Sparse Controller Realization with Small Roundoff Noise

G. Li [1], J. Wu [2] and S. Chen [3] *

[1] School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798

[2] National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou, 310027, P. R. China

[3] School of Electronics and Computer Science
University of Southampton
Highfield, Southampton SO17 1BJ, U.K.

Abstract

In this paper, the effect of roundoff noise in a digital controller is analyzed for a digital feedback control system. An analytical expression for the roundoff noise gain, defined as the ratio between the variances of the output error and the rounding error, is obtained. The problem of identifying the minimum roundoff noise realizations can be solved using an existing procedure. Noting that the optimal realizations are fully parametrized, based on a polynomial operator approach a new sparse controller realization is derived. This realization is a generalization of the direct forms in the classical shift operator and the prevailing delta operator. It provides us more degrees of freedom to reduce the roundoff noise. The problem of finding optimal polynomial operators can be solved with exhaustive search, and a design example is given. It is shown that with the proposed sparse realization the optimal polynomial operators can outperform the shift- and delta-operators.

Keywords: finite word length, digital feedback control systems, roundoff noise, polynomial operators, optimal realizations, sparse realizations.

1 Introduction

In most of the discrete-time control systems, the designed digital controller has to be implemented with 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 one is perturbation of controller parameters implemented with FWL and the second one 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

*Contact author: Tel./Fax: +44 (0)23 8059 6660/4508; E-mail: sqc@ecs.soton.ac.uk

classically studied with a transfer function sensitivity measure [1],[2]. More 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 [3],[4] and those based on pole sensitivity [5]-[11]. The second type of FWL errors 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. However, it was not until the late 1980s that the problem of optimal digital controller realizations minimizing the roundoff noise gain was addressed. For example, a roundoff noise gain was derived for a control system with state-estimate feedback controller and the corresponding optimal realization problem was solved in [1], while the roundoff error effect on the LQG performance was investigated in [7] and the optimal solution was obtained by Liu *et al* [8]. The problem of finding the optimum roundoff noise structures of digital controllers in sampled-data system was investigated in [9].

Recently, the delta operator based realizations have been studied in [12]-[14]. It was shown that under certain mild conditions the realizations in the delta operator yield a better performance against the FWL effects than those realizations in the shift operator. It should be pointed out that the optimal realizations obtained so far are fully parametrized and that from a practical point of view it is desired to implement the controller in such a realization that not only yields a very good performance against the FWL effects but also possesses as many trivial parameters[1] to be implemented as possible. The problem of finding sparse controller realizations was considered by several researchers. In [15], the "optimal" sparse controller realization was computed with loop opened. This realization is obviously not optimal in the sense that it does not minimize the roundoff noise in the closed-loop system. The sparse controller structures that maximize the stability robustness of closed-loop were investigated in [5],[16]. It should be pointed out that in these approaches complicated numerical algorithms were utilized and the positions of trivial parameters are not predicted.

The use of delta operator, defined as $\delta = \frac{z-1}{T_s}$ with T_s the sampling period, was first promoted by Peterka [17] and Middleton and Goodwin [18] in estimation and control applications. Two major advantages are claimed for the use of this operator: a theoretically interesting unified formation of continuous-time and discrete-time filtering and control theory, and a range of practically interesting numerical advantages connected with FWL effects. Later on, the numerical properties of the delta operator, where T_s is replaced by a positive factor Δ, were investigated in [2] from a pure algebraic point of view, where it was found that one can make the transfer function in delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$. This means that the delta operator based realizations may not yield a better performance if some of poles of the

[1]By *trivial parameters* we mean those that are 0 and ± 1, which can be digitally implemented exactly and cause no rounding errors. Other parameters are, therefore, referred to as *non-trivial parameters*.

transfer function are far away from $z = +1$. In [2], the concept of polynomial operators was proposed. One of the main objectives in this paper is based on this concept to derive a sparse controller realization and analyze its performance in terms of roundoff noise for a discrete-time feedback control system.

