15th Triennial World Congress, Barcelona, Spain

OPTIMAL FINITE-PRECISION CONTROLLER AND FILTER IMPLEMENTATIONS USING FLOATING-POINT ARITHMETIC

James F Whidborne [*] **Da-Wei Gu** [**]

[*] *Department of Mechanical Engineering, King's College London, Strand, London WC2R 2LS, UK.*
email: `james.whidborne@kcl.ac.uk`
[**] *Department of Engineering, University of Leicester, Leicester LE1 7RH, U.K.*
email: `dag@leicester.ac.uk`

Abstract: In this paper, eigenvalue sensitivity measures are proposed that are suitable for assessing the fragility of digital controllers and filters implemented using floating-point arithmetic. Floating-point arithmetic parameter uncertainty is shown to be multiplicative. 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. The problem for the closed-loop case is solved using nonlinear programming. The problems are illustrated with a numerical example.

Keywords: finite-precision, digital controller, digital filter, eigenvalue sensitivity, floating-point arithmetic, fragility

1. INTRODUCTION

The reducing cost and increasing speed of computer hardware means that there is an increasing tendency for digital controller implementations to be implemented using machines which utilize floating-point arithmetic. Although the effects on the control system due to the finite precision resulting from the finite word-length have been extensively studied for fixed-point implementations, see Istepanian and Whidborne (2001) for a review, there has been little work looking explicitly at the finite-precision effects for floating-point digital controller implementations. Some exceptions include Rink and Chong (1979), Molchanov and Bauer (1995), Faris *et al.* (1998) and Williamson (2001). There has been more work in the signal processing area, for example Rao (1996).

It is known that some controller/filter realizations are very sensitive to small errors in the parameters, and these small errors may even lead to instability. The source of the parameter errors is the finite precision of the computing device. Such controller realizations can be described as *fragile* (Keel and Bhattacharryya, 1997). However, a linear 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. The effect of the finite precision is actually 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 of the finite precision on the system.

In the next section, floating-point arithmetic is discussed and shown to result in multiplicative perturba-

tions on the filter/controller parameters. In Section 3, an upper bound on the eigenvalue perturbation magnitudes is obtained. 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. A necessary condition for minimising the measure is obtained. In the subsequent section, non-linear programming is shown to be effective for solving the problem.

Notation

$\lfloor x \rfloor$ denotes the floor function, that is, the largest integer less than or equal to x
$A \circ B = \left[a_{ij} b_{ij}\right]$ denotes the Hadamard product of A and B
$A^{\mathscr{T}}$ denotes the transpose of a matrix A
$A^{\mathscr{H}}$ denotes the complex conjugate transpose of a matrix A
$\mathrm{vec}(A)$ denotes the column stacking operator of a matrix A
$\|A\|_F = \sum_{i,j} a_{ij}^2$ denotes the Frobenius norm of a matrix A

2. FLOATING-POINT REPRESENTATION

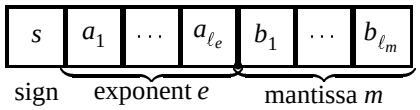

Fig. 1. Floating-point number representation

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

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

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

$$x = \pm m \times 2^e \tag{1}$$

where the mantissa is usually normalized so that $m \in [.5, 1)$. 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, $\mathscr{F}$, is given by

$$\mathscr{F} := \left\{ (-1)^s \left(0.5 + \sum_{i=1}^{\ell_m} b_i 2^{-(i+1)} \right) \times 2^e : \right.$$
$$\left. s \in \{0,1\}, b_i \in \{0,1\}, e \in \mathbb{N}_{[\underline{e}, \overline{e}]} \right\} \cup \{0\} \tag{2}$$

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

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

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

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

The quantization error, ε, is defined as

$$\varepsilon := |x - q(x)|. \tag{4}$$

It can be shown easily that the quantization error is bounded by $\varepsilon < |x| \, 2^{-(\ell_m+1)}$. Thus, when a number is implemented in finite-precision floating-point arithmetic, it may be perturbed to

$$q(x) = x(1+\delta), \quad |\delta| < \delta_{\max}. \tag{5}$$

where $\delta_{\max} = 2^{-(\ell_m+1)}$. Thus 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 bee proposed for fixed-point digital controller and filter implementations (Mantey, 1968; Gevers and Li, 1993; Li, 1998; Istepanian *et al.*, 1998; Whidborne *et al.*, 2001; Wu *et al.*, 2001).

