Optimum structures of digital controllers in sampled-data systems: a roundoff noise analysis

G. Li, J. Wu, S. Chen and K.Y. Zhao

Abstract: The effect of roundoff noise in a digital controller is analysed for a sampled-data system in which the digital controller is implemented in a state-space realisation. A new measure, called averaged roundoff noise gain, is derived. Unlike the traditionally used measure, where the analysis is performed based on an equivalent digital control system, this newly defined averaged roundoff noise gain allows one to take consideration of the inter-sample behaviour. It is shown that this measure is a function of the state-space realisation. Noting the fact that the state-space realisations of a digital controller are not unique, the problem of optimum controller structure is to identify those realisations that minimise the averaged roundoff noise gain subject to the l_2-scaling constraint which is for preventing the signals in the controller from overflow. An analytical solution to the problem is presented and a design example is given. Both theoretical analysis and simulation results show that the optimum controller realisations obtained with the proposed approach are superior to those obtained with the traditional analysis based on a digital control system.

1 Introduction

A sampled-data system (see Fig. 1) consists of a continuous-time plant $P(s)$ and a sampled-data controller which is composed of a sampler (A/D converter) S, the digital controller $C_d(z)$ to be designed and a hold (D/A converter) H. A digital controller is usually obtained by one of the following two ways: the first one is to design the controller in the continuous-time domain and then perform a digital implementation of the controller, while the second is to design the digital controller based on a discretised model of the plant. The designed digital controller has to be implemented with a digital device such as a digital control processor. Due to the finite word length (FWL) effects, the actually implemented controller is different from the designed one. Therefore, the actual performance of the system may be very different from the desired one. Generally speaking, there are two types of FWL errors in the digital controller. The first is perturbation of controller parameters implemented with FWL while the second is the rounding errors that occur in arithmetic operations. Typically, effects of these two types of errors are investigated separately.

The effects of the first type of FWL errors are classically studied with a transfer function sensitivity measure. In [1, 2], the analysis was performed based on the discrete-time counterpart of the sampled-data system. The corresponding sensitivity measure does not take the inter-sample behaviour into account. To overcome this, Madievski *et al.* [3] derived a sensitivity measure based on a hybrid operator (transfer function) of the sampled-data system. The corresponding optimal realisation problem was made tractable with a two-step procedure: very fast sampling at a multiple of the sampling frequency followed by 'blocking' or 'lifting' to achieve a single-rate (discrete-time) system. The stability of the sampled-data system may be lost due to the FWL errors of the digital controller parameters, which are not considered when the digital controller is designed. Recently, the effects of the parameter errors have been investigated with some stability robustness related measures such as the one based on the complex stability radius [4, 5] and those based on pole sensitivity (see, e.g. [6–10]).

The second type of FWL error is usually measured with the so-called roundoff noise gain. The effects of roundoff noise have been well studied in digital signal processing, particularly in digital filter implementation (see e.g. [11–13]). However, it was not until the late 1980s that the problem of optimal digital controller realisations minimising the roundoff noise gain was addressed. In [14], the 'optimal' controller realisation was computed with the loop opened. This realisation is obviously not optimal in the sense that it does not minimise the roundoff noise in the closed-loop system. A roundoff noise gain was derived for a control system with state-estimate feedback controller and the corresponding optimal realisation problem was solved in [1]. The effect of FWL errors of the regulator parameter on the LQG performance was investigated in [15]. For the roundoff error effect on the same control

IEE Proceedings online no. 20020397

DOI: 10.1049/ip-cta:20020397

Paper first received 15th November 2000 and in revised form 8th March 2002

G. Li is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

J. Wu is with the National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, People's Republic of China

S. Chen is with the Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

K.Y. Zhao is with the Institute of Electrical and Control Engineering, Qingdao University, People's Republic of China

strategy, the optimal FWL-LQG design problem was studied and a sub-optimal solution was provided in [16], while the optimal solution was obtained by Liu *et al.* [17]. It should be pointed out that a common feature of these results is that the plant is assumed to be in discrete-time form and hence so is the closed-loop. In most applications, the system is hybrid, i.e. the digital controller is used to control a continuous-time plant. Applying these results directly to such a sampled-data system implies neglecting the inter-sample system behaviour and particularly the inter-sample ripple, which may degrade the actual performance of the control system. The main objective in this paper is to investigate the effect of the roundoff noise in the digital controller and to identify the optimum controller realisations for a sampled-data system.

2 Roundoff noise analysis

Throughout the paper, a bold type symbol denotes a vector or matrix with appropriate dimension. It is well known that the digital controller $C_d(z)$ can be implemented with its state-space equations:

$$\left.\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}u_k\\ y_k &= \mathbf{C}\mathbf{x}_k + du_k\end{aligned}\right\} \tag{1}$$

where $u_k = u(kT_s)$, T_s is the sampling period, $\mathbf{A} \in \mathcal{R}^{n_c \times n_c}$, $\mathbf{B} \in \mathcal{R}^{n_c \times 1}$, $\mathbf{C} \in \mathcal{R}^{1 \times n_c}$ and $d \in \mathcal{R}$. $\mathbf{R} \triangleq (\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$ is called a realisation of $C_d(z)$, satisfying

$$C_d(z) = d + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{2}$$

Denote $\mathcal{S}_{C_d}$ as the set of all realisations $(\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$. It should be pointed out that $\mathcal{S}_{C_d}$ is an infinite set. In fact, if $\mathbf{R}_0 = (\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, d) \in \mathcal{S}_{C_d}$, $\mathcal{S}_{C_d} = \{(\mathbf{A}, \mathbf{B}, \mathbf{C}, d)\}$ is characterised by

$$\mathbf{A} = \mathbf{T}^{-1}\mathbf{A}_0\mathbf{T}, \quad \mathbf{B} = \mathbf{T}^{-1}\mathbf{B}_0, \quad \mathbf{C} = \mathbf{C}_0\mathbf{T}, \tag{3}$$

where $\mathbf{T} \in \mathcal{R}^{n_c \times n_c}$ is any non-singular matrix. Usually, such a $\mathbf{T}$ is called a similarity transformation. Once an initial realisation $\mathbf{R}_0$ is given, different controller realisations correspond to different similarity transformations $\mathbf{T}$.

It should be pointed out that the state-space model (1) is the digital controller implemented with infinite precision. Though there exist different state-space realisations, they yield exactly the same performance – the desired one. In practice, however, a designed digital controller has to be implemented with finite precision. Assuming a fixed-point implementation of digital controllers, then a more practical digital controller model is

$$\left.\begin{aligned}\mathbf{x}^*_{k+1} &= \mathbf{A}Q[\mathbf{x}^*_k] + \mathbf{B}Q[u^*_k]\\ y^*_k &= \mathbf{C}Q[\mathbf{x}^*_k] + dQ[u^*_k]\end{aligned}\right\} \tag{4}$$

where $Q[p]$ is the quantiser that rounds p to B_s bits in fractional part. In this Section, we will investigate the effects of signal rounding off in the digital controller implemented with the model (4) on the output of the hybrid system depicted in Fig. 1, where $P(s)$ denotes the continuous-time plant, $u(t)$ is the continuous-time plant output, u_k is the input to the digital controller $C_d(z)$, y_k the digital controller output, $v(t)$ is the continuous-time control signal and $r(t)$ is the reference control signal.