2 Optimal roundoff noise realizations

Throughout the paper, a bold type symbol denotes a vector or matrix with appropriate dimension. The discrete-time feedback control system considered in this paper is depicted in Figure 1. 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 and y_k are the output of the digital plant $P_d(z)$ and controller $C_d(z)$, respectively, while r_k is the input signal of the closed-loop system. $\mathbf{A} \in \mathcal{R}^{p\times p}, \mathbf{B} \in \mathcal{R}^{p\times 1}, \mathbf{C} \in \mathcal{R}^{1\times p}$ and $d \in \mathcal{R}$. $\mathbf{R} \triangleq (\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$ is called a realization of $C_d(z)$, which satisfies

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

Denote $\mathcal{S}_{C_d}$ as the set of all realizations $(\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$. 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 characterized with

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

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

2.1 Dynamic scaling scheme

It is well know that signal scaling is crucial in digital implementation. When two's complement and less-than-double precision fixed-point arithmetic are used, summation nodes are allowed to overflow except before multiplier (see for example [19]). Under the assumption that the input r_k and the output u_k of the closed-loop system are properly pre-scaled, the only signals which may have overflow are the elements of the controller state vector $\mathbf{x}_k$, which have to be scaled.

Note that $u_k = P_d(z)[r_k + y_k]$. Assuming that $P_d(z)$ is strictly proper and has a realization $(\mathbf{A}_z, \mathbf{B}_z, \mathbf{C}_z)$, one has

$$\left.\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{A}_z\mathbf{v}_k + \mathbf{B}_z[r_k + y_k] \\ u_k &= \mathbf{C}_z\mathbf{v}_k \end{aligned}\right\} \tag{4}$$

where $\mathbf{A}_z \in \mathcal{R}^{q\times q}, \mathbf{B}_z \in \mathcal{R}^{q\times 1}$, and $\mathbf{C}_z \in \mathcal{R}^{1\times q}$. Denote

$$\mathbf{x}_{cl}^T(k) \triangleq \begin{pmatrix} \mathbf{v}_k^T & \mathbf{x}_k^T \end{pmatrix}. \tag{5}$$

It follows from (1) and (4) that

$$\mathbf{x}_{cl}(k+1) = \mathbf{A}_{cl}\mathbf{x}_{cl}(k) + \mathbf{B}_r r_k \tag{6}$$

where

$$\mathbf{A}_{cl} = \begin{pmatrix} \mathbf{A}_z + d\mathbf{B}_z\mathbf{C}_z & \mathbf{B}_z\mathbf{C} \\ \mathbf{B}\mathbf{C}_z & \mathbf{A} \end{pmatrix}, \quad \mathbf{B}_r = \begin{pmatrix} \mathbf{B}_z \\ \mathbf{0} \end{pmatrix}. \tag{7}$$

There exist different scaling schemes for preventing variables from overflow. One of the popularly used schemes is the l_2-scaling, which means that each element of the controller state vector $\mathbf{x}_k$ should have a unit variance when the input r_k is a white noise with a unit variance. This can be achieved if

$$\bar{\mathbf{K}}(j,j) = 1, \quad j = q+1, q+2, \cdots, q+p \tag{8}$$

where $\bar{\mathbf{K}}$ is given by

$$\bar{\mathbf{K}} = \sum_{k=0}^{+\infty} \mathbf{A}_{cl}^k \mathbf{B}_r \mathbf{B}_r^T \left(\mathbf{A}_{cl}^T\right)^k, \tag{9}$$

satisfying

$$\bar{\mathbf{K}} = \mathbf{A}_{cl}\bar{\mathbf{K}}\mathbf{A}_{cl}^T + \mathbf{B}_r\mathbf{B}_r^T. \tag{10}$$

In the sequel, all the realizations under discussion are assumed l_2-scaled.

