, an upper bound on the eigenvalue perturbations is obtained in section 3. In section 4, a measure of the relative stability based on this upper bound is proposed for digital filter implementations, and the problem of minimising this measure for state-space realizations is solved. In section 5, a similar measure for closed-loop controller implementations is proposed. Non-linear programming is proposed to obtain solutions to the closed-loop problem. The problems are illustrated by a numerical example in the penultimate section and non-linear programming is shown to be effective for the closed-loop problem.

1.1. Notations

$[x]$	floor function, that is the largest integer less than or equal to x
$A \circ B = [a_{ij}b_{ij}]$	Hadamard product of A and B
A^T	transpose of a matrix A
A^H	complex conjugate transpose of a matrix A

$\text{vec}(A)$	column stacking operator of a matrix A
$\ A\ _F = \sqrt{\sum_{i,j} a_{ij}^2}$	Frobenius norm of a matrix A for a matrix $A \geq 0$, the unique symmetric matrix satisfying $A^{1/2} \geq 0$ and $A^{1/2}A^{1/2} = A$
\mathbb{C}	set of all complex numbers
\mathbb{R}	set of all real numbers
\mathbb{Z}	set of all integers
$\mathcal{O}(x)$	'is of order x '

2. Floating-point representation

Numbers in a digital computer are represented by a finite number of bits—the word-length, $\ell \in \mathbb{Z}_+$. In a floating-point arithmetic, the word consists of three parts:

- (1) one bit, $s \in \{0, 1\}$, for the sign of the number,
- (2) $\ell_m \in \mathbb{Z}_+$ bits for the mantissa, $m \in \mathbb{R}$, and
- (3) $\ell_e \in \mathbb{Z}_+$ bits for the exponent, $e \in \mathbb{Z}$.

Therefore, $\ell = \ell_m + \ell_e + 1$. The number is typically stored as shown in figure 1 and, with this representation, the value x is interpreted as

$$x = (-1)^s \times m \times 2^e \quad (1)$$

where the mantissa is usually normalized so that $m \in [0.5, 1)$. Now, since ℓ_e and ℓ_m are finite (ℓ is typically 16, 32 or 64 bits), the set of numbers that is represented by a particular floating-point scheme is not dense on the real line. Thus, the set of possible floating-point numbers, \mathcal{F} , is given by

$$\mathcal{F} := \left\{ (-1)^s \left(0.5 + \sum_{i=1}^{\ell_m} b_i 2^{-(i+1)} \right) \times 2^e : s \in \{0, 1\}, \right. \\ \left. b_i \in \{0, 1\}, e \in \mathbb{Z}, \underline{e} \leq e \leq \bar{e} \right\} \cup \{0\} \quad (2)$$

where $\underline{e} \in \mathbb{Z}$ and $\bar{e} \in \mathbb{Z}$ represent the lower and upper limits of the exponent, respectively, and $\bar{e} - \underline{e} = 2^{\ell_e} - 1$. Note that, unlike fixed-point representation, underflow can occur in floating-point arithmetic.

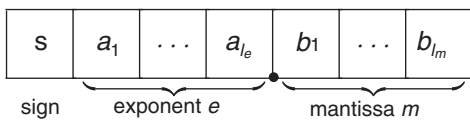


Figure 1. Floating-point number representation.

In the remainder of this paper, it is assumed that no underflow or overflow occurs, that is ℓ_e is unlimited, so $e \in \mathbb{Z}$. Define the floating-point rounding operator, $q : \mathbb{R} \rightarrow \mathcal{F}$, as

$$q(x) := \begin{cases} \text{sgn}(x)2^{(e-\ell_m-1)} \lfloor 2^{(\ell_m-e+1)}|x| + 0.5 \rfloor, & \text{for } x \neq 0 \\ 0, & \text{for } x = 0 \end{cases} \quad (3)$$

where $e = \lfloor \log_2 |x| \rfloor + 1$.

The rounding error, ε , is defined as

$$\varepsilon := |x - q(x)|. \quad (4)$$

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

$$\varepsilon < |x|2^{-(\ell_m+1)}. \quad (5)$$

Thus, when a number is implemented in finite-precision floating-point arithmetic, it may be perturbed to

$$q(x) = x(1 + \delta), \quad |\delta| < \delta_{\max}. \quad (6)$$

where $\delta_{\max} = 2^{-(\ell_m+1)}$. Thus, as is well-known (Wilkinson 1963), the perturbation is multiplicative, unlike the perturbation resulting using finite-precision fixed-point arithmetic, which is additive.

3. Eigenvalue sensitivity

In general, the perturbations on the controller parameters resulting from finite-precision implementation will be very small. Thus, perturbations on the closed-loop system eigenvalues can be approximated by considering the first-order term of a Taylor expansion, i.e. the eigenvalue sensitivities to changes in the controller parameters. A number of different eigenvalue sensitivity indices have been proposed for fixed-point digital controller and filter implementations (Mantey 1968, Gevers and Li 1993, Istepanian *et al.* 1998, Li 1998, Whidborne 2001, Wu *et al.* 2001).