Assume that H is a zero-order hold with the impulse response $h(t)$. The control signal $v(t)$ is given by

$$v(t) = \sum_{k=-\infty}^{+\infty} h(t - kT_s)y_k \tag{5}$$

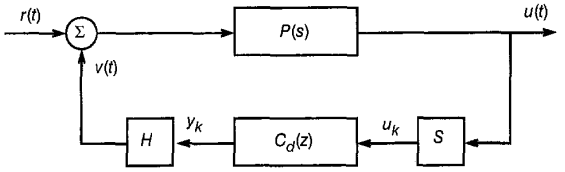

Fig. 1 *Block-diagram of a sampled-data feedback system consisting of a continuous-time plant P(s) and a sampled-data controller with sampler S, digital controller $C_d(z)$ and hold H. Here u(t) denotes the continuous-time plant output, u_k is the digital controller input, y_k the digital controller output, v(t) is the continuous-time control signal and r(t) the reference control signal*

The output of the closed-loop system is then

$$\begin{aligned}u(t) &= P(s)[r(t) + v(t)]\\ &= P(s)r(t) + P(s)\sum_{k=-\infty}^{+\infty} h(t - kT_s)y_k\end{aligned} \tag{6}$$

Since $e_y(k) \triangleq y^*_k - y_k$ is not zero, the actual plant output, denoted as $u^*(t)$, of the hybrid system is different from the ideal one and the difference between the two is

$$\Delta u(t) \triangleq u^*(t) - u(t) = P(s)\sum_{k=-\infty}^{+\infty} h(t - kT_s)e_y(k) \tag{7}$$

One of the main objectives in this Section is to evaluate the output error variance of the sampled-data system due to rounding off in the controller.

2.1 Derivation of averaged roundoff noise gain

Denote

$$\boldsymbol{\epsilon}_x(k) \triangleq Q[\mathbf{x}^*_k] - \mathbf{x}^*_k, \quad \epsilon_u(k) \triangleq Q[u^*_k] - u^*_k \tag{8}$$

as the quantisation errors. Traditionally, these quantisation errors are modelled as statistically independent white sequences (see, e.g. [11, 12]) and

$$E\left[\begin{pmatrix}\boldsymbol{\epsilon}_x(k+m)\\ \epsilon_u(k+m)\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}_x(m)\\ \epsilon_u(m)\end{pmatrix}^T\right] = \sigma_0^2\delta_d(k)\mathbf{I} \tag{9}$$

where $E[\cdot]$ denotes the ensemble average operator, ^{T} the transposed operator, $\sigma_0^2 = 2^{-2B_s}/12$ and $\mathbf{I}$ is the identity matrix of proper dimension. It follows from (1) and (4) that

$$\left.\begin{aligned}\mathbf{e}_x(k+1) &= \mathbf{A}\mathbf{e}_x(k) + \mathbf{B}e_u(k) + \mathbf{A}\boldsymbol{\epsilon}_x(k) + \mathbf{B}\epsilon_u(k)\\ e_y(k) &= \mathbf{C}\mathbf{e}_x(k) + de_u(k) + \mathbf{C}\boldsymbol{\epsilon}_x(k) + d\epsilon_u(k)\end{aligned}\right\} \tag{10}$$

where

$$e_u(k) = u^*_k - u_k, \quad \mathbf{e}_x(k) = \mathbf{x}^*_k - \mathbf{x}_k, \quad e_y(k) = y^*_k - y_k \tag{11}$$

Noting that u_k is the discretised version of $u(t)$, which is the output of the continuous-time plant $P(s)$, it can be computed with the following well-known results.

Theorem 1: Let $x(t)$ and $y(t)$ be the input and output of a continuous-time system $F(s)$. Denote $(\mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s, d_s)$ as a realisation of $F(s)$, i.e. $F(s) = d_s + \mathbf{C}_s(s\mathbf{I} - \mathbf{A}_s)^{-1}\mathbf{B}_s$. Suppose that $y(t)$ is sampled with a sampling frequency $f = 1/T_s$, then the discrete-time sequence $y_k \triangleq y(kT_s)$ can be computed by

$$\left.\begin{aligned}\mathbf{v}_{k+1} &= e^{\mathbf{A}_s T_s}\mathbf{v}_k + \int_0^{T_2} e^{\mathbf{A}_s \tau}\mathbf{B}_s x((k+1)T_s - \tau)d\tau\\ y_k &= \mathbf{C}_s\mathbf{v}_k + d_s x(kT_s)\end{aligned}\right\} \tag{12}$$

Corollary 1: If $x(t)=\sum_{k=-\infty}^{+\infty} h(t-kT_s)x_k$, where $h(t)$ is the response of the time-invariant zero-order hold H to the discrete-time unit impulse function $\delta_d(k)$:

$$h(t)=\begin{cases}1, & 0\le t<T_s\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

one has

$$\left.\begin{aligned}\mathbf{v}_{k+1}&=\mathbf{A}_z\mathbf{v}_k+\mathbf{B}_z x_k\\ y_k&=\mathbf{C}_z\mathbf{v}_k+dx_k\end{aligned}\right\} \tag{14}$$

that is

$$y_k=[d+\mathbf{C}_z(z\mathbf{I}-\mathbf{A}_z)^{-1}\mathbf{B}_z]x_k\triangleq F_d(z)x_k \tag{15}$$

where

$$\mathbf{A}_z=e^{\mathbf{A}_s T_s},\quad \mathbf{B}_z=\int_0^{T_s}e^{\mathbf{A}_s\tau}\mathbf{B}_s d\tau,\quad \mathbf{C}_z=\mathbf{C}_s,\quad d=d_s \tag{16}$$

Suppose that the plant $P(s)$ in Fig. 1 is a strictly proper transfer function, having a realisation $(\mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s)$ such that $P(s)=\mathbf{C}_s(s\mathbf{I}-\mathbf{A}_s)^{-1}\mathbf{B}_s$. Let

$$u_r(t)=P(s)r(t) \tag{17}$$

Noting $u(t)=u_r(t)+P(s)\sum_{m=-\infty}^{+\infty}h(t-mT_s)y_m$, it follows from (15) that

$$u_k=u(kT_s)=u_r(kT_s)+P_d(z)y_k \tag{18}$$

where

$$P_d(z)=\mathbf{C}_z(z\mathbf{I}-\mathbf{A}_z)^{-1}\mathbf{B}_z \tag{19}$$

with $(\mathbf{A}_z, \mathbf{B}_z, \mathbf{C}_z)$ computed according to (16). This leads to

$$e_u(k)=P_d(z)e_y(k) \tag{20}$$

Denote

$$\mathbf{x}_{cl}(k)\triangleq\begin{pmatrix}\mathbf{v}_k\\ \mathbf{e}_x(k)\end{pmatrix}$$

where $\mathbf{v}_k$ is the state vector in (14) with $y_k=e_u(k)$ and $x_k=e_y(k)$. With some manipulations, it can be shown that

$$\left.\begin{aligned}\mathbf{x}_{cl}(k+1)&=\mathbf{A}_{cl}\mathbf{x}_{cl}(k)+\mathbf{B}_{cl}\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\\ e_y(k)&=\mathbf{C}_{cl}\mathbf{x}_{cl}(k)+\mathbf{D}_{cl}\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\end{aligned}\right\} \tag{21}$$

where

$$\left.\begin{aligned}\mathbf{A}_{cl}&=\begin{pmatrix}\mathbf{A}_z+d\mathbf{B}_z\mathbf{C}_z & \mathbf{B}_z\mathbf{C}\\ \mathbf{B}\mathbf{C}_z & \mathbf{A}\end{pmatrix}\\ \mathbf{B}_{cl}&=\begin{pmatrix}\mathbf{B}_z\mathbf{C} & d\mathbf{B}_z\\ \mathbf{A} & \mathbf{B}\end{pmatrix}\\ \mathbf{C}_{cl}&=(d\mathbf{C}_z\quad \mathbf{C})\\ \mathbf{D}_{cl}&=(\mathbf{C}\quad d)\end{aligned}\right\} \tag{22}$$

Therefore

$$\begin{aligned}e_y(k)&=[\mathbf{D}_{cl}+\mathbf{C}_{cl}(z\mathbf{I}-\mathbf{A}_{cl})^{-1}\mathbf{B}_{cl}]\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\\ &\triangleq\mathbf{H}_y(z)\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\end{aligned} \tag{23}$$

Similarly, one has

$$\begin{aligned}e_u(k)&=(\mathbf{C}_z\quad \mathbf{0})(z\mathbf{I}-\mathbf{A}_{cl})^{-1}\mathbf{B}_{cl}\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\\ &\triangleq\mathbf{H}_u(z)\begin{pmatrix}\boldsymbol{\epsilon}_x(k)\\ \epsilon_u(k)\end{pmatrix}\end{aligned} \tag{24}$$

Remark 1: It is easy to see that $\mathbf{A}_{cl}$ is the transition matrix of the digital control system depicted in Fig. 2. This is actually the discrete-time counterpart of the hybrid control system in Fig. 1. It was shown in [18] that the hybrid system is stable if and only if its discrete-time counterpart is stable. Therefore, $\mathbf{H}_y(z)$ and $\mathbf{H}_u(z)$ are stable transfer functions.