2.2 Roundoff noise analysis

Another practical issue in digital implementation is rounding. Assume that a fixed-point implementation is considered, where the coefficients in the realization are implemented with B_c bits. If the microcontroller's accumulator has $B_c + B_s$ bits, any intermediate signal has to be rounded into B_s bits *before* multiplied with a non-trivial coefficient. Let $\mathbf{P} = \mathbf{P}_i + \mathbf{P}_f$, where $\mathbf{P}_i$ is a trivial matrix and $\mathbf{P}_f$ contains no trivial elements except 0s, for $\mathbf{P} = \mathbf{A}, \mathbf{B}, \mathbf{C}$. Then a more practical digital controller model is[2]

$$\left.\begin{aligned} \mathbf{x}_{k+1}^* &= \mathbf{A}_i\mathbf{x}_k^* + \mathbf{A}_f Q[\mathbf{x}_k^*] + \mathbf{B}_i u_k^* + \mathbf{B}_f Q[u_k^*] \\ y_k^* &= \mathbf{C}_i\mathbf{x}_k^* + \mathbf{C}_f Q[\mathbf{x}_k^*] + dQ[u_k^*] \end{aligned}\right\} \tag{11}$$

where $Q[x]$ is the quantizer that rounds the number x to B_s bits.

Denote

$$\boldsymbol{\epsilon}_x(k) \triangleq Q[\mathbf{x}_k^*] - \mathbf{x}_k^*, \quad \epsilon_u(k) \triangleq Q[u_k^*] - u_k^* \tag{12}$$

[2]Here, it is assumed that the ideal controller realization can be implemented exactly with B_c bits.

as the quantization errors. Traditionally, these quantization errors are modeled as statistically independent white sequences (see, e.g., [20],[21]) and

$$E\left[\begin{pmatrix} \boldsymbol{\epsilon}_x(k+m) \\ \epsilon_u(k+m) \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon}_x(m) \\ \epsilon_u(m) \end{pmatrix}^T\right] = \sigma_0^2 \delta_d(k)\mathbf{I}, \tag{13}$$

where $E[\cdot]$ denotes the ensemble average operator, ^{T} the transposed operator, $\delta_d(k)$ the discrete Dirac delta function, $\sigma_0^2 = 2^{-2B_s}/12$ and $\mathbf{I}$ is the identity matrix of proper dimension. It follows from (1) and (11) that

$$\left.\begin{aligned} \mathbf{e}_x(k+1) &= \mathbf{A}\mathbf{e}_x(k) + \mathbf{B}e_u(k) + \mathbf{A}_f\boldsymbol{\epsilon}_x(k) + \mathbf{B}_f\epsilon_u(k) \\ e_y(k) &= \mathbf{C}\mathbf{e}_x(k) + de_u(k) + \mathbf{C}_f\boldsymbol{\epsilon}_x(k) + d\epsilon_u(k) \end{aligned}\right\} \tag{14}$$

where

$$e_u(k) = u_k^* - u_k, \;\; \mathbf{e}_x(k) = \mathbf{x}_k^* - \mathbf{x}_k, \;\; e_y(k) = y_k^* - y_k. \tag{15}$$

Noting $e_u(k) = P_d(z)e_y(k)$, one can show that

$$e_u(k) = \mathbf{C}_{cl}(z\mathbf{I} - \mathbf{A}_{cl})^{-1}\mathbf{B}_{cl} \begin{pmatrix} \boldsymbol{\epsilon}_x(k) \\ \epsilon_u(k) \end{pmatrix} \triangleq \mathbf{H}_u(z) \begin{pmatrix} \boldsymbol{\epsilon}_x(k) \\ \epsilon_u(k) \end{pmatrix} \tag{16}$$

where $\mathbf{A}_{cl}$ is given in (7) and

$$\mathbf{B}_{cl} = \begin{pmatrix} \mathbf{B}_z\mathbf{C}_f & d\mathbf{B}_z \\ \mathbf{A}_f & \mathbf{B}_f \end{pmatrix}, \;\; \mathbf{C}_{cl}^T = \begin{pmatrix} \mathbf{C}_z^T \\ \mathbf{0} \end{pmatrix}. \tag{17}$$