Assume that a controller/filter realization $x = \mathrm{vec}(X)$ is implemented with floating-point arithmetic with finite precision, that is the actual realization will be $q(x)$. Then, from (5), 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 $\delta = \left[\delta_i\right]$.

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

$$|f(\tilde{x}) - f(x)| \leq \delta_{\max} \|g\| \|x\| + \left\| \mathscr{O}(\delta_{\max}^2) \right\| \tag{6}$$

where $|\delta_i| < \delta_{\max}$ for all i, $\mathscr{O}$ represents the second and higher order terms of the Taylor series and $g(x)$ is the gradient vector, *i.e.*,

$$g(x) := \frac{\partial f(x)}{\partial x} = \left[\frac{\partial f}{\partial x_i}\right]_x \tag{7}$$

evaluated at x.

Proof: Taking a first-order Taylor series approximation:

$$f(\tilde{x}) = f(x) + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)_x (\tilde{x}_i - x_i) + \mathscr{O}(\delta_{\max}^2) \tag{8}$$

Now, from (5), $\tilde{x}_i = x_i(1+\delta_i)$, so

$$f(\tilde{x}) - f(x) = \sum_{i=1}^{n_x} g_i(x) x_i \delta_i + \mathscr{O}(\delta_{\max}^2). \tag{9}$$

Hence

$$|f(\tilde{x}) - f(x)| \leq \sum_{i=1}^{n_x} |g_i(x)|\,|x_i|\,|\delta_i| + \| \mathscr{O}(\delta_{\max}^2) \| \tag{10}$$

$$< \delta_{\max} \sum_{i=1}^{n_x} |g_i(x)|\,|x_i| + \| \mathscr{O}(\delta_{\max}^2) \| \tag{11}$$

which, by the Cauchy-Schwartz inequality, gives

$$|f(\tilde{x}) - f(x)| < \delta_{\max} \| g(x) \| \| x \| + \| \mathscr{O}(\delta_{\max}^2) \|. \tag{12}$$

□ □

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 implementation.

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 is 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

$$\Phi = \left\| A_f \right\|_F^2 \sum_{k=1}^{n} w_k \Phi_k \tag{13}$$

where w_k is a non-negative real scalar weighting and

$$\Phi_k = \left\| \frac{\partial \lambda_k}{\partial A_f} \right\|_F^2 \tag{14}$$

where $\{\lambda_i : i = 1, \ldots, n\}$ represents the set of unique eigenvalues of A_f. The measure Φ is dependent upon the filter realization, that is, given $A_f = T^{-1}A_f^0T$,

$$\Phi(T) := \left\| T^{-1}A_f^0T \right\|_F^2 \sum_{k=1}^{n} w_k \Phi_k(T) \tag{15}$$

where, (Gevers and Li, 1993; Li, 1998),

$$\Phi_k(T) = \mathrm{tr}\left(R_k^{\mathscr{H}} T^{-\mathscr{T}} T^{-1} R_k\right) \mathrm{tr}\left(L_k^{\mathscr{H}} T T^{\mathscr{T}} L_k\right) \tag{16}$$

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

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

$$\Phi_{\min} = \min_{\substack{T\in\mathbb{R}^{n\times n} \\ \det(T)\neq 0}} \Phi(T) \tag{17}$$