Assume that a controller/filter realization, $x = \text{vec}(X)$, is implemented with floating-point arithmetic with finite precision, that is the actual realization will be $q(x)$. Then, from (6), each element of x will be perturbed to $x_i(1 + \delta_i)$, $|\delta_i| < \delta_{\max} = 2^{-(\ell_m+1)}$ and the realization vector will be perturbed to $x + x \circ \delta$, where $\delta = [\delta_i]$.

Proposition 1: *Let $f(x) \in \mathbb{C}$ be a differentiable function of $x \in \mathbb{R}^{n_x}$. Assume that x is perturbed to \tilde{x} where $\tilde{x}_i = x_i(1 + \delta_i)$. Then, to a first-order Taylor series approximation*

$$|f(\tilde{x}) - f(x)| \leq \delta_{\max} \|g\|_2 \|x\|_2 + |\mathcal{O}(\delta_{\max}^2)| \quad (7)$$

where $|\delta_i| < \delta_{\max}$ for all i and $g(x)$ is the gradient vector, i.e.

$$g(x) := \frac{\partial f(x)}{\partial x} = \left[\frac{\partial f}{\partial x_i} \right]_x \quad (8)$$

evaluated at x .

Proof: Taking a first-order Taylor series approximation:

$$f(\tilde{x}) = f(x) + \sum_{i=1}^{n_x} \left(\frac{\partial f}{\partial x_i} \right)_x (\tilde{x}_i - x_i) + \mathcal{O}(\delta_{\max}^2) \quad (9)$$

Now, from (6), $\tilde{x}_i = x_i(1 + \delta_i)$, so

$$f(\tilde{x}) - f(x) = \sum_{i=1}^{n_x} g_i(x) x_i \delta_i + \mathcal{O}(\delta_{\max}^2). \quad (10)$$

Hence

$$|f(\tilde{x}) - f(x)| \leq \sum_{i=1}^{n_x} |g_i(x)| |x_i| |\delta_i| + |\mathcal{O}(\delta_{\max}^2)| \quad (11)$$

$$< \delta_{\max} \sum_{i=1}^{n_x} |g_i(x)| |x_i| + |\mathcal{O}(\delta_{\max}^2)| \quad (12)$$

which, by the Cauchy-Schwartz inequality, gives

$$|f(\tilde{x}) - f(x)| < \delta_{\max} \|g(x)\|_2 \|x\|_2 + |\mathcal{O}(\delta_{\max}^2)|. \quad (13)$$

□

If $f(\cdot)$ is the system pole/eigenvalue, x is the infinite-precision parameter vector and \tilde{x} is the finite-precision parameter vector, then Proposition 1 can be used to measure the relative system stability when subject to finite-precision implementation using floating-point arithmetic. Based on Proposition 1, tractable eigenvalue sensitivity indices can be formulated which are appropriate for finite-precision floating-point digital controller and filter implementations.

4. Optimal digital filter realizations

Consider the problem of implementing a digital filter, $F(z) = C_f(zI - A_f)^{-1}B_f + D_f$, where $A_f \in \mathbb{R}^{n \times n}$ and has no repeated eigenvalues, $B_f \in \mathbb{R}^{n \times q}$, $C_f \in \mathbb{R}^{l \times n}$ and $D_f \in \mathbb{R}^{l \times q}$. In this paper, (A_f, B_f, C_f, D_f) is also called a realization of $F(z)$. The realizations of $F(z)$ are not unique, if $(A_f^0, B_f^0, C_f^0, D_f^0)$ is a realization of $F(z)$, then so is $(T^{-1}A_f^0T, T^{-1}B_f^0, C_f^0T, D_f^0)$ for any non-singular

similarity transformation $T \in \mathbb{R}^{n \times n}$. The system poles are simply the eigenvalues of A_f . The problem under consideration is to find the similarity transformation such that the realization has a minimal eigenvalue sensitivity when implemented using finite word-length floating-point arithmetic.

Based on Proposition 1, the following tractable eigenvalue sensitivity index, Φ , is proposed

$$\Phi = \|A_f\|_F^2 \sum_{k=1}^n w_k \Phi_k \quad (14)$$

where w_k is a non-negative real scalar weighting and

$$\Phi_k = \left\| \frac{\partial \lambda_k}{\partial A_f} \right\|_F^2 \quad (15)$$

where $\{\lambda_i: i=1, \dots, n\}$ represents the set of unique eigenvalues of A_f . The weights, w_k , $k=1, \dots, n$, are generally chosen so that the eigenvalues closer to the unit circle have the larger values. The measure Φ is dependent upon the filter realization, that is, given $A_f = T^{-1}A_f^0T$,

$$\Phi(T) := \left\| T^{-1}A_f^0T \right\|_F^2 \sum_{k=1}^n w_k \Phi_k(T) \quad (16)$$

where (Gevers and Li 1993, Li 1998),

$$\Phi_k(T) = \text{tr}(R_k^H T^{-T} T^{-1} R_k) \text{tr}(L_k^H T T^T L_k) \quad (17)$$

and where R_k and L_k are the right and left eigenvectors, respectively, for the k th eigenvalue of A_f^0 .