Since the quantization errors $\boldsymbol{\epsilon}_x(k)$ and $\epsilon_u(k)$ are statistically independent white sequences (see (13)), 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 [22], one has

$$\Gamma_u(z) = \mathbf{H}_u(z)\mathbf{H}_u^T(z^{-1})\sigma_0^2 \tag{18}$$

and

$$\begin{aligned} E[e_u^2(k)] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_u(e^{j\omega})d\omega = tr\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{H}_u^T(e^{-j\omega})\mathbf{H}_u(e^{j\omega})d\omega\right]\sigma_0^2 \\ &\triangleq tr\left[\mathbf{B}_{cl}^T\bar{\mathbf{W}}\mathbf{B}_{cl}\right]\sigma_0^2 \end{aligned} \tag{19}$$

where $tr[\cdot]$ denotes the trace operator and $\bar{\mathbf{W}}$ is the observability gramian of the realization of $\mathbf{H}_u(z)$ (see (16)) defined by:

$$\bar{\mathbf{W}} = \sum_{k=0}^{+\infty} \left(\mathbf{A}_{cl}^T\right)^k \mathbf{C}_{cl}^T\mathbf{C}_{cl}\mathbf{A}_{cl}^k, \tag{20}$$

which satisfies

$$\bar{\mathbf{W}} = \mathbf{A}_{cl}^T\bar{\mathbf{W}}\mathbf{A}_{cl} + \mathbf{C}_{cl}^T\mathbf{C}_{cl}. \tag{21}$$

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

$$G \triangleq \frac{E[e_u^2(k)]}{\sigma_0^2} \tag{22}$$

and therefore

$$G = tr\left[\mathbf{B}_{cl}^T \bar{\mathbf{W}} \mathbf{B}_{cl}\right]. \tag{23}$$

For a given l_2-scaled controller realization, one can compute the corresponding roundoff noise gain. Different controller realizations yield different G values. The interesting problem is to identify those realizations that give the minimum G. This problem is very difficult due to the decomposition $\mathbf{P} = \mathbf{P}_i + \mathbf{P}_f$ for $\mathbf{P} = \mathbf{A}, \mathbf{B}, \mathbf{C}$. In the remainder of this section, we will consider the case where $\mathbf{P}_i = \mathbf{0}$ for $\mathbf{P} = \mathbf{A}, \mathbf{B}, \mathbf{C}$, which corresponds to the classical shift operator parameterizations and makes the optimization problem tractable.

2.3 Realization dependence and optimal realizations

Let $(\mathbf{A}_{cl}, \mathbf{B}_{cl}, \mathbf{C}_{cl})$ and $(\mathbf{A}_{cl}^0, \mathbf{B}_{cl}^0, \mathbf{C}_{cl}^0)$ be two realizations of $\mathbf{H}_u(z)$ defined by (7), (16) and (17), corresponding to the two digital controller realizations $\mathbf{R} \triangleq (\mathbf{A}, \mathbf{B}, \mathbf{C}, d)$ 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

$$\left.\begin{aligned} \mathbf{A}_{cl} &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}^{-1} \mathbf{A}_{cl}^0 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}, \\ \mathbf{B}_{cl} &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}^{-1} \mathbf{B}_{cl}^0 \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \\ \mathbf{C}_{cl} &= \mathbf{C}_{cl}^0 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}. \end{aligned}\right\} \tag{24}$$

It then turns out that

$$\mathbf{H}_u(z) = \mathbf{H}_u^0(z) \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{25}$$

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

$$\bar{\mathbf{W}} = \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^T \bar{\mathbf{W}}^0 \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{26}$$

where $\bar{\mathbf{W}}^0$ is similar to $\bar{\mathbf{W}}$ defined in (21) but corresponds to the controller realization $\mathbf{R}_0$.