Since the quantisation errors $\boldsymbol{\epsilon}_x(k)$ and $\epsilon_u(k)$ are statistically independent white sequences (see (9)), it follows that $e_y(k)$ and $e_u(k)$ are wide-sense stationary sequences. Denote

$$\gamma_u(l)\triangleq E[e_u(k)e_u(k-l)]$$

as the autocorrelation function of $e_u(k)$ and $\Gamma_u(z)$ the corresponding spectral density function. According to the well known results in [19], one has

$$\Gamma_u(z)=\mathbf{H}_u(z)\mathbf{H}_u^T(z^{-1})\sigma_0^2 \tag{25}$$

and

$$\begin{aligned}E[e_u^2(k)]&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Gamma_u(e^{j\omega})d\omega\\ &=\mathrm{tr}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathbf{H}_u^T(e^{-j\omega})\mathbf{H}_u(e^{j\omega})d\omega\right]\sigma_0^2\\ &\triangleq\mathrm{tr}[\mathbf{B}_{cl}^T\bar{\mathbf{W}}_0^d\mathbf{B}_{cl}]\sigma_0^2\end{aligned} \tag{26}$$

where tr[·] denotes the trace operator and $\bar{\mathbf{W}}_0^d$ is the observability gramian of the realisation of $\mathbf{H}_u(z)$ (see (24)):

$$\bar{\mathbf{W}}_0^d=\sum_{k=0}^{+\infty}(\mathbf{A}_{cl}^T)^k\begin{pmatrix}\mathbf{C}_z^T\\ 0\end{pmatrix}(\mathbf{C}_z\quad 0)\mathbf{A}_{cl}^k$$

satisfying

$$\bar{\mathbf{W}}_0^d=\mathbf{A}_{cl}^T\bar{\mathbf{W}}_0^d\mathbf{A}_{cl}+\begin{pmatrix}\mathbf{C}_z^T\\ \mathbf{0}\end{pmatrix}(\mathbf{C}_z\quad \mathbf{0}) \tag{27}$$

Remark 2: The expression given in (26) for the output error variance is based on the digital control system shown in Fig. 2, since the variance is evaluated at the sampling points. This means that the inter-sample behaviour of the output error is not taken into account.

Looking at (7), one can see that the output error of the sampled-data system is the output of the plant excited via a zero-order hold H with an error sequence evaluated with the digital system (23), where the sampling frequency is $f_s=1/T_s$. Denote $H(s)$ as the Laplace transform of $h(t)$ defined in (13). It turns out that

$$\Delta u(t)=\sum_{k=-\infty}^{+\infty}\phi(t-kT_s)e_y(k) \tag{28}$$

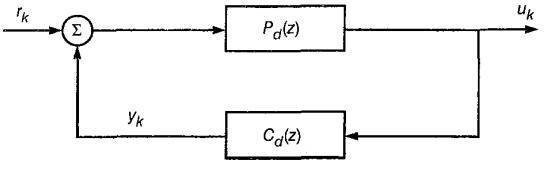

Fig. 2 *Block-diagram of the equivalent discrete-time feedback system*

where $\phi(t)$ is the impulse response of the continuous-time system $P(s)H(s)$. Therefore, the output error variance of the sampled-data system is

$$E[(\Delta u(t))^2] = \sum_{k=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} \phi(t-kT_s)\phi(t-mT_s)\gamma_y(k-m) \tag{29}$$

with $\gamma_y(l) = E[e_y(m+l)e_y(m)]$ the autocorrelation function of $e_y(k)$.

Remark 3:

- It is interesting to note that $E[(\Delta u(t))^2]$ is a periodic function in t with period T_s.
- $\gamma_y(l)$ can readily be evaluated and, since $\mathbf{H}_y(z)$ is stable, $e_y(l)$ is a well-behaved sequence.
- When the plant is unstable, $\phi(t)$ is an unbounded sequence [Note 1]. This is a serious problem for evaluating $E[(\Delta u(t))^2]$ numerically. To overcome this problem we can replace the continuous time plan $P(s)$ with its digital connterpart but with a much higher sampling frequency, denoted as $\bar{f}_s = 1/\bar{T}_s$, than f_s which is used in digital controller.

Let $\bar{f}_s = Nf_s$ with N a large integer. It follows from (15) and (16) that

$$\left.\begin{aligned} \mathbf{v}_{m+1} &= \bar{\mathbf{A}}_z\mathbf{v}_m + \bar{\mathbf{B}}_z\bar{e}_y(m) \\ \Delta u(m\bar{T}_s) &= \bar{\mathbf{C}}_z\mathbf{v}_m \end{aligned}\right\} \tag{30}$$

where $(\bar{\mathbf{A}}_z, \bar{\mathbf{B}}_z, \bar{\mathbf{C}}_z)$ is given by (16) with T_s replaced with $\bar{T}_s = T_s/N$ and

$$\bar{e}_y(m) \triangleq \sum_{k=-\infty}^{+\infty} e_y(k)\eta_N(m-kN) \tag{31}$$

with $\eta_N(m)$ the (discrete-time) window function:

$$\eta_N(m) \triangleq \begin{cases} 1, & 0 \le m < N \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

Denote $\bar{z}$ as the shift operator, corresponding to sampling frequency $\bar{f}_s$, and $\bar{E}_y(\bar{z})$ the $\bar{z}$-transform of $\bar{e}_y(m)$, we then have $\bar{E}_y(\bar{z}) = \sum_{m=-\infty}^{+\infty} \bar{e}_y(m)\bar{z}^{-m} = [(1-\bar{z}^{-N})/(1-\bar{z}^{-1})]E_y(\bar{z}^N)$, where $E_y(z)$ is the z-transform of $e_y(k)$ corresponding to the sampling frequency f_s. According to (23),

$$E_y(z) = \mathbf{H}_y(z)\begin{pmatrix} \mathbf{E}_x(z) \\ E_u(z) \end{pmatrix} \tag{33}$$

with $\boldsymbol{E}_x(z)$ and $E_u(z)$ the z-transforms of $\boldsymbol{\epsilon}_x(k)$ and $\epsilon_u(k)$, respectively. Therefore, the $\bar{z}$-transform of $\Delta u(m\bar{T}_s)$ is equal to

$$\bar{P}_d(\bar{z})\mathbf{H}_y(\bar{z}^N)\begin{pmatrix} \dfrac{1-\bar{z}^{-N}}{1-\bar{z}^{-1}}\mathbf{E}_x(\bar{z}^N) \\ \dfrac{1-\bar{z}^{-N}}{1-\bar{z}^{-1}}E_u(\bar{z}^N) \end{pmatrix} \tag{34}$$

where $\bar{P}_d(\bar{z})$ is the discrete-time counterpart of $P(s)$, obtained from (15) and (16) by substituting T_s with $\bar{T}_s = T_s/N$. This means that

$$\Delta u(m\bar{T}_s) = \bar{P}_d(\bar{z})\mathbf{H}_y(\bar{z}^N)\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m) \\ \bar{\epsilon}_u(m) \end{pmatrix} \triangleq \bar{\mathbf{H}}_u(\bar{z})\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m) \\ \bar{\epsilon}_u(m) \end{pmatrix} \tag{35}$$

