International Journal of Systems Science. 10 June, 2005, Vol 36 Iss 7, p405-413

Optimal Controller and Filter Realisations using Finite-precision, Floating-point Arithmetic

James F. Whidborne
Department of Aerospace Sciences, Cranfield University
Bedfordshire MK43 0AL, U.K.
email:j.f.whidborne@cranfield.ac.uk
Da-Wei Gu
Department of Engineering, University of Leicester,
Leicester LE1 7RH, U.K.
email: dag@leicester.ac.uk
Jun Wu
National Key Laboratory of Industrial Control Technology,
Institute of Industrial Process Control,
Zhejiang University, Hangzhou, 310027, P. R. China.
email: jwu@iipc.zju.edu.cn
Sheng Chen
Department of Electronics and Computer Science, University of Southampton,
Highfield, Southampton SO17 1BJ, U.K.
email: sqc@ecs.soton.ac.uk

February 14, 2005

Abstract

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 realisation. Problems of obtaining the optimal realisation 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 nonlinear programming. The problems are illustrated with a numerical example.

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 quantisation errors resulting from the quantisation 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 quantisation. 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, e.g.) that quantisation and rounding effects with floating point arithmetic is of a different nature to that of fixed point. Fixed-point quantisation error results in additive noise independent of the signal, but with floating-point arithmetic, the quantisation error is correlated with the signal that is being quantised. 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 optimisation of finite-precision filter and controller implementations needs to take the arithmetic into account.

The quantisation error effect on digital filters resulting from the finite precision using floating-point arithmetic has been fairly extensively studied over the last 4 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, Smith and Joseph, 1997; Tsai, 1997; Ko and Bitmead, 2004), see Kontro, Kalliojärvi and Neuvo (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 quantisation errors have been analysed by Rink and Chong (1979*a*,*b*), and by Vanwingerden and De Koning (1984) for optimal controllers. Miller, Mousa and Michel

(1988) have also analysed the quantisation errors, but notably also include the inter-sample behaviour. A method to design optimal controllers that minimise the quantisation 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 realisations 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 realisations can be described as *fragile* (Keel and Bhattacharryya, 1997). However, a dynamical system has an infinite number of equivalent realisations. 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 realisation for any non-singular matrix T. It so happens that the effect of the finite precision is partially dependent upon the realisation. Thus, in order to ensure a non-fragile implementation, it is of interest to know the realisation, or matrix T, which minimises the effect on the system of the finite precision.

One approach to obtaining non-fragile realisations is to minimise the sensitivity of the system eigenvalues. This approach has been extensively investigated for fixed-point realisations. 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 realisations based on a norm for the open-loop eigenvalue sensitivities. The case of the closed-loop system eigenvalue sensitivity for state-space controller realisations was first considered by Li (1998) and has subsequently been thoroughly investigated (Istepanian et al., 1998; Chen et al., 1999; Wu, Istepanian and Chen, 1999; Istepanian et al., 2000; Wu et al., 2000; Whidborne, Istepanian and Wu, 2001).

In this paper, a simple eigenvalue sensitivity measure is considered for both filter and controller realisations. 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 minimisation 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 realisation established. However additive perturbations on the coefficients are assumed, and this index is actually an upper bound on an index proposed by Whidborne, Istepanian and Wu (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 realisations is solved. In Section 5, a similar measure for closed-loop controller implementations is proposed. Nonlinear 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.

Notations

$\lfloor x \rfloor$ denotes the floor function, that is, the largest integer less than or equal to x
$A \circ B = [a_{ij} b_{ij}]$ denotes the Hadamard product of A and B
$A^{\mathcal{T}}$ denotes the transpose of a matrix A
$A^{\mathcal{H}}$ denotes the complex conjugate transpose of a matrix A
$\mathrm{vec}(A)$ denotes the column stacking operator of a matrix A
$\| A \|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ denotes the Frobenius norm of a matrix A
$A^{1/2}$ denotes, for a matrix $A \geq 0$, the unique symmetric matrix satisfying $A^{1/2} \geq 0$ and $A^{1/2} A^{1/2} = A$
$\mathbb{C}$ denotes the set of all complex numbers
$\mathbb{R}$ denotes the set of all real numbers
$\mathbb{Z}$ denotes the set of all integers
$\mathcal{O}(x)$ denotes "is of order x"

2 Floating-point representation

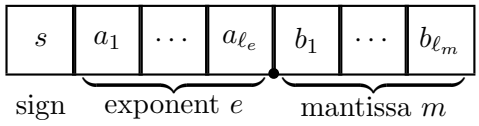

Figure 1: Floating-point number representation

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

$$x = (-1)^s \times m \times 2^e \tag{1}$$

where the mantissa is usually normalised 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, $\mathcal{F}$, is given by

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

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.