Clearly,

$$\mathbf{B}_{cl}^T \bar{\mathbf{W}} \mathbf{B}_{cl} = \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^T \left(\mathbf{B}_{cl}^0\right)^T \bar{\mathbf{W}}^0 \mathbf{B}_{cl}^0 \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{27}$$

Let

$$\left(\mathbf{B}_{cl}^0\right)^T \bar{\mathbf{W}}^0 \mathbf{B}_{cl}^0 \triangleq \begin{pmatrix} \mathbf{W}_0 & \mathbf{W}_{12}^0 \\ \mathbf{W}_{21}^0 & \mathbf{Q}_0 \end{pmatrix} \tag{28}$$

have the same partition as

$$\begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{29}$$

It is easy to see that

$$G = tr[\mathbf{T}^T\mathbf{W}_0\mathbf{T}] + tr[\mathbf{Q}_0] \tag{30}$$

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

Similarly, one has

$$\bar{\mathbf{K}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}^{-1} \bar{\mathbf{K}}^0 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}^{-T} \tag{31}$$

Let

$$\bar{\mathbf{K}} \triangleq \begin{pmatrix} \mathbf{K} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K} \end{pmatrix} \tag{32}$$

and

$$\bar{\mathbf{K}}^0 \triangleq \begin{pmatrix} \mathbf{K}^0 & \mathbf{K}_{12}^0 \\ \mathbf{K}_{21}^0 & \mathbf{K}^0 \end{pmatrix} \tag{33}$$

have the same partition as

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}. \tag{34}$$

Clearly, $\mathbf{K} = \mathbf{T}^{-1}\mathbf{K}^0\mathbf{T}^{-T}$.

The optimal realizations of the digital controller are the solutions to the following minimization problem:

$$\min_{\substack{\mathbf{R}\in S_{C_d} \\ (\mathbf{T}^{-1}\mathbf{K}^0\mathbf{T}^{-T})(j,j)=1,\ \forall j}} \quad G \tag{35}$$

In [9], a solution to the above problem was given. It has been found that the optimal realizations are generally fully parametrized with non-trivial parameters. This is a disadvantage since it will increase the implementation complexity and slow down the processing. From a practical point of view, it is desired to implement the digital controller with such a realization that not only has a very robust performance against the FWL effects but also possesses as few non-trivial parameters as possible. In the next section, based on the concept of the polynomial approach we will derive a new controller structure which has very nice numerical properties and a very sparse form.

3 Derivation of a new sparse controller structure

The digital controller $C_d(z)$ is usually given by its transfer function

$$C_d(z) = \frac{b_0 z^p + b_1 z^{p-1} + \cdots + b_{p-1}z + b_p}{z^p + a_1 z^{p-1} + \cdots + a_{p-1} + a_p} \triangleq \frac{N(z)}{D(z)}. \tag{36}$$

Define

$$\rho_j \triangleq \frac{z-\gamma_j}{\Delta_j},\ j=1,2,\cdots,p, \tag{37}$$

where γ_j, $\forall j$, takes values from the trivial parameter set $\{-1,0,1\}$, and $\{\Delta_j > 0\}$, as to be seen later, are determined for l_2-scaling. The denominator and numerator of $C_d(z)$ given by (36) can be re-parametrized with

$$D(z) = \mathcal{K}\left[\prod_{j=1}^{p}\rho_j + \alpha_1\prod_{j=2}^{p}\rho_j + \cdots + \alpha_{p-1}\rho_p + \alpha_p\right], \tag{38}$$

$$N(z) = \mathcal{K}\left[\beta_0\prod_{j=1}^{p}\rho_j + \beta_1\prod_{j=2}^{p}\rho_j + \cdots + \beta_{p-1}\rho_p + \beta_p\right], \tag{39}$$