Note 1: Though $\phi(t)$ (that is, $P(s)H(s)$) may be unstable, $\Delta u(t)$ is a stable sequence since the sampled-data system is assumed stable. In fact, there is a pole-zero 'cancellation' between $P(s)H(s)$ and $\mathbf{H}_y(z)$.

where $\bar{\boldsymbol{\epsilon}}_x(m)$ and $\bar{\epsilon}_u(m)$ are defined in the same way as $\bar{e}_y(m)$ in (31). It follows from (9) that

$$E\left[\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m+l) \\ \bar{\epsilon}_u(m+l) \end{pmatrix}\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m) \\ \bar{\epsilon}_u(m) \end{pmatrix}^T\right] = \sigma_0^2\mathbf{I}\sum_{k=-\infty}^{+\infty}\eta_N(m+l-kN) \times \eta_N(m-kN) \tag{36}$$

which is periodic in m with period equal to N.

Theorem 2: Let $\mathbf{w}_m$ be the output of a stable transfer function $\mathbf{H}(z) = \sum_{k=0}^{+\infty}\mathbf{H}_k z^{-k}$ excited with a sequence $\boldsymbol{\epsilon}_m$. If $E[\boldsymbol{\epsilon}_{m+l}\boldsymbol{\epsilon}_m^T]$ is a periodic function in m with a period N, so is $E[\mathbf{w}_{m+l}\mathbf{w}_m^T]$. Denote

$$\gamma_\epsilon(l) \triangleq \frac{1}{N}\sum_{m=0}^{N-1}E[\boldsymbol{\epsilon}_{m+l}\boldsymbol{\epsilon}_m^T],\quad \gamma_w(l) \triangleq \frac{1}{N}\sum_{m=0}^{N-1}E[\mathbf{w}_{m+l}\mathbf{w}_m^T] \tag{37}$$

Let $\boldsymbol{\Gamma}_\epsilon(z)$ and $\boldsymbol{\Gamma}_w(z)$ be their corresponding z-transforms. Then

$$\boldsymbol{\Gamma}_w(z) = \mathbf{H}(z)\boldsymbol{\Gamma}_\epsilon(z)\mathbf{H}^T(z^{-1}) \tag{38}$$

Proof: It follows from $\mathbf{w}_m = \sum_{k=0}^{+\infty}\mathbf{H}_k\boldsymbol{\epsilon}_{m-k}$ that

$$E[\mathbf{w}_{m+l}\mathbf{w}_m^T] = \sum_{k_1=0}^{+\infty}\sum_{k_2=0}^{+\infty}\mathbf{H}_{k_1}E\left[\boldsymbol{\epsilon}_{(m-k_2)+(l-k_1+k_2)}\boldsymbol{\epsilon}_{(m-k_2)}^T\right]\mathbf{H}_{k_2}^T \tag{39}$$

Clearly, it is periodic if $E[\boldsymbol{\epsilon}_{m+l}\boldsymbol{\epsilon}_m^T]$ is periodic. Based on the above equation, one has

$$\begin{aligned}\gamma_w(l) &= \sum_{k_1=0}^{+\infty}\sum_{k_2=0}^{+\infty}\mathbf{H}_{k_1}\left\{\frac{1}{N}\sum_{m=0}^{N-1}E\left[\boldsymbol{\epsilon}_{(m-k_2)+(l-k_1+k_2)}\boldsymbol{\epsilon}_{(m-k_2)}^T\right]\right\}\mathbf{H}_{k_2}^T \\ &= \sum_{k_1=0}^{+\infty}\sum_{k_2=0}^{+\infty}\mathbf{H}_{k_1}\gamma_\epsilon(l-k_1+k_2)\mathbf{H}_{k_2}^T \end{aligned} \tag{40}$$

(38) follows by applying z-transformation to both sides of (40). □

Denote

$$\bar{\gamma}_\epsilon(l) \triangleq \frac{1}{N}\sum_{m=0}^{N-1}E\left[\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m+l) \\ \bar{\epsilon}_u(m+l) \end{pmatrix}\begin{pmatrix} \bar{\boldsymbol{\epsilon}}_x(m) \\ \bar{\epsilon}_u(m) \end{pmatrix}^T\right] \tag{41}$$

It can be shown that

$$\bar{\gamma}_\epsilon(l) = \frac{\sigma_0^2}{N}\sum_{m=0}^{N-1}\eta_N(l+m)\mathbf{I}$$

and that its $\bar{z}$-transform, denoted as $\bar{\boldsymbol{\Gamma}}_\epsilon(\bar{z})$, is

$$\bar{\boldsymbol{\Gamma}}_\epsilon(\bar{z}) = \frac{\sigma_0^2}{N}\sum_{m=0}^{N-1}\bar{z}^{-m}\sum_{m=0}^{N-1}\bar{z}^m\mathbf{I} \tag{42}$$

Let $\bar{\Gamma}_u(\bar{z})$ be the $\bar{z}$-transform of

$$\bar{\gamma}_u(l) \triangleq E\left[\frac{1}{N}\sum_{m=0}^{N-1}\Delta u((m+l)\bar{T}_s)\Delta u(m\bar{T}_s)\right]$$

applying theorem 2 leads to

$$\bar{\Gamma}_u(\bar{z}) = \bar{\mathbf{H}}_u(\bar{z})\left(\frac{\sigma_0^2}{N}\sum_{m=0}^{N-1}\bar{z}^{-m}\sum_{m=0}^{N-1}\bar{z}^m\right)\bar{\mathbf{H}}_u^T(\bar{z}^{-1}) \tag{43}$$

Similarly to (25) and (26), the averaged output error variance of the sampled-data system can be written as

$$\bar{\gamma}_u(0) = \frac{1}{N}\sum_{m=0}^{N-1}E[[\Delta u(m\bar{T}_s)]^2] = \mathrm{tr}[\bar{\mathbf{W}}]\sigma_0^2 \tag{44}$$

where

$$\bar{\mathbf{W}} \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{N}\left(\sum_{m=0}^{N-1}e^{-j\omega m}\right)\left(\sum_{m=0}^{N-1}e^{j\omega m}\right)\bar{\mathbf{H}}_u^T(e^{-j\omega})\bar{\mathbf{H}}_u(e^{j\omega})d\omega \tag{45}$$

The averaged roundoff noise gain, denoted as G, is defined as

$$G \triangleq \frac{\bar{\gamma}_u(0)}{\sigma_0^2}$$

and therefore

$$G = \mathrm{tr}[\bar{\mathbf{W}}] \tag{46}$$

2.2 Realisation dependence

Let $(\mathbf{A}_{cl}, \mathbf{B}_{cl}, \mathbf{C}_{cl}, \mathbf{D}_{cl})$ and $(\mathbf{A}_{cl}^0, \mathbf{B}_{cl}^0, \mathbf{C}_{cl}^0, \mathbf{D}_{cl}^0)$ be two realisations of $\mathbf{H}_y(z)$ defined by (21) and (22), corresponding to the two digital controller realisations

$$\mathbf{R} \triangleq (\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$$

and

$$\mathbf{R}_0 \triangleq (\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0, d)$$

that are related with (3), respectively. It can be shown that

$$\left.\begin{aligned}\mathbf{A}_{cl} &= \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix}^{-1}\mathbf{A}_{cl}^0\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix} \quad \mathbf{C}_{cl} = \mathbf{C}_{cl}^0\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix}\\ \mathbf{B}_{cl} &= \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix}^{-1}\mathbf{B}_{cl}^0\begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} \quad \mathbf{D}_{cl} = \mathbf{D}_{cl}^0\begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\end{aligned}\right\} \tag{47}$$

It then turns out that

$$\hat{\mathbf{H}}_u(\bar{z}) = \bar{\mathbf{H}}_u^0(\bar{z})\begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}$$

where $\bar{\mathbf{H}}_u^0(\bar{z})$ is independent of $\mathbf{T}$, and hence

$$\bar{\mathbf{W}} = \begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}^T\bar{\mathbf{W}}^0\begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} \tag{48}$$

where $\bar{\mathbf{W}}^0$ is similar to $\bar{\mathbf{W}}$ defined in (45) but corresponds to the controller realisation $\mathbf{R}_0$.