Problem 1: Given an initial realization $(A_f^0, B_f^0, C_f^0, D_f^0)$, calculate

$$\Phi_{\min} = \min_{\substack{T \in \mathbb{R}^{n \times n} \\ \det(T) \neq 0}} \Phi(T) \quad (18)$$

and calculate a subsequent similarity transformation T_{\min} such that $\Phi_{\min} = \Phi(T_{\min})$.

Theorem 1: The solution to Problem 1 is given by

$$\Phi_{\min} = \sum_{k=1}^n |\lambda_k|^2 \sum_{k=1}^n w_k \quad (19)$$

and

$$T_{\min} = (RWR^H)^{1/2} V \quad (20)$$

where $R=[R_i]$ is the matrix of right eigenvectors of A_f^0 , $W = \text{diag}(w_1, \dots, w_n)$ is a diagonal matrix of the weights and V is an arbitrary orthogonal matrix.

Proof: From Lemma 6.2 and Theorem 6.1 of Gevers and Li (1993, pp. 137–138), it follows that $\Phi_k \geq 1$ with equality for all k if A_f is normal. From Horn and Johnson (1985, p. 101),

$$\|A_f\|_F^2 \geq \sum_{k=1}^n |\lambda_k|^2 \quad (21)$$

with equality if A_f is normal. Clearly, if A_f is normal, Φ is minimal and (19) holds. Theorem 6.2 of Gevers and Li (1993, p. 141) gives (20). \square

Remark 1: The requirement for minimal eigenvalue sensitivity for FWL fixed-point arithmetic is also that the transition matrix A_f is in the normal form (Gevers and Li 1993, p. 139).

5. Optimal digital controller realizations

Consider the linear discrete-time feedback control system shown in figure 2. Let the plant be $P(z)$ and let the controller be $C(z, X)$, where X is the parametrization of the controller.

Let $(A_p, B_p, C_p, 0)$ be a state space description of the strictly proper plant $P(z) = C_p(zI - A_p)^{-1}B_p$, $A_p \in \mathbb{R}^{m \times m}$, $B_p \in \mathbb{R}^{m \times l}$ and $C_p \in \mathbb{R}^{q \times m}$. Let (A_c, B_c, C_c, D_c) be a state space description of $C(z) = C_c(zI - A_c)^{-1}B_c + D_c$, where $A_c \in \mathbb{R}^{n \times n}$, $B_c \in \mathbb{R}^{n \times q}$, $C_c \in \mathbb{R}^{l \times n}$ and $D_c \in \mathbb{R}^{l \times q}$.

The transition matrix of the closed loop system is

$$\begin{aligned} \bar{A} &= \begin{bmatrix} A_p + B_p D_c C_p & B_p C_c \\ B_c C_p & A_c \end{bmatrix} \\ &= \begin{bmatrix} A_p & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B_p & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix} \begin{bmatrix} C_p & 0 \\ 0 & I_n \end{bmatrix}, \quad (22) \\ &=: M_0 + M_1 X M_2 = \bar{A}(X), \end{aligned}$$

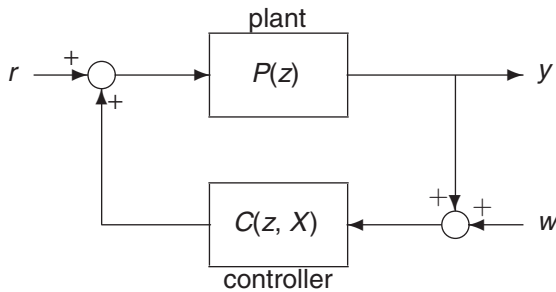


Figure 2. Feedback control system.

where

$$X := \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix}, \quad (23)$$

In the sequel, it is assumed that \bar{A} has no repeated eigenvalues.

Let the realization $(A_c^0, B_c^0, C_c^0, D_c^0)$ of $C(z)$ be represented by

$$X_0 = \begin{bmatrix} D_c^0 & C_c^0 \\ B_c^0 & A_c^0 \end{bmatrix}, \quad (24)$$

then any realization is given by

$$X = \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}^{-1} \begin{bmatrix} D_c^0 & C_c^0 \\ B_c^0 & A_c^0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}, \quad (25)$$

$$=: T_I^{-1} X_0 T_I, \quad (26)$$

for some non-singular $T \in \mathbb{R}^{n \times n}$.