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, $q : \mathbb{R} \to \mathcal{F}$, as

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

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

The rounding error, ε, is defined as

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

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

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

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

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

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 bee proposed for fixed-point digital controller and filter implementations (Mantey, 1968; Gevers and Li, 1993; Li, 1998; Istepanian et al., 1998; Whidborne, Istepanian and Wu, 2001; Wu et al., 2001).

Assume that a controller/filter realisation $x = \mathrm{vec}(X)$ is implemented with floating-point arithmetic with finite precision, that is the actual realisation 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 realisation 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*

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

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

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

evaluated at x.

Proof. Taking a first-order Taylor series approximation:

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

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

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

Hence

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

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

which, by the Cauchy-Schwartz inequality, gives

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

□

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 realisations

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 realisation of $F(z)$. The realisations of $F(z)$ are not unique, if $(A_f^0, B_f^0, C_f^0, D_f^0)$ is a realisation 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 realisation 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

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

where w_k is a non-negative real scalar weighting and

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

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

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

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

$$\Phi_k(T) = \mathrm{tr}\left(R_k^{\mathcal{H}} T^{-\mathcal{T}} T^{-1} R_k\right) \mathrm{tr}\left(L_k^{\mathcal{H}} T T^{\mathcal{T}} L_k\right) \tag{17}$$

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 realisation* (A_f^0, B_f^0, C_f^0, D_f^0), *calculate*

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

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*

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

and

$$T_{\min} = \left(RWR^{\mathcal{H}}\right)^{1/2} V \tag{20}$$

where $R = [R_i]$ *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),

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

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, p141) gives (20). □

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 realisations

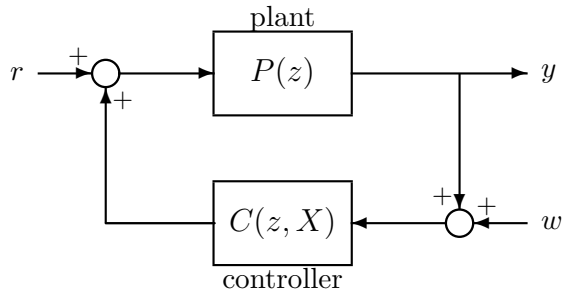

Figure 2: Feedback control system

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 parametrisation 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

$$\begin{aligned}\bar{A} &= \begin{bmatrix} A_p + B_pD_cC_p & B_pC_c \\ B_cC_p & A_c \end{bmatrix} \\ &= \begin{bmatrix} A_p & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B_p & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix} \begin{bmatrix} C_p & 0 \\ 0 & I_n \end{bmatrix}, \\ &=: M_0 + M_1XM_2 = \bar{A}(X), \end{aligned} \tag{22}$$

where

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

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

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

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

then any realisation is given by

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

$$=: T_I^{-1}X_0T_I, \tag{26}$$

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

Let $R_k = (R_k^{\mathcal{T}}(1)\ R_k^{\mathcal{T}}(2))^{\mathcal{T}}$ and $L_k = (L_k^{\mathcal{T}}(1)\ L_k^{\mathcal{T}}(2))^{\mathcal{T}}$ 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 (23). Then, it can be shown (Li, 1998; Whidborne, Istepanian and Wu, 2001) that

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

$$\left(\frac{\partial \lambda_k}{\partial B_c}\right)^{\mathcal{T}} = C_pR_k(1)L_k^{\mathcal{H}}(2), \tag{28}$$

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

$$\left(\frac{\partial \lambda_k}{\partial D_c}\right)^{\mathcal{T}} = C_pR_k(1)L_k^{\mathcal{H}}(1)B_p, \tag{30}$$

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

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

where w_k is a non-negative real scalar weighting and

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

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

$$\| X \|_F^2 = \mathrm{tr}(P^{-1} A_c^0 P A_c^{0\mathcal{T}}) + \mathrm{tr}(P^{-1} B_c^0 B_c^{0\mathcal{T}}) + \mathrm{tr}(P C_c^{0\mathcal{T}} C_c^0) + \mathrm{tr}(D_c^0 D_c^{0\mathcal{T}}), \tag{33}$$