Let

$$\bar{\mathbf{W}} \triangleq \begin{pmatrix}\mathbf{W} & \mathbf{W}_{12}\\ \mathbf{W}_{21} & \mathbf{Q}\end{pmatrix}$$

and

$$\bar{\mathbf{W}}^0 \triangleq \begin{pmatrix}\mathbf{W}_0 & \mathbf{W}_{12}^0\\ \mathbf{W}_{21}^0 & \mathbf{Q}_0\end{pmatrix}$$

have the same partition as

$$\begin{pmatrix}\mathbf{T} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}$$

It is easy to see that $\mathbf{W} = \mathbf{T}^T\mathbf{W}_0\mathbf{T}$, $\mathbf{Q} = \mathbf{Q}_0$ and hence

$$G = \mathrm{tr}[\mathbf{T}^T\mathbf{W}_0\mathbf{T}] + \mathrm{tr}[\mathbf{Q}_0] \tag{49}$$

Remark 4: For the equivalent discrete-time feedback system shown in Fig. 2, the averaged roundoff noise gain, denoted was G_d, is defined by $G_d \triangleq E[e_u^2(k)]/\sigma_0^2$. It follows from (26) that $G_d = \mathrm{tr}[\mathbf{B}_{cl}\bar{\mathbf{W}}_0^d\mathbf{B}_{cl}]$. Similarly, one can show that

$$G_d = \mathrm{tr}[\mathbf{T}^T\mathbf{W}_0^d\mathbf{T}] + \mathrm{tr}[\mathbf{Q}_0^d] \tag{50}$$

where $\mathbf{W}_0^d$ and $\mathbf{Q}_0^d$ are independent of $\mathbf{T}$.

Clearly, the averaged roundoff noise gain G (or G_d) can be divided into two parts: one is a function of the controller realisation, and the other is a constant, having nothing to do with the controller structure.

3 Dynamic range of states and optimal controller realisations

On the one hand, G is a function of realisation and hence can be made as small as possible with 'small' $\mathbf{T}$. The dynamic range of the states in (1), on the other hand, varies dramatically with realisations. From a practical point of view [Note 2], one wishes that all the states have the same dynamic range. To do so, the actually implemented realisation must be scaled.

The classical l_2-scaling on the states implies that the variance of each state is all equal to one when the input signal $r(t)$ is a white noise with unit variance. Denote $\mathbf{K} \triangleq E[\mathbf{x}_k\mathbf{x}_k^T]$ as the covariance matrix of the state vector of the controller, corresponding to realisation $\mathbf{R}$. The l_2-scaling (see, e.g. [11, 12]) implies that the diagonal elements of $\mathbf{K}$ satisfy

$$\mathbf{K}(i, i) = 1, \forall i \tag{51}$$

Let us re-visit (18). According to Theorem 1, $u_r(kT_s) = \mathbf{C}_z(z\mathbf{I} - \mathbf{A}_z)^{-1}\mathbf{s}_k$, where

$$\mathbf{s}_k \triangleq \int_0^{T_s} e^{\mathbf{A}_s\tau}\mathbf{B}_s r((k+1)T_s - \tau)d\tau \tag{52}$$

Therefore, $u_k = \mathbf{C}_z(z\mathbf{I} - \mathbf{A}_z)^{-1}(\mathbf{s}_k + \mathbf{B}_z y_k)$, which is equivalent to

$$\left.\begin{aligned}\mathbf{v}_{k+1} &= \mathbf{A}_z\mathbf{v}_k + \mathbf{B}_z y_k + \mathbf{s}_k\\ u_k &= \mathbf{C}_z\mathbf{v}_k\end{aligned}\right\} \tag{53}$$

Combining (53) with (1), one has

$$\begin{pmatrix}\mathbf{v}_{k+1}\\ \mathbf{x}_{k+1}\end{pmatrix} = \mathbf{A}_{cl}\begin{pmatrix}\mathbf{v}_k\\ \mathbf{x}_k\end{pmatrix} + \begin{pmatrix}\mathbf{s}_k\\ \mathbf{0}\end{pmatrix} \tag{54}$$

This means that

$$\begin{pmatrix}\mathbf{v}_k\\ \mathbf{x}_k\end{pmatrix} = (z\mathbf{I} - \mathbf{A}_{cl})^{-1}\begin{pmatrix}\mathbf{s}_k\\ \mathbf{0}\end{pmatrix} \tag{55}$$

Since $r(t)$ is a white noise of unit variance, i.e. $E[r(t+\tau)r(t)] = \delta(\tau)$. It follows from the expression (52) for $\mathbf{s}_k$ that

$$E[\mathbf{s}_{k+m}\mathbf{s}_m^T] = \int_0^{T_s} e^{\mathbf{A}_s\tau}\mathbf{B}_s\left\{\int_0^{T_s} E[r((k+m+1)T_s - \tau) \times r((m+1)T_s - \xi)]\mathbf{B}_s^T e^{\mathbf{A}_s^T\xi}d\xi\right\}d\tau \tag{56}$$

that is

$$E[\mathbf{s}_{k+m}\mathbf{s}_m^T] = \int_0^{T_s} e^{\mathbf{A}_s\tau}\mathbf{B}_s\left\{\int_{kT_s-\tau}^{(k+1)T_s-\tau}\delta(\zeta)\mathbf{B}_s^T e^{\mathbf{A}_s^T(\zeta+\tau-kT_s)}d\zeta\right\}d\tau \tag{57}$$

Note 2: For example, to avoid any overflow effect.

It is easy to see that, in the above equation, the term inside $\{\cdot\}$ is equal to zero for all k, except $k=0$ for which it is equal to $\mathbf{B}_s^T e^{\mathbf{A}_s^T\tau}$. We then have

$$\gamma_s(k) \triangleq E\left[\mathbf{s}_{k+m}\mathbf{s}_m^T\right] = \int_0^{T_s} e^{\mathbf{A}_s\tau}\mathbf{B}_s\mathbf{B}_s^T e^{\mathbf{A}_s^T\tau}d\tau\delta_d(k) \triangleq \mathbf{\Gamma}_0\delta_d(k) \quad (58)$$

which implies that $\mathbf{s}_k$ is a wide-sense stationary sequence and so is

$$\begin{pmatrix}\mathbf{v}_k\\ \mathbf{x}_k\end{pmatrix}$$

It follows that the latter has a power density function given by

$$\mathbf{\Gamma}(z) \triangleq (z\mathbf{I}-\mathbf{A}_{cl})^{-1}\begin{pmatrix}\mathbf{\Gamma}_0 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}(z^{-1}\mathbf{I}-\mathbf{A}_{cl}^T)^{-1} \quad (59)$$

Denoting

$$\mathbf{B}_\kappa \triangleq \begin{pmatrix}\mathbf{\Gamma}_0^{1/2}\\ \mathbf{0}\end{pmatrix}$$

with $\mathbf{\Gamma}_0^{1/2}$ square root matrix of $\mathbf{\Gamma}_0$, one can see that

$$\bar{\mathbf{K}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathbf{\Gamma}(e^{j\omega})d\omega = \bar{\mathbf{W}}_c \quad (60)$$

where $\bar{\mathbf{W}}_c$, called the controllability gramian corresponding to $(\mathbf{A}_{cl}, \mathbf{B}_\kappa)$, satisfies

$$\bar{\mathbf{W}}_c = \mathbf{A}_{cl}\bar{\mathbf{W}}_c\mathbf{A}_{cl}^T + \mathbf{B}_\kappa\mathbf{B}_\kappa \quad (61)$$