Let $R_k = (R_k^T(1) R_k^T(2))^T$ and $L_k = (L_k^T(1) L_k^T(2))^T$ be the right and left eigenvectors respectively, for the k th eigenvalue of \bar{A} partitioned such that $R_k(1), L_k(1) \in \mathbb{C}^m$ and $R_k(2), L_k(2) \in \mathbb{C}^n$, i.e. the partitions correspond to the partitions of X defined by (23). Then, it can be shown (Li 1998, Whidborne *et al.* 2001) that

$$\left(\frac{\partial \lambda_k}{\partial A_c} \right)^T = R_k(2) L_k^H(2), \quad (27)$$

$$\left(\frac{\partial \lambda_k}{\partial B_c} \right)^T = C_p R_k(1) L_k^H(2), \quad (28)$$

$$\left(\frac{\partial \lambda_k}{\partial C_c} \right)^T = R_k(2) L_k^H(1) B_p, \quad (29)$$

$$\left(\frac{\partial \lambda_k}{\partial D_c} \right)^T = C_p R_k(1) L_k^H(1) B_p, \quad (30)$$

where $\{\lambda_k: k = 1, \dots, n+m\}$ represents the set of unique eigenvalues of \bar{A} .

Based on Proposition 1, the following tractable eigenvalue sensitivity index, Υ , is proposed

$$\Upsilon(X) := \|X\|_F^2 \sum_{k=1}^{n+m} w_k \Upsilon_k \quad (31)$$

where w_k is a non-negative real scalar weighting and

$$\Upsilon_k = \left\| \frac{\partial \lambda_k}{\partial A_c} \right\|_F^2 + \left\| \frac{\partial \lambda_k}{\partial B_c} \right\|_F^2 + \left\| \frac{\partial \lambda_k}{\partial C_c} \right\|_F^2 + \left\| \frac{\partial \lambda_k}{\partial D_c} \right\|_F^2. \quad (32)$$

The weights, w_k , $k = 1, \dots, n + m$, are generally chosen so that the eigenvalues closer to the unit circle have the larger values. The measure Υ is dependent upon the controller realization. Given an initial realization $(A_f^0, B_f^0, C_f^0, D_f^0)$, then it can be easily shown that

$$\|X\|_F^2 = \text{tr}(P^{-1}A_c^0PA_c^{0T}) + \text{tr}(P^{-1}B_c^0B_c^{0T}) + \text{tr}(PC_c^{0T}C_c^0) + \text{tr}(D_c^0D_c^{0T}), \quad (33)$$

where $P = TT^T$ and, from (27)–(30), that

$$\begin{aligned} \Upsilon_k &= \text{tr}(R_k^{0H}(2)P^{-1}R_k^0(2)) \text{tr}(L_k^{0H}(2)PL_k^0(2)) \\ &+ \alpha_k \text{tr}(L_k^{0H}(2)PL_k^0(2)) \\ &+ \beta_k \text{tr}(R_k^{0H}(2)P^{-1}R_k^0(2)) + \alpha_k \beta_k, \end{aligned} \quad (34)$$

where $\alpha_k = \text{tr}(R_k^{0H}(1)C_p^H C_p R_k^0(1))$ and $\beta_k = \text{tr}(L_k^{0H}(1)B_p B_p^H L_k^0(1))$. Rearranging gives

$$\begin{aligned} \Upsilon(P) &= \left(\text{tr}(P^{-1}A_c^0PA_c^{0T}) + \text{tr}(P^{-1}B_c^0B_c^{0T}) + \text{tr}(PC_c^{0T}C_c^0) \right. \\ &+ \left. \text{tr}(D_c^0D_c^{0T}) \right) \left(\sum_{k=1}^{n+m} \text{tr}(P^{-1}M_{R_k}) \text{tr}(PM_{L_k}) \right) \\ &+ \text{tr}(PW_L) + \text{tr}(P^{-1}W_R) + c \end{aligned} \quad (35)$$

where

$$M_{R_k} = w_k^{1/2} R_k^0(2) R_k^{0H}(2) \quad (36)$$

$$M_{L_k} = w_k^{1/2} L_k^0(2) L_k^{0H}(2) \quad (37)$$

$$W_L = L^0(2) \text{diag}(w_1 \alpha_1, \dots, w_{n+m} \alpha_{n+m}) L^{0H}(2), \quad (38)$$

$$W_R = R^0(2) \text{diag}(w_1 \beta_1, \dots, w_{n+m} \beta_{n+m}) R^{0H}(2), \quad (39)$$

are all Hermitian and

$$c = \sum_{k=1}^{n+m} \alpha_k \beta_k. \quad (40)$$

Problem 2: Given an initial realization $(A_c^0, B_c^0, C_c^0, D_c^0)$, calculate

$$\Upsilon_{\min} = \min_{\substack{P \in \mathbb{R}^{n \times n} \\ P = P^T > 0}} \Upsilon(P) \quad (41)$$

where $P = TT^T$ and calculate a subsequent similarity transformation T_{\min} such that $\Upsilon_{\min} = \Upsilon(T_{\min} T_{\min}^T)$.