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

$$\Phi_{\min} = \sum_{k=1}^{n} |\lambda_k|^2 \sum_{k=1}^{n} w_k \tag{18}$$

and

$$T_{\min} = \left(RWR^{\mathscr{H}}\right)^{1/2} V \tag{19}$$

where $R = \left[R_i\right]$ is the matrix of right eigenvectors of A_f^0, $W = \mathrm{diag}(w_1, \ldots, w_n)$ 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, pp137-138), it follows that $\Phi_k \geq 1$ with equality for all k if A_f is normal. From Horn and Johnson (1985, p101),

$$\left\| A_f \right\|_F^2 \geq \sum_{k=1}^{n} |\lambda_k|^2 \tag{20}$$

with equality if A_f is normal. Clearly, if A_f is normal, Φ is minimal and (18) holds. Theorem 6.2 of Gevers and Li (1993, p141) gives (19). □ □

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, p139).

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

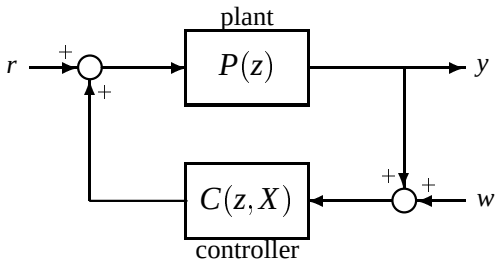

Fig. 2. Feedback control system

the controller be $C(z,X)$ where X is the parameterization 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

$$\begin{aligned}\bar{A}&=\begin{bmatrix}A_p+B_pD_cC_p & B_pC_c\\ B_cC_p & A_c\end{bmatrix},\\ &=\begin{bmatrix}A_p & 0\\ 0 & 0\end{bmatrix}+\begin{bmatrix}B_p & 0\\ 0 & I_n\end{bmatrix}\begin{bmatrix}D_c & C_c\\ B_c & A_c\end{bmatrix}\begin{bmatrix}C_p & 0\\ 0 & I_n\end{bmatrix},\\ &=M_0+M_1XM_2=\bar{A}(X),\end{aligned}\tag{21}$$

where

$$X:=\begin{bmatrix}D_c & B_c\\ C_c & A_c\end{bmatrix},\tag{22}$$

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

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

then any realization is given by

$$\begin{aligned}X&=\begin{bmatrix}I & 0\\ 0 & T\end{bmatrix}^{-1}\begin{bmatrix}D_c^0 & C_c^0\\ B_c^0 & A_c^0\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & T\end{bmatrix},&(24)\\ &=T_I^{-1}X_0T_I,&(25)\end{aligned}$$

where $T\in\mathbb{R}^{n\times n}$ is non-singular.

Let $R_k=(R_k^{\mathscr{T}}(1)\;\;R_k^{\mathscr{T}}(2))^{\mathscr{T}}$ and $L_k=(L_k^{\mathscr{T}}(1)\;L_k^{\mathscr{T}}(2))^{\mathscr{T}}$ be the right and left eigenvectors respectively for the kth 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 (22). Then, it can be shown (Li, 1998; Whidborne *et al.*, 2001) that

$$\left(\frac{\partial\lambda_k}{\partial A_c}\right)^{\mathscr{T}}=R_k(2)L_k^{\mathscr{H}}(2),\tag{26}$$

$$\left(\frac{\partial\lambda_k}{\partial B_c}\right)^{\mathscr{T}}=C_pR_k(1)L_k^{\mathscr{H}}(2),\tag{27}$$

$$\left(\frac{\partial\lambda_k}{\partial C_c}\right)^{\mathscr{T}}=R_k(2)L_k^{\mathscr{H}}(1)B_p,\tag{28}$$

$$\left(\frac{\partial\lambda_k}{\partial D_c}\right)^{\mathscr{T}}=C_pR_k(1)L_k^{\mathscr{H}}(1)B_p,\tag{29}$$

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

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

$$\Upsilon(X):=\|X\|_F^2\sum_{k=1}^{n+m}w_k\Upsilon_k\tag{30}$$

where w_k is a non-negative real scalar weighting and

$$\begin{aligned}\Upsilon_k=&\left\|\frac{\partial\lambda_k}{\partial A_c}\right\|_F^2+\left\|\frac{\partial\lambda_k}{\partial B_c}\right\|_F^2\\ &+\left\|\frac{\partial\lambda_k}{\partial C_c}\right\|_F^2+\left\|\frac{\partial\lambda_k}{\partial D_c}\right\|_F^2.\end{aligned}\tag{31}$$