where $\mathcal{K} = \prod_{j=1}^{p}\Delta_j$. Therefore, $C_d(z)$ can be re-parametrized with $\{\alpha_m, \beta_m\}$ in the polynomials $\{\rho_j\}$, called *polynomial operators*:

$$C_d(z) = \frac{\beta_0 + \beta_1\rho_1^{-1} + \cdots + \beta_{p-1}\prod_{j=1}^{p-1}\rho_j^{-1} + \beta_p\prod_{j=1}^{p}\rho_j^{-1}}{1 + \alpha_1\rho_1^{-1} + \cdots + \alpha_{p-1}\prod_{j=1}^{p-1}\rho_j^{-1} + \alpha_p\prod_{j=1}^{p}\rho_j^{-1}}. \tag{40}$$

Denoting

$$\mathbf{V}_a \triangleq \begin{pmatrix}1\\a_1\\\vdots\\a_p\end{pmatrix},\ \mathbf{V}_b \triangleq \begin{pmatrix}b_0\\b_1\\\vdots\\b_p\end{pmatrix},\ \mathbf{V}_\alpha \triangleq \begin{pmatrix}1\\\alpha_1\\\vdots\\\alpha_p\end{pmatrix},\ \mathbf{V}_\beta \triangleq \begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix}, \tag{41}$$

one has

$$\left.\begin{aligned}\mathbf{V}_a &= \mathcal{K}\mathbf{M}\mathbf{V}_\alpha\\ \mathbf{V}_b &= \mathcal{K}\mathbf{M}\mathbf{V}_\beta\end{aligned}\right\} \tag{42}$$

or

$$\left.\begin{aligned}\mathbf{V}_\alpha &= \mathcal{K}^{-1}\mathbf{M}^{-1}\mathbf{V}_a\\ \mathbf{V}_\beta &= \mathcal{K}^{-1}\mathbf{M}^{-1}\mathbf{V}_b\end{aligned}\right\} \tag{43}$$

where $\mathbf{M} \in R^{(p+1)\times(p+1)}$ is a lower triangular matrix whose mth column is determined by the coefficients of the polynomial $\prod_{j=m}^{p}\rho_j$ for $m=1,2,\cdots,p$ and $M(p+1,p+1)=1$.

The corresponding input and output relationship to (40) is then

$$\begin{aligned} y_k &= -\alpha_1\rho_1^{-1}y_k - \cdots - \alpha_{p-1}\prod_{j=1}^{p-1}\rho_j^{-1}y_k - \alpha_p\prod_{j=1}^{p}\rho_j^{-1}y_k \\ &\quad +\beta_0u_k + \beta_1\rho_1^{-1}u_k + \cdots + \beta_{p-1}\prod_{j=1}^{p-1}\rho_j^{-1}u_k + \beta_p\prod_{j=1}^{p}\rho_j^{-1}u_k. \end{aligned} \tag{44}$$

It is easy to see that the output can be computed with the following equations (also see Figure 2)

$$\left.\begin{aligned} y_k &= \beta_0u_k + w_{1,k},\\ w_{j,k} &= \rho_j^{-1}[\beta_ju_k - \alpha_jy_k + w_{j+1,k}],\\ w_{p,k} &= \rho_p^{-1}[\beta_pu_k - \alpha_py_k]. \end{aligned}\right\} \tag{45}$$

3.1 Equivalent state-space realization

From the definition (37), one can implement ρ_j^{-1} with the realization depicted in Figure 3, where the "input" to the operator ρ_j^{-1} is $\beta_j u_k - \alpha_j y_k + w_{j+1,k}$ for $j = p, p-1, \cdots, 1$ with $w_{p+1,k} = 0$ (see (45)). We choose $\{x_{j,k}\}$ indicated in Figure 3 as the state variables and denote $\mathbf{x}_k$ as the state vector. Noting