Remark 2: The function $\Upsilon(P)$ is everywhere differentiable over the set $\{\Upsilon(P): P = P^T > 0\}$. Hence, it is

proposed that non-linear programming is used to find local solutions to Problem 2. The problem of finding a global solution remains open.

To solve the problem using non-linear programming, a search is required over $n \times n$ real, positive definite symmetric matrices. This can be accomplished by utilising a Cholesky factorization given by the following theorem (Golub and Van Loan 1989, p. 141).

Theorem 2 (Cholesky factorization): For $P \in \mathbb{R}^{n \times n}$, $P = P^T$, $P > 0$, there exists a unique lower triangular $G \in \mathbb{R}^{n \times n}$ with positive diagonal entries such that $P = GG^T$.

Thus a search can be made over the set

$$\left\{ \begin{bmatrix} g_a \\ g_b \end{bmatrix} : g_a \in \mathbb{R}^{(n-1)n/2}, g_b \in \mathbb{R}_+^n \right\}.$$

Remark 3: Since $VV^T = I$ where V is any orthogonal matrix, then $P_{\min} = G_{\min} VV^T G_{\min}^T$ and so

$$T_{\min} = G_{\min} V. \quad (42)$$

This provides an extra degree of freedom which could be utilized to find, for example, sparse realizations (Li *et al.* 1992).

6. Example

The following numerical example is taken from Gevers and Li (1993, pp. 236–237). The discrete time system to be controlled is given by

$$A_p = \begin{bmatrix} 3.7156 & -5.4143 & 3.6525 & -0.9642 \\ 1.000 & 0 & 0 & 0 \\ 0 & 1.000 & 0 & 0 \\ 0 & 0 & 1.000 & 0 \end{bmatrix}, \quad (43)$$

$$B_p = [1 \ 0 \ 0 \ 0]^T, \quad (44)$$

$$C_p = [0.1116 \ 0.0043 \ 0.1088 \ 0.0014] \times 10^{-5}. \quad (45)$$

A pole-placement controller is designed to place the closed-loop poles at

$$0.9844 \pm 0.0357j, 0.9643 \pm 0.0145j, \quad (46)$$

and a state observer is designed with poles located at

$$0.7152 \pm 0.6348j, 0.3522 \pm 0.2857j. \quad (47)$$

The initial realization of the feedback controller $C(z)$ is given by (to 4 decimal places)

$$A_c^0 = A_p + B_p C_c^0 - B_c^0 C_p \quad (48)$$

$$= \begin{bmatrix} 2.6743 & -5.7443 & 2.5096 & -0.9176 \\ 0.2877 & -0.0273 & -0.6947 & -0.0088 \\ -0.3377 & 0.9871 & -0.3294 & -0.0042 \\ -0.0830 & -0.0032 & 0.9190 & -0.0010 \end{bmatrix}, \quad (49)$$

$$B_c^0 = [1.0963 \quad 0.6385 \quad 0.3027 \quad 0.0744]^T \times 10^6, \quad (50)$$

$$C_c^0 = [0.1818 \quad -0.2831 \quad 0.0500 \quad 0.0617], \quad (51)$$

$$D_c^0 = 0. \quad (52)$$

The weights are set to $w_i = (1 - \lambda_{\max})/(1 - |\lambda_i|)$ where $\lambda_{\max} = \max_i \{|\lambda_i|\}$ and $\{\lambda_i\}$ are the eigenvalues of A_c^0 and \bar{A} (from (22)) for the open-loop and closed-loop sensitivity indices, respectively. Thus, the eigenvalues closer to the unit circle have the larger weighting values.

The initial realization has an open-loop pole sensitivity, $\Phi = 1.5737 \times 10^6$. From Theorem 1, the optimal open-loop pole sensitivity $\Phi_{\min} = 6.1746$, which can be achieved with the realization (to 4 decimal places):

$$A_c = \begin{bmatrix} 0.6194 & -0.1992 & -0.0835 & -0.1265 \\ 0.1346 & 0.6052 & -0.2297 & 0.0171 \\ 0.0508 & 0.1650 & 0.5315 & -0.2813 \\ 0.2047 & 0.0653 & 0.2218 & 0.5605 \end{bmatrix}, \quad (53)$$

$$B_c = [0.6508 \quad 0.0048 \quad 2.0020 \quad 0.2961]^T \times 10^6, \quad (54)$$