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

$$\begin{aligned}\|X\|_F^2=&\operatorname{tr}(P^{-1}A_c^0PA_c^{0\mathscr{T}})+\operatorname{tr}(P^{-1}B_c^0B_c^{0\mathscr{T}})\\ &+\operatorname{tr}(PC_c^{0\mathscr{T}}C_c^0)+\operatorname{tr}(D_c^0D_c^{0\mathscr{T}}),\end{aligned}\tag{32}$$

where $P=TT^{\mathscr{T}}$ and, from (26) – (29), that

$$\begin{aligned}\Upsilon_k=&\operatorname{tr}(R_k^{0\mathscr{H}}(2)P^{-1}R_k^0(2))\operatorname{tr}(L_k^{0\mathscr{H}}(2)PL_k^0(2))\\ &+\alpha_k\operatorname{tr}(L_k^{0\mathscr{H}}(2)PL_k^0(2))\\ &+\beta_k\operatorname{tr}(R_k^{0\mathscr{H}}(2)P^{-1}R_k^0(2))+\alpha_k\beta_k,\end{aligned}\tag{33}$$

where

$$\alpha_k=\operatorname{tr}(R_k^{0\mathscr{H}}(1)C_p^{\mathscr{H}}C_pR_k^0(1)),\tag{34}$$

$$\beta_k=\operatorname{tr}(L_k^{0\mathscr{H}}(1)B_pB_p^{\mathscr{H}}L_k^0(1)).\tag{35}$$

Rearranging gives

$$\begin{aligned}\Upsilon(P)=&\Big(\operatorname{tr}(P^{-1}A_c^0PA_c^{0\mathscr{T}})+\operatorname{tr}(P^{-1}B_c^0B_c^{0\mathscr{T}})\\ &+\operatorname{tr}(PC_c^{0\mathscr{T}}C_c^0)+\operatorname{tr}(D_c^0D_c^{0\mathscr{T}})\Big)\\ &\times\Big(\sum_{k=1}^{n+m}\operatorname{tr}(P^{-1}M_{R_k})\operatorname{tr}(PM_{L_k})\\ &+\operatorname{tr}(PW_L)+\operatorname{tr}(P^{-1}W_R)+c\Big)\end{aligned}\tag{36}$$

where

$$M_{R_k}=w_k^{1/2}R_k^0(2)R_k^{0\mathscr{H}}(2)\tag{37}$$

$$M_{L_k}=w_k^{1/2}L_k^0(2)L_k^{0\mathscr{H}}(2)\tag{38}$$

$$W_L=L^0(2)\operatorname{diag}(w_1\alpha_1,\ldots,w_{n+m}\alpha_{n+m})L^{0\mathscr{H}}(2),\tag{39}$$

$$W_R=R^0(2)\operatorname{diag}(w_1\beta_1,\ldots,w_{n+m}\beta_{n+m})R^{0\mathscr{H}}(2),\tag{40}$$

are all Hermitian, and

$$c = \sum_{k=1}^{n+m} \alpha_k \beta_k. \tag{41}$$

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

$$\Upsilon_{\min} = \min_{\substack{P \in \mathbb{R}^{n\times n} \\ P = P^{\mathscr{T}} \geq 0}} \Upsilon(P) \tag{42}$$