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

$$\begin{aligned} \Upsilon_k = \mathrm{tr}(R_k^{0\mathcal{H}}(2) P^{-1} R_k^0(2)) \, \mathrm{tr}(L_k^{0\mathcal{H}}(2) P L_k^0(2)) + \alpha_k \, \mathrm{tr}(L_k^{0\mathcal{H}}(2) P L_k^0(2)) \\ + \beta_k \, \mathrm{tr}(R_k^{0\mathcal{H}}(2) P^{-1} R_k^0(2)) + \alpha_k \beta_k, \end{aligned} \tag{34}$$

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

$$\begin{aligned} \Upsilon(P) = \Big(\mathrm{tr}(P^{-1} A_c^0 P A_c^{0\mathcal{T}}) + \mathrm{tr}(P^{-1} B_c^0 B_c^{0\mathcal{T}}) + \mathrm{tr}(P C_c^{0\mathcal{T}} C_c^0) + \mathrm{tr}(D_c^0 D_c^{0\mathcal{T}}) \Big) \\ \times \Big(\sum_{k=1}^{n+m} \mathrm{tr}(P^{-1} M_{R_k}) \, \mathrm{tr}(P M_{L_k}) + \mathrm{tr}(P W_L) + \mathrm{tr}(P^{-1} W_R) + c \Big) \end{aligned} \tag{35}$$

where

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

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

$$W_L = L^0(2) \, \mathrm{diag}(w_1 \alpha_1, \ldots, w_{n+m} \alpha_{n+m}) L^{0\mathcal{H}}(2), \tag{38}$$

$$W_R = R^0(2) \, \mathrm{diag}(w_1 \beta_1, \ldots, w_{n+m} \beta_{n+m}) R^{0\mathcal{H}}(2), \tag{39}$$

are all Hermitian, and

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

Problem 2. *Given an initial realisation* (A_c^0, B_c^0, C_c^0, D_c^0), *calculate*

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

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

Remark 2. *The function $\Upsilon(P)$ is everywhere differentiable over the set $\{\Upsilon(P) : P = P^{\mathcal{T}} > 0\}$. Hence it is proposed that nonlinear programming is used to find local solutions to Problem 2. The problem of finding a global solution remains open.*

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

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

Thus a search can be made over the set $\left\{\left[\begin{smallmatrix} g_a \\ g_b \end{smallmatrix}\right] : g_a \in \mathbb{R}^{(n-1)n/2}, g_b \in \mathbb{R}^n_+\right\}$.

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

$$T_{\min} = G_{\min}V. \tag{42}$$

This provides an extra degree of freedom which could be utilised to find, for example, sparse realisations (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

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

$$B_p = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^{\mathcal{T}}, \tag{44}$$

$$C_p = \begin{bmatrix} 0.1116 & 0.0043 & 0.1088 & 0.0014 \end{bmatrix} \times 10^{-5}. \tag{45}$$

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

$$0.9844 \pm 0.0357j, 0.9643 \pm 0.0145j, \tag{46}$$

and a state observer is designed with poles located at

$$0.7152 \pm 0.6348j, 0.3522 \pm 0.2857j. \tag{47}$$

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

$$A_c^0 = A_p + B_p C_c^0 - B_c^0 C_p \tag{48}$$

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

$$B_c^0 = \begin{bmatrix} 1.0963 & 0.6385 & 0.3027 & 0.0744 \end{bmatrix}^{\mathcal{T}} \times 10^6, \tag{50}$$

$$C_c^0 = \begin{bmatrix} 0.1818 & -0.2831 & 0.0500 & 0.0617 \end{bmatrix}, \tag{51}$$

$$D_c^0 = 0. \tag{52}$$

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 realisation 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 realisation (to 4 decimal places):

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

$$B_c = \begin{bmatrix} 0.6508 & 0.0048 & 2.0020 & 0.2961 \end{bmatrix}^{\mathcal{T}} \times 10^6, \tag{54}$$

$$C_c = \begin{bmatrix} 0.1100 & 0.0222 & -0.0142 & -0.0168 \end{bmatrix}. \tag{55}$$

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

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

$$B_c = \begin{bmatrix} 203.1819 & -63.5703 & 32.0424 & -4.1143 \end{bmatrix}^{\mathcal{T}}, \tag{57}$$

$$C_c = \begin{bmatrix} 203.1819 & 63.5703 & 32.0424 & 4.1143 \end{bmatrix}. \tag{58}$$