Let

$$\bar{\mathbf{K}} = \begin{pmatrix}\mathbf{K}_{11} & \mathbf{K}_{12}\\ \mathbf{K}_{21} & \mathbf{K}\end{pmatrix} \quad (62)$$

where $\mathbf{K}$ has the same dimension as that of $\mathbf{A}$. It follows from (47) that

$$\bar{\mathbf{K}} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix}^{-1}\bar{\mathbf{K}}_0\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{pmatrix}^{-T} \quad (63)$$

that is

$$\mathbf{K} = \mathbf{T}^{-1}\mathbf{K}_0\mathbf{T}^{-T} \quad (64)$$

where $\mathbf{K}_0$ is a positive-definite matrix independent of $\mathbf{T}$, corresponding to the controller realisation $\mathbf{R}_0$.

The above equation means that for a given controller realisation, say $\mathbf{R}_0$, the l_2-scaling can be achieved by applying a diagonal transformation $\mathbf{T}_d = \text{diag}\{\eta_1, \ldots, \eta_k, \ldots, \eta_{n_c}\}$ with

$$\eta_k = \sqrt{1/\mathbf{K}_0(k,k)}, \forall k$$

This transformation leads to an l_2-scaled realisation, denoted as $\mathbf{R}_0^{sc}$, that has almost the same structure as $\mathbf{R}_0$ and can prevent the controller from overflow. But it is usually not the best one since it may have a very large averaged roundoff noise gain G.

The optimum realisations of the digital controller to the hybrid system depicted in Fig. 1 are the solutions to the following minimisation problem:

$$\min_{\substack{\mathbf{R}\in\mathcal{S}_{C_d}\\ \text{subject to (51)}}} \quad G \quad (65)$$

In the next Section, we will discuss how to compute the newly derived measure G and hence to solve for the optimum realisation problem (65).

4 Computing optimal realisations

The l_2 dynamic range constraint (51) defines a set of controller realisations, denoted as $\mathcal{S}_{C_d}^{sc}$, in which each realisation satisfies

$$(\mathbf{T}^{-1}\mathbf{K}_0\mathbf{T}^{-T})(i,i) = 1, \quad \text{for } i = 1, 2, \ldots, n_c \quad (66)$$

Noting the fact that $\text{tr}[\mathbf{Q}_0]$ is independent of $\mathbf{T}$, the optimum controller realisation problem under l_2-scaling constraint can be specified as

$$\min_{\substack{\mathbf{T}:\det(\mathbf{T})\neq 0\\ \text{subject to (66)}}} \text{tr}\left[\mathbf{T}^T\mathbf{W}_0\mathbf{T}\right] \quad (67)$$

This problem was solved for independently in [11, 12]. In what follows, we present an alternative approach to solve the optimisation problem (67) and provide an analytical solution.

Lemma 1: Let $\mathbf{K}_0 \geq 0$ be a given $n_c \times n_c$ matrix and $\mathbf{T} = \mathbf{T}_1\mathbf{V}$ a non-singular matrix of the same dimension, where $\mathbf{V}$ is an orthogonal matrix. There exists a $\mathbf{T}$ such that (66) holds if an only if

$$\text{tr}\left[\mathbf{T}_1^{-1}\mathbf{K}_0\mathbf{T}_1^{-T}\right] = n_c \quad (68)$$

Proof: The necessary condition is obvious, and we prove the sufficient condition. By a singular value decomposition (SVD), $\mathbf{T}_1^{-1}\mathbf{K}_0\mathbf{T}_1^{-T} = \mathbf{V}_0^T\mathbf{\Sigma}\mathbf{V}_0$, where $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \ldots, \sigma_{n_c}\} \geq 0$ and $\mathbf{V}_0$ is some orthogonal matrix. So, $\mathbf{T}^{-1}\mathbf{K}_0\mathbf{T}^{-T} = \bar{\mathbf{V}}^T\mathbf{\Sigma}\bar{\mathbf{V}}$ with $\bar{\mathbf{V}} = \mathbf{V}_0\mathbf{V}$. Using the numerical algorithm given in [12], one can find a $\bar{\mathbf{V}}$ such that $\bar{\mathbf{V}}^T\mathbf{\Sigma}\bar{\mathbf{V}}$ has its diagonal elements all equal to one, which means $\mathbf{V} = \mathbf{V}_0^T\bar{\mathbf{V}}$. □

With Lemma 1, (67) can be rewritten as

$$\min_{\substack{\mathbf{T}_1:\det(\mathbf{T}_1)\neq 0\\ n_c=\text{tr}[\mathbf{T}_1^{-1}\mathbf{K}_0\mathbf{T}_1^{-T}]}} \text{tr}\left[\mathbf{T}_1^T\mathbf{W}_0\mathbf{T}_1\right] \quad (69)$$

This problem can be solved for using the Lagrange multiplier method. Noting $\text{tr}[\mathbf{T}_1^T\mathbf{W}_0\mathbf{T}_1] = \text{tr}[\mathbf{W}_0\mathbf{P}_1]$ and $\text{tr}[\mathbf{T}_1^{-1}\mathbf{K}_0\mathbf{T}_1^{-T}] = \text{tr}[\mathbf{K}_0\mathbf{P}_1^{-1}]$, where

$$\mathbf{P}_1 \triangleq \mathbf{T}_1\mathbf{T}_1^T$$

we define the Lagrangian

$$L(\mathbf{P}_1, \lambda) \triangleq \text{tr}[\mathbf{W}_0\mathbf{P}_1] + \lambda\left(\text{tr}[\mathbf{K}_0\mathbf{P}_1^{-1}] - n_c\right) \quad (70)$$

The optimal $\mathbf{P}_1$ should satisfy $\partial L/\partial\mathbf{P}_1 = \mathbf{0}$ and $\partial L/\partial\lambda_1 = 0$, which leads to

$$\left.\begin{aligned}\mathbf{W}_0 &= \lambda\mathbf{P}_1^{-1}\mathbf{K}_0\mathbf{P}_1^{-1}\\ n_c &= \text{tr}[\mathbf{K}_0\mathbf{P}_1^{-1}]\end{aligned}\right\} \quad (71)$$

The first equation of (71) implies $\lambda > 0$ and can be rewritten as

$$\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2} = (\lambda^{1/2}\mathbf{K}_0^{1/2}\mathbf{P}_1^{-1}\mathbf{K}_0^{1/2})(\lambda^{1/2}\mathbf{K}_0^{1/2}\mathbf{P}_1^{-1}\mathbf{K}_0^{1/2}) \quad (72)$$

which suggests that $\lambda^{1/2}\mathbf{K}_0^{1/2}\mathbf{P}_0^{-1}\mathbf{K}_0^{1/2}$ is a square root matrix of $\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2}$. Since the latter is a positive-definite matrix, its square root matrix is unique. Therefore

$$(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{1/2} = \lambda^{1/2}\mathbf{K}_0^{1/2}\mathbf{P}_1^{-1}\mathbf{K}_0^{1/2} \quad (73)$$

which leads to

$$\mathbf{P}_1 = \lambda^{1/2}\mathbf{K}_0^{1/2}(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{-1/2}\mathbf{K}_0^{1/2} \quad (74)$$

With the second equation of (71), one has

$$\lambda^{1/2} = \frac{\mathrm{tr}[(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{1/2}]}{n_c} \tag{75}$$

Therefore

$$\mathbf{P}_1 = \frac{\mathrm{tr}[(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{1/2}]}{n_c}\mathbf{K}_0^{1/2}(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{-1/2}\mathbf{K}_0^{1/2} \tag{76}$$

As a summary, we then have the following theorem:

Theorem 3: The solutions to (67) are characterised by

$$\mathbf{T}_{opt} = \mathbf{P}_1^{1/2}\mathbf{V} \tag{77}$$

where $\mathbf{P}_1$ is given by (76) and $\mathbf{V}$ is any orthogonal matrix such that the diagonal elements of $\mathbf{V}^T\mathbf{P}_1^{-1/2}\mathbf{K}_0\mathbf{P}_1^{-1/2}\mathbf{V}$ are all equal to one; the minimum of the averaged roundoff power gain with l_2 scaling is equal to

$$G_{min} = \frac{\left(\sum_{k=1}^{n_c}\sigma_k\right)^2}{n_c} + \mathrm{tr}[\mathbf{Q}_0] \tag{78}$$

where $\{\sigma_k^2\}$ is the eigenvalues set of $\mathbf{K}_0\mathbf{W}_0$.