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

Theorem 2. A necessary condition for the solution of Problem 2 is given by

$$\begin{aligned}
&\Big(\sum_{k=1}^{n+m} \operatorname{tr}(P^{-1}M_{R_k}) \operatorname{tr}(PM_{L_k}) \\
&\qquad + \operatorname{tr}(PW_L) + \operatorname{tr}(P^{-1}W_R) + c\Big) \\
&\qquad \times \Big(A_c^0 P^{-1} A_c^{0,\mathscr{T}} + C_c^{0,\mathscr{T}} C_c^0\Big) \\
&\quad + \Big(\operatorname{tr}(P^{-1}A_c^0 P A_c^{0,\mathscr{T}}) + \operatorname{tr}(P^{-1}B_c^0 B_c^{0,\mathscr{T}}) \\
&\qquad + \operatorname{tr}(PC_c^{0,\mathscr{T}} C_c^0) + \operatorname{tr}(D_c^0 D_c^{0,\mathscr{T}})\Big) \\
&\qquad \times \Big(\sum_{k=1}^{n+m} \operatorname{tr}(P^{-1}M_{R_k}) M_{L_k} + W_L\Big) \\
&= P^{-1}\Bigg(\Big(\sum_{k=1}^{n+m} \operatorname{tr}(P^{-1}M_{R_k}) \operatorname{tr}(PM_{L_k}) \\
&\qquad + \operatorname{tr}(PW_L) + \operatorname{tr}(P^{-1}W_R) + c\Big) \\
&\qquad \times \Big(A_c^0 P A_c^{0,\mathscr{T}} + B_c^0 B_c^{0,\mathscr{T}}\Big) \\
&\quad + \Big(\operatorname{tr}(P^{-1}A_c^0 P A_c^{0,\mathscr{T}}) + \operatorname{tr}(P^{-1}B_c^0 B_c^{0,\mathscr{T}}) \\
&\qquad + \operatorname{tr}(PC_c^{0,\mathscr{T}} C_c^0) + \operatorname{tr}(D_c^0 D_c^{0,\mathscr{T}})\Big) \\
&\qquad \times \Big(\sum_{k=1}^{n+m} \operatorname{tr}(PM_{L_k}) M_{R_k} + W_R\Big)\Bigg) P^{-1}.
\end{aligned} \tag{43}$$

Proof: Differentiating Υ with respect to P gives

$$\begin{aligned}
\frac{\partial \Upsilon}{\partial P} &= \Big(\sum_{k=1}^{n+m} \operatorname{tr}(P^{-1}M_{R_k}) \operatorname{tr}(PM_{L_k}) \\
&\qquad + \operatorname{tr}(PW_L) + \operatorname{tr}(P^{-1}W_R) + c\Big) \\
&\qquad \times \Big(A_c^0 P^{-1} A_c^{0,\mathscr{T}} - P^{-1} A_c^0 P A_c^{0,\mathscr{T}} P^{-1} \\
&\qquad - P^{-1} B_c^0 B_c^{0,\mathscr{T}} P^{-1} + C_c^{0,\mathscr{T}} C_c^0\Big) \\
&\quad + \Big(\operatorname{tr}(P^{-1}A_c^0 P A_c^{0,\mathscr{T}}) + \operatorname{tr}(P^{-1}B_c^0 B_c^{0,\mathscr{T}}) \\
&\qquad + \operatorname{tr}(PC_c^{0,\mathscr{T}} C_c^0) + \operatorname{tr}(D_c^0 D_c^{0,\mathscr{T}})\Big) \\
&\quad \times \Big(\sum_{k=1}^{n+m} \operatorname{tr}(P^{-1}M_{R_k}) M_{L_k}^{\mathscr{T}} - \operatorname{tr}(PM_{L_k}) P^{-1} M_{R_k}^{\mathscr{T}} P^{-1} \\
&\qquad + W_L^{\mathscr{T}} - P^{-1} W_R^{\mathscr{T}} P^{-1}\Big).
\end{aligned} \tag{44}$$

Any solution P of (42) must necessarily satisfy $\frac{\partial \Upsilon}{\partial P} = 0$, so (43) is obtained. □ □

There does not appear to be an analytic solution to (43), hence nonlinear programming is used to find a solution to Problem 2. From the optimal $P_{\min}$, a corresponding optimal transformation matrix $T_{\min}$ where $P_{\min} = T_{\min} T_{\min}^{\mathscr{T}}$ can be constructed as (Li *et al.*, 1992) $T_{\min} = P_{\min}^{1/2} V$ for any orthogonal matrix V.