The closed-loop pole sensitivity for the balanced realisation 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 about 30 minutes to converge. An optimal closed-loop pole sensitivity value of $\Upsilon_{\min} = 4.3366 \times 10^8$ was obtained with a realisation (to 4 decimal places):

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

$$B_c = \begin{bmatrix} 129.2367 & -137.2672 & 56.4560 & -23.7868 \end{bmatrix}^{\mathcal{T}}, \tag{60}$$

$$C_c = \begin{bmatrix} 155.1427 & -119.5560 & 32.1475 & 1.0368 \end{bmatrix}. \tag{61}$$

7 Discussion and conclusions

In previous works, the eigenvalue sensitivity approach to obtain optimal digital filter and controller realisations 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 realisations; and the work of Whidborne, Istepanian and Wu (2001) is extended to obtain optimal floating-point digital controller realisations.

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

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*,

1997, 44(11):952–955.

S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures", *IEEE Trans. Autom. Control*, 1999, 44(11):2149–2153.

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*. 2001, 74(15):1501–1509.

D. Faris, T. Pare, A. Packard, K.A. Ali and J.P. How, "Controller Fragility: What's All The Fuss?", *Proc. Annual Allerton Conference on Communication Control and Computing*, 1998. Vol. 36 Monticello, Illinois: 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*, 1989, 34(12):1319–1324.

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.*, 1998, 145(5):472–478.

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, pp. 1–12, 2001.

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.*, 2000, 31(4):429–438.

K. Kalliojärvi and J. Astola, "Required coefficient word length in floating-point and logarithmic digital filters", *IEEE Sig. Proc. Letters*, 1994, 1(3):52–54.

E.P.F. Kan and J.K. Aggarwal, "Error analysis of digital filter employing floating-point arithmetic", *IEEE Trans. Circuit Theory*, 1971, 18(6):678–686.

T. Kaneko and B. Liu, "Effect of coefficient rounding in floating-point digital filters", *IEEE Trans. Aerosp. Electron. Syst.*, 1971, 7(5):995–1003.

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

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*, 2004, 34(4):1923–1932.

S. Ko and R. R. Bitmead, "Covariance calculation for floating-point state-space realizations", *IEEE Trans. Sig. Proc.*, 2004, 52(12):3370–3377.

J. Kontro, K. Kalliojärvi and Y. Neuvo, "Floating-point arithmetic in signal processing", *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.*, 1975, 22(12):927–936.

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

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*, 1992, 39(5):365–377.

B. Liu, "Effects of finite word-length on the accuracy of digital filters — a review", *IEEE Trans. Circuit Theory*, 1971, 18(6):670–677.

B. Liu and T. Kaneko, "Error analysis of digital filters realised with floating-point arithmetic", *Proc. IEEE*, 1969, 57(10):1735–1747.

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

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*, 1988, 33:698–704.

A.P. Molchanov and P.H. Bauer, "Robust stability of digital feedback control systems with floating point arithmetic", *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.*, 1976, 23:551–562.

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

R.E. Rink and H.Y. Chong, "Covariance equation for a floating-point regulator system", *IEEE Trans. Autom. Control*, 1979*a*, 24:980 – 982.

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

I.W. Sandberg, "Floating-point-roundoff accumulation in digital-filter realizations", *Bell Syst. Tech. J.*, 1967, 46(8):1775–1791.

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*, 1992, 39(2):90–98.

C. M. Tsai, "Floating-point roundoff noises of first- and second-order sections in parallel form digital filters", *IEEE Trans. Circuits & Syst. II*, 1997, 44(9):774–779.

A.J.M. Vanwingerden and W.L. De Koning, "The influence of finite word-length on digital optimal-control", *IEEE Trans. Autom. Control*, 1984, 29(5):385–391.

J.F. Whidborne and D.-W. Gu, "Optimal finite-precision controller and filter implementations using floating-point arithmetic", *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*, 2001, 46(2):320–325.

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*, 1999, 72(15):1331–1342.

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*, 2000, 45(8):1550 – 1554.

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*, 2001, 46(7):1662–1666.

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*, 2003, 48(5):816–822.

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*, 2004, 77(5):427–440.

B. Zeng and Y. Neuvo, "Analysis of floating point roundoff errors using dummy multiplier coefficient sensitivities", *IEEE Trans. Circuits & Syst.*, 1991, 38:590–601.