$$C_c = [0.1100 \quad 0.0222 \quad -0.0142 \quad -0.0168]. \quad (55)$$

The closed-loop pole sensitivity for the initial realization is $\Upsilon = 3.9903 \times 10^{22}$ and for the open-loop optimal realization, it is $\Upsilon = 9.8156 \times 10^{21}$. It is a fairly common practice to implement controllers using a balanced realization. Using the MATLAB[®] routine `balreal.m` a balanced realization was obtained (to 4 decimal places):

$$A_c = \begin{bmatrix} 0.1119 & 0.5408 & -0.1954 & -0.0531 \\ -0.5408 & 0.7216 & 0.1647 & 0.0350 \\ -0.1954 & -0.1647 & 0.7643 & -0.1298 \\ 0.0531 & 0.0350 & 0.1298 & 0.7189 \end{bmatrix}, \quad (56)$$

$$B_c = [203.1819 \quad -63.5703 \quad 32.0424 \quad -4.1143]^T, \quad (57)$$

$$C_c = [203.1819 \quad 63.5703 \quad 32.0424 \quad 4.1143]. \quad (58)$$

The closed-loop pole sensitivity for the balanced realization is $\Upsilon = 1.2546 \times 10^{11}$.

The MATLAB[®] routine `fminsearch.m` was used with the Cholesky factorization of Theorem 2 to solve Problem 2. The routine `fminsearch.m` implements the Nelder-Mead simplex method. Using a 350 MHz Pentium PC, from a random starting point, the routine took ~ 30 minutes to converge. An optimal closed-loop pole sensitivity value of $\Upsilon_{\min} = 4.3366 \times 10^8$ was obtained with a realization (to 4 decimal places):

$$A_c = \begin{bmatrix} -1.0614 & -0.9631 & -0.0054 & -0.0018 \\ 2.2892 & 1.7570 & -0.0235 & 0.0057 \\ -1.4089 & 0.4759 & 0.6716 & -0.0868 \\ 1.7421 & -2.3837 & 0.4706 & 0.9494 \end{bmatrix}, \quad (59)$$

$$B_c = [129.2367 \quad -137.2672 \quad 56.4560 \quad -23.7868]^T, \quad (60)$$

$$C_c = [155.1427 \quad -119.5560 \quad 32.1475 \quad 1.0368]. \quad (61)$$

7. Discussion and conclusions

In previous works, the eigenvalue sensitivity approach to obtain optimal digital filter and controller realizations so as to account for the finite precision inherent in digital computing devices has been thoroughly investigated. However, there has been an assumption that the parameter uncertainty is additive. This assumption is perfectly valid for filter and controller implementations that use fixed-point arithmetic, however, for floating-point arithmetic, the parameter uncertainty is multiplicative. It is becoming increasingly common to use floating-point arithmetic for digital filters and controllers. Thus, in this paper, the work of Gevers and Li (1993) is extended to obtain optimal floating-point digital filter realizations; and the work of Whidborne *et al.* (2001) is extended to obtain optimal floating-point digital controller realizations.

The methods are demonstrated on a numerical example of a control system. Both the initial realization of the controller and the optimal open-loop realization result in very high closed-loop pole sensitivities. This is significantly reduced by using a balanced realization. However, the closed-loop pole sensitivity of the balanced controller realization can be reduced by three orders of magnitude by the optimal closed-loop realization.

References

- B.W. Bomar, L.M. Smith and R.D. Joseph, "Roundoff noise analysis of state-space digital filters implemented on floating-point