6. EXAMPLE

The following numerical example is taken from Gevers and Li (1993, pp236-237). The discrete time system to be controlled is given by

$$A_p = \begin{bmatrix} 3.7156 & -5.4143 & 3.6525 & -0.9642 \\ 1.000 & 0 & 0 & 0 \\ 0 & 1.000 & 0 & 0 \\ 0 & 0 & 1.000 & 0 \end{bmatrix}, \tag{45}$$

$$B_p = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^{\mathscr{T}}, \tag{46}$$

$$C_p = \begin{bmatrix} 0.1116 & 0.0043 & 0.1088 & 0.0014 \end{bmatrix} \times 10^{-5}. \tag{47}$$

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

$$0.9844 \pm 0.0357j, 0.9643 \pm 0.0145j, \tag{48}$$

and a state observer is designed with poles located at

$$0.7152 \pm 0.6348j, 0.3522 \pm 0.2857j. \tag{49}$$

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

$$\begin{aligned}
A_c^0 &= A_p + B_p C_c^0 - B_p C_c^0 \\
&= \begin{bmatrix} 2.6743 & -5.7443 & 2.5096 & -0.9176 \\ 0.2877 & -0.0273 & -0.6947 & -0.0088 \\ -0.3377 & 0.9871 & -0.3294 & -0.0042 \\ -0.0830 & -0.0032 & 0.9190 & -0.0010 \end{bmatrix}, \\
B_c^0 &= \begin{bmatrix} 1.0963 & 0.6385 & 0.3027 & 0.0744 \end{bmatrix}^{\mathscr{T}} \times 10^6, \\
C_c^0 &= \begin{bmatrix} 0.1818 & -0.2831 & 0.0500 & 0.0617 \end{bmatrix}, \\
D_c^0 &= 0.
\end{aligned}$$

The weights are set to $w_i = (1-\lambda_{\max})/(1-|\lambda_i|)$ where $\lambda_{\max} = \max_i\{|\lambda_i|\}$. 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):

$$\begin{aligned}
A_c &= \begin{bmatrix} 0.6194 & -0.1992 & -0.0835 & -0.1265 \\ 0.1346 & 0.6052 & -0.2297 & 0.0171 \\ 0.0508 & 0.1650 & 0.5315 & -0.2813 \\ 0.2047 & 0.0653 & 0.2218 & 0.5605 \end{bmatrix}, \\
B_c &= \begin{bmatrix} 0.6508 & 0.0048 & 2.0020 & 0.2961 \end{bmatrix}^{\mathscr{T}} \times 10^6, \\
C_c &= \begin{bmatrix} 0.11000.0222 & -0.0142 & -0.0168 \end{bmatrix}.
\end{aligned}$$

The closed-loop pole sensitivity for the initial realization is $\Upsilon = 2.3396 \times 10^{10}$ and for the open-loop optimal realization is $\Upsilon = 2.1720 \times 10^9$. Using the MATLAB routine `balreal.m`, a balanced realization was obtained (to 4 decimal places):

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

$$B_c = \begin{bmatrix} 203.1819 & -63.5703 & 32.0424 & -4.1143 \end{bmatrix}^{\mathscr{T}},$$

$$C_c = \begin{bmatrix} 203.1819 & 63.5703 & 32.0424 & 4.1143 \end{bmatrix}.$$

The closed-loop pole sensitivity for the balanced realization is $\Upsilon = 1.3528 \times 10^6$.