Proof: First of all, all the solutions to (67) belong to the similarity transformation set defined by (77). It is easy to verify that, with $\mathbf{T}_{opt}$ given by (77), $\mathrm{tr}[\mathbf{T}_{opt}^T\mathbf{W}_0\mathbf{T}_{opt}] = \mathrm{tr}[\mathbf{P}_1^{1/2}\mathbf{W}_0\mathbf{P}_1^{1/2}] = (\mathrm{tr}[(\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2})^{1/2}])^2/n_c$. This means that $\mathbf{T}_{opt}$ defined in (77) is actually the complete solution set of (67). Noting the fact that $\mathbf{K}_0^{1/2}\mathbf{W}_0\mathbf{K}_0^{1/2}$ and $\mathbf{W}_0\mathbf{K}_0$ have the same eigenvalues, (78) follows. □

With such a $\mathbf{T}_{opt}$ and the given initial realisation $\mathbf{R}_0$, one can obtain an optimal controller realisation of the sampled-data system, denoted as $\mathbf{R}_{opt}$, using (3).

It is easy to understand that the roundoff noise gain G_d of the digital closed-loop system can be minimised and the corresponding optimal transformations can be obtained in the same way. The optimal controller realisations so obtained are denoted as $\mathbf{R}_{opt}^d$ [Note 3]. Compared with $\mathbf{R}_{opt}$, $\mathbf{R}_{opt}^d$ is 'locally' optimal since G_d is a measure that does not take into account the inter-sample behaviour of the sampled-data system.

5 Design example

We now present a design example to illustrate the design procedure. The transfer function of the plant is

$$P(s) = \frac{1.6188s^2 - 0.1575s - 43.9425}{s^5 + 1.1736s^4 + 28.0737s^3 + 27.9187s^2 + 0.0186s}$$

A stabilising (continuous-time) controller $C_s(s)$ is designed and the transfer function is

$$C_s(s) = \frac{0.046s^6 + 1.5862s^5 + 3.09s^4 + 44.3s^3 + 42.7785s^2 + 0.02867s + 1.58\times10^{-4}}{s^6 + 3.766s^5 + 34.9509s^4 + 106.2s^3 + 179.2s^2 + 166.43s + 0.0033}$$

With $f_s = 1$ Hz, we obtained the discretised plant $P_d(z)$ and controller $C_d(z)$, both are presented with their controllable realisations, denoted as $\mathbf{R}_z$ and $\mathbf{R}_c$, respectively. $\mathbf{R}_z$ is given by

$$\mathbf{A}_z = \begin{pmatrix} 3.3555 & -4.9154 & 4.0734 & -1.8227 & 0.3093 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{B}_z = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{C}_z = (-0.1183 \quad -0.7249 \quad 0.6878 \quad -0.6510 \quad -0.0425)$$

$\mathbf{R}_c$ is given by

$$\mathbf{A}_c = \begin{pmatrix} 2.1016 & -2.2306 & 1.4467 & -0.4901 & 0.1954 & -0.0231 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{B}_c = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{C}_c = (0.1971 \quad -0.7401 \quad 1.1527 \quad -1.0041 \quad 0.4857 \quad -0.0912),$$

$$d_c = 0.0460$$

We point out that the coefficients of $P_d(z)$ and $C_d(z)$ are presented in *FORMAT SHORT* (MATLAB) and hence only the first four significant digits in fractional part of each parameter are displayed. In the sequel, the *FORMAT SHORT* display is assumed. The poles of the ideal closed-loop system and the corresponding absolute values can be computed and they are all inside the unit circle. Clearly, this digital closed-loop system is stable and hence so is the sampled-data system.

With $\mathbf{R}_c$ as the initial digital controller realisation and $N = 111$ in (45), we compute the corresponding $\mathbf{K}_0$, $\mathbf{W}_0$ and $\mathbf{W}_0^d$, based on which three l_2-scaled controller realisations $\mathbf{R}_c^{sc}$, $\mathbf{R}_{opt}^d$ and $\mathbf{R}_{opt}$ are obtained, where $\mathbf{R}_c^{sc}$ is obtained from $\mathbf{R}_c$ with a diagonal transformation, $\mathbf{R}_{opt}^d$ and $\mathbf{R}_{opt}$, as defined before, are the optimal realisations that minimise G_d and G, respectively. Table 1 shows the averaged roundoff noise gain of each realisation. The results in Table 1 are self-explanatory. Both the $\mathbf{R}_{opt}^d$ and $\mathbf{R}_{opt}$ yield a much smaller averaged roundoff noise gain than $\mathbf{R}_c^{sc}$. Comparing

Table 1: Comparison of the averaged roundoff noise gains for the three l_2-scaled realisations

Realisation	$\mathbf{R}_c^{sc}$	$\mathbf{R}_{opt}^d$	$\mathbf{R}_{opt}$
G	8.7344×10^{13}	5.0269×10^{6}	4.1116×10^{6}

Note 3: $\mathbf{R}_{opt}^d$ are different from the classical optimal realisations which are obtained by minimising G_d with the constraint $\mathbf{K}_d(i,i) = 1$, $\forall i$, where $\mathbf{K}_d$ is similar to the $\mathbf{K}$ matrix but corresponding to the equivalent digital control system.

the optimum realisation $\mathbf{R}_{opt}$ with $\mathbf{R}^d_{opt}$, one can see that the averaged roundoff noise gain of the former is about 80% of the latter.

Though it is very hard to compute the averaged roundoff noise gain (see (44)–(46)) with simulation data, it is expected that $\mathbf{R}_{opt}$ or $\mathbf{R}^d_{opt}$ should have a much smaller output error variance than $\mathbf{R}^{sc}_c$. Also $\mathbf{R}_{opt}$ should have a smaller output error variance than $\mathbf{R}^d_{opt}$. To confirm these, some simulations have been conducted. In Fig. 1, the input signal $r(t)$ is replaced with a white sequence of 100 000 points, generated with unit variance using the command *randn* in MATLAB. The continuous-time plant is replaced with its discrete-time counterpart obtained using fast sampling ($\bar{f}_s = 10 f_s = 10$ Hz). The digital controller is implemented with (4), where the quantiser $Q[p]$ rounds the fractional part of signal p into 16 bits. The variance of the error sequence between the ideal output and the actual one of the sampled-data system for each of the three controller realisations is computed with the same input sequence. For $\mathbf{R}^{sc}_c$, the variance is 3.2708×10^6, while for $\mathbf{R}^d_{opt}$ and $\mathbf{R}_{opt}$, we have 8.1846 and 6.2318, respectively.