- digital signal processors", *IEEE Trans. Circuits & Syst. II*, 44, pp. 952–955, 1997.
- S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures", *IEEE Trans. Autom. Control*, 44, pp. 2149–2153, 1999.
- M.C. de Oliveira and R.E. Skelton, "State feedback control of linear systems in the presence of devices with finite signal-to-noise ratio", *Int. J. Control*, 74, pp. 1501–1509, 2001.
- D. Faris, T. Pare, A. Packard, K.A. Ali and J.P. How, "Controller fragility: what's all the fuss?", in *Proc. Annual Allerton Conference on Communication Control and Computing*, Monticello, Illinois, 1998, Vol. 36, pp. 600–609.
- J.A. Farrell and A.N. Michel, "Estimates of asymptotic trajectory bounds in digital implementations of linear feedback control systems", *IEEE Trans. Autom. Control*, 34, pp. 1319–1324, 1989.
- M. Gevers and G. Li, *Parametrizations in Control, Estimations and Filtering Problems: Accuracy Aspects*, Berlin: Springer-Verlag, 1993.
- G.H. Golub and C.F. Van Loan, *Matrix Computations*, Baltimore, MD: John Hopkins University Press, 1989.
- R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge, UK: Cambridge University Press, 1985.
- 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", *IEE Proc. Control Theory and Appl.*, 145, pp. 472–478, 1998.
- R.S.H. Istepanian and J.F. Whidborne, "Finite-precision computing for digital control systems: current status and future paradigms", in *Digital Controller Implementation and Fragility: A Modern Perspective*, London, UK: Springer-Verlag, 2001, pp. 1–12.
- R.S.H. Istepanian, S. Chen, J. Wu and J.F. Whidborne, "Optimal finite-precision controller realization of sampled-data systems", *Int. J. Systems Sci.*, 31, pp. 429–438, 2000.
- K. Kalliojärvi and J. Astola, "Required coefficient word length in floating-point and logarithmic digital filters", *IEEE Sig. Proc. Letters*, 1, pp. 52–54, 1994.
- E.P.F. Kan and J.K. Aggarwal, "Error analysis of digital filter employing floating-point arithmetic", *IEEE Trans. Circuit Theory*, 18, pp. 678–686, 1971.
- T. Kaneko and B. Liu, "Effect of coefficient rounding in floating-point digital filters", *IEEE Trans. Aerosp. Electron. Syst.*, 7, pp. 995–1003, 1971.
- L.H. Keel and S.P. Bhattacharyya, "Robust, fragile, or optimal?", *IEEE Trans. Autom. Control*, 42, pp. 1098–1105, 1997.
- H.-J. Ko and W.-S. Yu, "Guaranteed robust stability of the closed-loop systems for digital controller implementations via orthogonal Hermitian transform", *IEEE Trans. Syst. Man & Cybernetics – B*, 34, pp. 1923–1932, 2004.
- S. Ko and R.R. Bitmead, "Covariance calculation for floating-point state-space realizations", *IEEE Trans. Sig. Proc.*, 52, pp. 3370–3377, 2004.
- J. Kontro, K. Kalliojärvi and Y. Neuvo, "Floating-point arithmetic in signal processing", in *Proc 1992 IEEE Int. Symp. Circuits and Syst.* San Diego, CA, 1992, pp. 1784–1791.
- W. Ku and S.-M. Ng, "Floating-point coefficient sensitivity and roundoff noise of recursive digital filters realized in ladder structures", *IEEE Trans. Circuits & Syst.*, 22, pp. 927–936, 1975.
- G. Li, "On the structure of digital controllers with finite word length consideration", *IEEE Trans. Autom. Control*, 43, pp. 689–693, 1998.
- G. Li, B.D.O. Anderson, M. Gevers and J.E. Perkins, "Optimal FWL design of state space digital systems with weighted sensitivity minimization and sparseness consideration", *IEEE Trans. Circuits & Syst. II*, 39, pp. 365–377, 1992.
- B. Liu, "Effects of finite word-length on the accuracy of digital filters—a review", *IEEE Trans. Circuit Theory*, 18, pp. 670–677, 1971.
- B. Liu and T. Kaneko, "Error analysis of digital filters realised with floating-point arithmetic", in *Proc. IEEE*, 57, pp. 1735–1747, 1969.
- P.E. Mantey, "Eigenvalue sensitivity and state-variable selection", *IEEE Trans. Autom. Control*, 13, pp. 263–269, 1968.
- R.K. Miller, M.S. Mousa and A.N. Michel, "Quantization and overflow effects in digital implementations of linear dynamic controllers", *IEEE Trans. Autom. Control*, 33, pp. 698–704, 1988.
- A.P. Molchanov and P.H. Bauer, "Robust stability of digital feedback control systems with floating point arithmetic", in *Proc. 34th IEEE Conf. Decision Contr.*, New Orleans, LA, 1995, pp. 4251–4258.
- C.T. Mullis and R.A. Roberts, "Synthesis of minimum round off noise fixed-point digital filters", *IEEE Trans. Circuits & Syst.*, 23, pp. 551–562, 1976.
- B.D. Rao, "Roundoff noise in floating point digital filters", *Control and Dynamic Systems*, 75, pp. 79–103, 1996.
- R.E. Rink and H.Y. Chong, "Covariance equation for a floating-point regulator system", *IEEE Trans. Autom. Control*, 24, pp. 980–982, 1979a.
- R.E. Rink and H.Y. Chong, "Performance of state regulator systems with floating point computation", *IEEE Trans. Autom. Control*, 24, pp. 411–421, 1979b.
- I.W. Sandberg, "Floating-point-roundoff accumulation in digital-filter realizations", *Bell Syst. Tech. J.*, 46, pp. 1775–1791, 1967.
- L.M. Smith, B.W. Bomar, R.D. Joseph and G.C.-J. Yang, "Floating-point roundoff noise analysis of second-order state-space digital filter structures", *IEEE Trans. Circuits & Syst. II*, 39, pp. 90–98, 1992.
- C.M. Tsai, "Floating-point roundoff noises of first- and second-order sections in parallel form digital filters", *IEEE Trans. Circuits & Syst. II*, 44, pp. 774–779, 1997.
- A.J.M. Vanwingerden and W.L. De Koning, "The influence of finite word-length on digital optimal-control", *IEEE Trans. Autom. Control*, 29, pp. 385–391, 1984.
- J.F. Whidborne and D.-W. Gu, "Optimal finite-precision controller and filter implementations using floating-point arithmetic", in *Proc. 15th IFAC World Congress*, Barcelona, 2002 CD-ROM Paper 990.
- J.F. Whidborne, R.S.H. Istepanian and J. Wu, "Reduction of controller fragility by pole sensitivity minimization", *IEEE Trans. Autom. Control*, 46, pp. 320–325, 2001.
- J.H. Wilkinson, *Rounding Errors in Algebraic Processes*. London, UK: HMSO, 1963.
- J. Wu, R.H. Istepanian and S. Chen, "Stability issues of finite-precision controller structures for sampled-data systems", *Int. J. Control*, 72, pp. 1331–1342, 1999.
- J. Wu, S. Chen, G. Li and J. Chu, "Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems", *IEEE Trans. Autom. Control*, 45, pp. 1550–1554, 2000.
- 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. Autom. Control*, 46, pp. 1662–1666, 2001.
- J. Wu, S. Chen, J.F. Whidborne and J. Chu, "A unified closed-loop stability measure for finite-precision digital controller realizations implemented in different realization schemes", *IEEE Trans. Autom. Control*, 48, pp. 816–822, 2003.
- J. Wu, S. Chen, J.F. Whidborne and J. Chu, "Optimal realizations of floating-point implemented digital controllers with finite word length considerations", *Int. J. Control*, 77, pp. 427–440, 2004.
- B. Zeng and Y. Neuvo, "Analysis of floating point roundoff errors using dummy multiplier coefficient sensitivities", *IEEE Trans. Circuits & Syst.*, 38, pp. 590–601, 1991.