The MATLAB routine `fminsearch.m` was used to solve Problem 2. An optimal closed-loop pole sensitivity value of $\Upsilon_{\min} = 5.3002 \times 10^3$ was obtained with a realization (to 4 decimal places):

$$A_c = \begin{bmatrix} 0.6841 & -0.0813 & 2.4433 & -0.8279 \\ 0.5783 & 0.9234 & -1.7750 & 2.7260 \\ -0.0096 & 0.0006 & -0.2040 & -0.1434 \\ 0.0765 & -0.0035 & 3.0433 & 0.9131 \end{bmatrix},$$

$$B_c = \begin{bmatrix} 113.9572 & -50.4959 & -64.7777 & 160.9055 \end{bmatrix}^{\mathscr{T}},$$

$$C_c = \begin{bmatrix} 36.9396 & -0.9027 & -134.6869 & 157.0591 \end{bmatrix}.$$

7. DISCUSSION AND CONCLUSIONS

In previous works, the eigenvalue sensitivity approach to obtaining 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, it is shown in this paper that 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. It is shown that the optimal open-loop and the balanced realization result in high closed-loop pole sensitivities. The closed-loop pole sensitivity of the balanced controller realization is reduced by three orders of magnitude using the proposed method.

8. REFERENCES

Faris, D., T. Pare, A. Packard, K.A. Ali and J.P. How (1998). Controller fragility: What's all the fuss?. In: *Proc. Annual Allerton Conference on Communication Control and Computing*. Vol. 36. Monticello, Illinois. pp. 600–609.

Gevers, M. and G. Li (1993). *Parametrizations in Control, Estimations and Filtering Problems: Accuracy Aspects*. Springer-Verlag. Berlin.

Horn, R.A. and C.R. Johnson (1985). *Matrix Analysis*. Cambridge University Press. Cambridge, U.K.

Istepanian, R.H., G. Li, J. Wu and J. Chu (1998). Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds. *IEE Proc. Control Theory and Appl.* **145**(5), 472–478.

Istepanian, R.S.H. and J.F. Whidborne (2001). Finite-precision computing for digital control systems: Current status and future paradigms. In: *Digital Controller Implementation and Fragility: A Modern Perspective* (R.S.H. Istepanian and J.F. Whidborne, Eds.). Chap. 1, pp. 1–12. Springer-Verlag. London, UK.

Keel, L.H. and S.P. Bhattacharryya (1997). Robust, fragile, or optimal?. *IEEE Trans. Autom. Control* **42**(8), 1098–1105.

Li, G. (1998). On the structure of digital controllers with finite word length consideration. *IEEE Trans. Autom. Control* **43**(5), 689–693.

Li, G., B.D.O. Anderson, M. Gevers and J.E. Perkins (1992). Optimal FWL design of state space digital systems with weighted sensitivity minimization and sparseness consideration. *IEEE Trans. Circuits & Syst. II* **39**(5), 365–377.

Mantey, P.E. (1968). Eigenvalue sensitivity and state-variable selection. *IEEE Trans. Autom. Control* **13**(3), 263–269.

Molchanov, A.P. and P.H. Bauer (1995). Robust stability of digital feedback control systems with floating point arithmetic. In: *Proc. 34th IEEE Conf. Decision Contr.*. New Orleans, LA. pp. 4251–4258.

Rao, B.D. (1996). Roundoff noise in floating point digital filters. *Control and Dynamic Systems* **75**, 79–103.

Rink, R.E. and H.Y. Chong (1979). Performance of state regulator systems with floating point computation. *IEEE Trans. Autom. Control* **24**, 411–421.

Whidborne, J.F., R.S.H. Istepanian and J. Wu (2001). Reduction of controller fragility by pole sensitivity minimization. *IEEE Trans. Autom. Control* **46**(2), 320–325.

Williamson, D. (2001). Implementation of a class of low complexity low sensitivity digital controllers using adaptive fixed-point arithmetic. In: *Digital Controller Implementation and Fragility: A Modern Perspective* (R.S.H. Istepanian and J.F. Whidborne, Eds.). Chap. 4, pp. 43–74. Springer-Verlag. Godalming, UK.

Wu, J., S. Chen, G. Li, R.H. Istepanian and J. Chu (2001). An improved closed-loop stability related measure for finite-precision digital controller realizations. *IEEE Trans. Autom. Control* **46**(7), 1662–1666.