Fig. 3 shows the unit-step responses of the sampled-data system, where the solid line is for the ideal response, while the dashed and the dotted lines are for 16-bit implemented $\mathbf{R}_{opt}$ and $\mathbf{R}^{sc}_c$, respectively. Clearly, the response corresponding to $\mathbf{R}^{sc}_c$ is far away from the desired one, while the one corresponding to $\mathbf{R}_{opt}$ is very close to the ideal response. Fig. 4 compares the ideal unit-step response of the sampled-data system with those of 16-bit implemented $\mathbf{R}_{opt}$ and $\mathbf{R}^d_{opt}$, respectively. It can just be seen visually that the output error for $\mathbf{R}^d_{opt}$ is larger than that for $\mathbf{R}_{opt}$.

6 Conclusions

We have addressed the optimum digital controller structure problem in a hybrid system with roundoff noise consideration. Our contribution has been threefold. The first one is to have given a thorough analysis of the effect of roundoff noise in the digital controller on the output of the system. Based on this analysis, a new measure has been proposed. This measure, unlike the existing ones, is derived for the hybrid system rather than its discrete-time counterpart and hence can take the inter-sample behaviour into account. The second contribution is to have given a method to evaluate this measure by fast sampling plant, which can avoid the numerical problem involved in computing directly the newly defined measure. The exact expression for the covariance matrix of controller state vector has also been derived in order to scale the realisations with l_2 norm. It is shown that the proposed new measure is controller realisation dependent, and the third contribution of this paper is to have presented an analytical solution to the optimum controller structure problem. A design example has been given to illustrate the design procedure and to confirm the theoretical results.

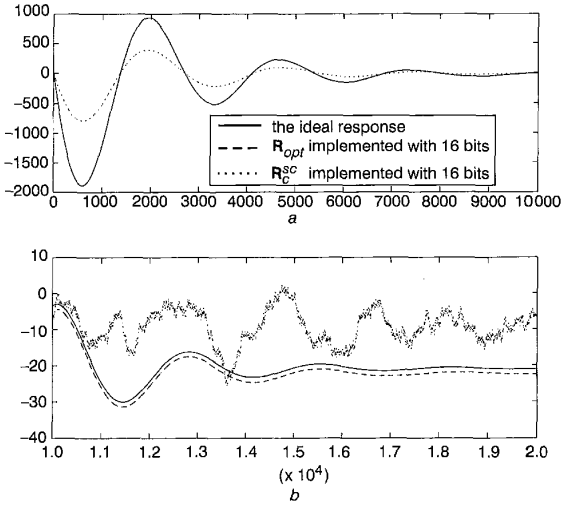

Fig. 3 *Unit-step responses of the sampled-data system (x-axis in second)*

Note that in the first 10000 seconds (*a*), with the given *y*-axis range, the differences for the ideal response and that of $\mathbf{R}_{opt}$ implemented with 16 bits cannot visually be seen

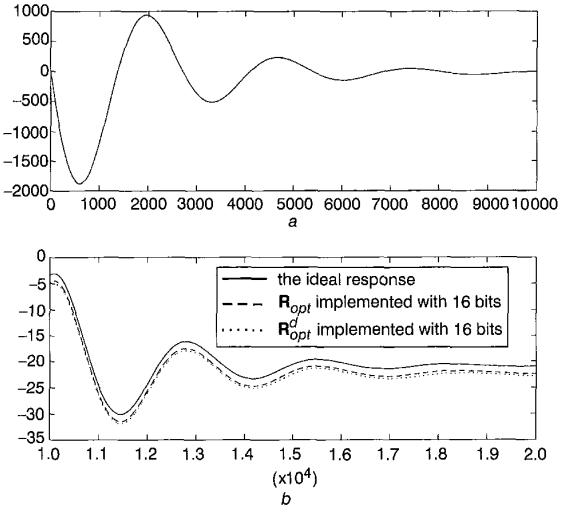

Fig. 4 *Unit-step responses of the sampled-data system (x-axis in second)*

Note that in the first 10000 seconds (*a*), with the given *y*-axis range, the differences for the three responses cannot visually be seen

7 Acknowledgments

J. Wu and S. Chen are grateful for the support of the UK Royal Society under a KC Wong fellowship (RL/ART/CN/XFI/KCW/11949).

8 References

1 LI, G., and GEVERS, M.: 'Optimal finite precision implementation of a state-estimate feedback controller', *IEEE Trans. Circuits Syst.*, 1990, **38**, (12), pp. 1487–1499
2 GEVERS, M., and LI, G.: 'Parametrisations in control, estimation and filtering problems: accuracy aspects' (Springer Verlag, London, Communication and Control Engineering Series, 1993)
3 MADIEVSKI, A.G., ANDERSON, B.D.O., and GEVERS, M.: 'Optimum realizations of sampled-data controllers for FWL sensitivity minimization', *Automatica*, 1995, **31**, (3), pp. 367–379
4 FIALHO, I.J., and GEORGIOU, T.T.: 'On stability and performance of sampled-data systems subject to wordlength constraint', *IEEE Trans. Autom. Control*, 1994, **39**, pp. 2476–2481
5 FIALHO, I.J., and GEORGIOU, T.T.: 'Optimal finite wordlength digital controller realizations'. Proc. American Control Conf., San Diego, USA, 2–4 June 1999, pp. 4326–4327
6 LI, G.: 'On the structure of digital controllers with finite word length consideration', *IEEE Trans. Autom. Control*, 1998, **43**, (5), pp. 689–693
7 ISTEPANIAN, R.H., LI, G., WU, J., and CHU, J.: 'Analysis of sensitivity measures of finite-precision digital controller structures with closed-loop stability bounds', *IEE Proc., Control Theory Appl.*, 1998, **145**, (5), pp. 472–478
8 CHEN, S., WU, J., ISTEPANIAN, R.H., and CHU, J.: 'Optimizing stability bounds of finite-precision PID controller structures', *IEEE Trans. Autom. Control*, 1999, **44**, (11), pp. 2149–2153

9 WU, J., CHEN, S., LI, G., and CHU, J.: 'Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems', *IEEE Trans. Autom. Control*, 2000, **45**, (8), pp. 1550–1554
10 CHEN, S., ISTEPANIAN, R.H., WU, J., and CHU, J.: 'Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators', *Syst. Control Lett.*, 2000, **40**, (3), pp. 153–163
11 MULLIS, C.T., and ROBERTS, R.A.: 'Synthesis of minimum roundoff noise fixed-point digital filters', *IEEE Trans. Circuits Syst.*, 1976, **23**, pp. 551–562
12 HWANG, S.Y.: 'Minimum uncorrelated unit noise in state-space digital filtering', *IEEE Trans. Acoust., Speech Signal Process.*, 1977, **25**, (4), pp. 273–281
13 ROBERTS, R.A., and MULLIS, C.T.: 'Digital signal processing' (Addison Wesley, 1987)
14 AMI, G., and SHAKED, U.: 'Small roundoff realization of fixed-point digital filters and controllers', *IEEE Trans. Acoust., Speech Signal Process.*, 1988, **36**, (6), pp. 880–891
15 MORONEY, P., WILLSKY, A.S., and HOUPT, P.K.: 'The digital implementation of control compensators: the coefficient wordlength issue', *IEEE Trans. Autom. Control*, 1980, **25**, (4), pp. 621–630
16 WILLIAMSON, D., and KADIMAN, K.: 'Optimal finite wordlength linear quadratic regulation', *IEEE Trans. Autom. Control*, 1989, **34**, (12), pp. 1218–1228
17 LIU, K., SKELTON, R., and GRIGORIADIS, K.: 'Optimal controllers for finite wordlength implementation', *IEEE Trans. Autom. Control*, 1992, **37**, pp. 1294–1304
18 CHEN, T., and FRANCIS, B.A.: 'Input-output stability of sampled data systems', *IEEE Trans. Autom. Control*, 1991, **36**, (1), pp. 50–58
19 ÅSTRÖM, K.J.: 'Introduction to stochastic control theory' (Academic Press, New York, 1970)