$$\left.\begin{aligned} x_{j,k+1} &= \gamma_j x_{j,k} + \beta_j u_k - \alpha_j y_k + w_{j+1,k}, \\ w_{j,k} &= \Delta_j x_{j,k}, \\ y_k &= \beta_0 u_k + \Delta_1 x_{1,k}, \end{aligned}\right\} \tag{46}$$

it can be shown that (45) is equivalent to the following state-space realization

$$\left.\begin{aligned} \mathbf{x}_{k+1} &= \tilde{\mathbf{A}}\mathbf{x}_k + \tilde{\mathbf{B}}u_k \\ y_k &= \tilde{\mathbf{C}}\mathbf{x}_k + du_k \end{aligned}\right\} \tag{47}$$

where

$$\tilde{\mathbf{A}} = \begin{pmatrix} \gamma_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \gamma_2 & 0 & \cdots & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & \gamma_{p-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \gamma_p \end{pmatrix} + \begin{pmatrix} -\Delta_1\alpha_1 & \Delta_2 & 0 & \cdots & 0 & 0 \\ -\Delta_1\alpha_2 & 0 & \Delta_3 & \cdots & 0 & 0 \\ & & & \vdots & & \\ -\Delta_1\alpha_{p-1} & 0 & 0 & \cdots & 0 & \Delta_p \\ -\Delta_1\alpha_p & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \triangleq \tilde{\mathbf{A}}_i + \tilde{\mathbf{A}}_f, \tag{48}$$

$$\tilde{\mathbf{B}} = \mathbf{V}_\beta(2:p+1) - \beta_0\mathbf{V}_\alpha(2:p+1), \quad \tilde{\mathbf{C}} = \begin{pmatrix} \Delta_1 & 0 & \cdots & 0 \end{pmatrix}, \tag{49}$$

with $\mathbf{V}_x(2:p+1)$ representing the vector formed with 2nd to $(p+1)$th elements of the vector $\mathbf{V}_x$ for $x = \alpha, \beta$. Clearly, the controller realization $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, d)$ is very sparse with $\tilde{\mathbf{A}}_i$ diagonal and possesses only $3p+1$ non-trivial parameters.

It can be verified easily that the realization $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, d)$ can be transformed with a diagonal similarity transformation, denoted as $\mathbf{T}_{sc}$, into a realization $(\tilde{\mathbf{A}}^0, \tilde{\mathbf{B}}^0, \tilde{\mathbf{C}}^0, d)$, which has exactly the same form as $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, d)$ given in (48) and (49) except that the parameters $\{\alpha_j, \beta_j\}$ correspond to $\Delta_j = 1, \forall j$, that is $\tilde{\mathbf{A}}^0 = \mathbf{T}_{sc}^{-1}\tilde{\mathbf{A}}\mathbf{T}_{sc}$, $\tilde{\mathbf{B}}^0 = \mathbf{T}_{sc}^{-1}\tilde{\mathbf{B}}$, $\tilde{\mathbf{C}}^0 = \tilde{\mathbf{C}}\mathbf{T}_{sc}$, or

$$\tilde{\mathbf{A}} = \mathbf{T}_{sc}\tilde{\mathbf{A}}^0\mathbf{T}_{sc}^{-1}, \ \tilde{\mathbf{B}} = \mathbf{T}_{sc}\tilde{\mathbf{B}}^0, \ \tilde{\mathbf{C}} = \tilde{\mathbf{C}}^0\mathbf{T}_{sc}^{-1}, \tag{50}$$

where

$$\mathbf{T}_{sc} = diag(d_1, d_2, \cdots, d_{p-1}, d_p), \ \ d_j = \prod_{m=1}^{j} \Delta_m^{-1}, \ \forall j. \tag{51}$$

As mentioned before, the controller realization (47) should be l_2-scaled and this can be achieved by choosing $\{\Delta_j\}$ properly. In fact, with (7), (24) and (10) it can be shown that

$$\tilde{\mathbf{K}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{sc} \end{pmatrix} \tilde{\mathbf{K}}_0 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{sc} \end{pmatrix}^T, \tag{52}$$