James F. Whidborne was born in Zimbabwe in 1960. He received a B.A. degree in Engineering from Cambridge University in 1982, and an M.Sc. and a Ph.D. in Systems and Control from UMIST, Manchester, in 1987 and 1992 respectively. From 1982 to 1985 he worked in industry, and from 1986 to 1991, he was with the Control Systems Centre, UMIST. From 1991 to 1994, he held a position of Research Associate with the Department of Engineering, University of Leicester. From 1994 to 2003, he was a Lecturer, then Senior Lecturer with the Department of Mechanical Engineering, King's College London. James Whidborne is currently a Senior Lecturer in the Department of Aerospace Sciences at Cranfield University, UK. His research interests include optimal finite-precision controller implementations, multi-objective robust control design, fluid flow control and control of UAVs. He has nearly 100 research publications, including three books. He is a Chartered Engineer and a Member of the IEE and a Member of the IEEE. He currently serves on the Editorial Board of the International Journal of Systems Science and on the Editorial Board of the International Journal of Automation and Computing, and is Chair of the IEEE (UK&RI) Control Systems Chapter.



Da-Wei Gu graduated from the Department of Mathematics, Fudan University, Shanghai, China, in 1979, and received an M.Sc. degree in applied mathematics from Shanghai Jiao Tong University, China, in 1981, and a Ph.D. degree in control system theory from Department of Electrical Engineering, Imperial College of Science and Technology, London, U.K., in 1985. During 1981–1982, he was a lecturer in Shanghai Jiao Tong University. He was a postdoctoral research assistant in the Department of Engineering Science, Oxford University, U.K., from 1985 to 1989. In 1989, he was appointed to a University Lectureship in the Department of Engineering at Leicester University, UK, and has since been promoted to a Senior Lecturer and to a Reader at Leicester. His current research interests include robust and optimal control, optimization algorithms, control system computer-aided design and numerical perturbation analysis, with applications in aerospace and manufacturing industry. He has over 230 research publications, including two books. He is a Chartered Engineer and a Member of the Institution of Electrical Engineers.



Dr Jun Wu is currently a professor in National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Department of Control Science and Engineering at Zhejiang University, China. His research interests include finite word length implementation of digital controllers, robust control, networked control and model reduction. He has published over 50 papers in international journals and conference proceedings. He was awarded a Royal Society K.C. Wong fellowship in 2002 and the 1997/1998 IEE Heaviside Premium.



Sheng Chen received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982 and the Ph.D. degree in control engineering from the City University, London, U.K., in 1986. He joined the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in September 1999. He previously held research and academic appointments at the University of Sheffield, Sheffield, U.K., the University of Edinburgh, Edinburgh, U.K., and University of Portsmouth, Portsmouth, U.K. His recent research works include adaptive nonlinear signal processing, wireless communications, modeling and identification of nonlinear systems, neural networks and machine learning, finite-precision digital controller design, evolutionary computation methods, and optimization. He has published over 240 research papers. In the database of the world's most highly cited researchers in various disciplines, compiled by Institute for Scientific Information (ISI) of the USA, Dr. Chen is on the list of the highly cited researchers in the engineering category.