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{K}}_0$ are the controllability gramians, corresponding to the controller realizations $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, d)$ and $(\tilde{\mathbf{A}}^0, \tilde{\mathbf{B}}^0, \tilde{\mathbf{C}}^0, d)$, respectively. Clearly, the l_2 scaling is achieved if $d_j^2 \tilde{\mathbf{K}}^0(j,j) = 1$ for $j = q+1, q+2, \cdots, q+p$, which leads to

$$\Delta_1 = \sqrt{\tilde{\mathbf{K}}_0(q+1, q+1)},\ \Delta_{j-q} = \sqrt{\frac{\tilde{\mathbf{K}}_0(j,j)}{\tilde{\mathbf{K}}_0(j-1, j-1)}},\ j = q+2, q+3, \cdots, q+p. \tag{53}$$

3.2 Optimal polynomial operators

As mentioned before, γ_j takes value from the finite set S_γ:

$$S_\gamma \stackrel{\triangle}{=} \{-1,\ 0,\ 1\}. \tag{54}$$

For given $\{\gamma_j\}$ and $\Delta_j = 1,\ \forall j$, one can compute the corresponding $\{\alpha_j, \beta_j\}$ with (43), and hence the realization $(\tilde{\mathbf{A}}^0, \tilde{\mathbf{B}}^0, \tilde{\mathbf{C}}^0, d)$ and $\tilde{\mathbf{K}}_0$ can be computed with (48)-(49) and (10), respectively. The l_2 scaling matrix $\mathbf{T}_{sc}$ can be obtained with (53) and hence the l_2-scaled realization $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, d)$ can be computed steadily, with which the corresponding $\tilde{\mathbf{W}}$ can be evaluated with (21). Therefore, the roundoff noise gain G can be calculated with (23), where $\mathbf{B}_{cl}$ is given by (17) with $\mathbf{A}_f$ by (48) and $\mathbf{B}_f = \tilde{\mathbf{B}}, \mathbf{C}_f = \tilde{\mathbf{C}}$.

Different sets of $\{\gamma_j\}$ yield different values of roundoff noise gain. The interesting problem is to find the optimal sets of $\{\gamma_j\}$, which are the solutions to

$$\min_{\gamma_j \in S_\gamma,\ \forall j} G. \tag{55}$$

Though G is a highly nonlinear function of $\{\gamma_j\}$, the problem can be solved easily since the space $\{\gamma_j :\ \gamma_j \in S_\gamma\}$ is finite.

4 A design example

We now present a design example to illustrate the design procedure. The digital plant $P_d(z)$ and controller $C_d(z)$ are presented with the following canonical forms, denoted as $\mathbf{R}_z$ and $\mathbf{R}_c$, respectively. $\mathbf{R}_z$ is given by:

$$\mathbf{A}_z = \begin{pmatrix} 3.3555 & 1 & 0 & 0 & 0 \\ -4.9154 & 0 & 1 & 0 & 0 \\ 4.0734 & 0 & 0 & 1 & 0 \\ -1.8227 & 0 & 0 & 0 & 1 \\ 0.3093 & 0 & 0 & 0 & 0 \end{pmatrix},\ \mathbf{B}_z = \begin{pmatrix} -0.1183 \\ -0.7249 \\ 0.6878 \\ -0.6510 \\ -0.0425 \end{pmatrix},$$

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

$\mathbf{R}_c$ is given by:

$$
\begin{aligned}
\mathbf{A}_c &= \begin{pmatrix} 2.1016 & 1 & 0 & 0 & 0 & 0 \\ -2.2306 & 0 & 1 & 0 & 0 & 0 \\ 1.4467 & 0 & 0 & 1 & 0 & 0 \\ -0.4901 & 0 & 0 & 0 & 1 & 0 \\ 0.1954 & 0 & 0 & 0 & 0 & 1 \\ -0.0231 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ \mathbf{B}_c = \begin{pmatrix} 0.1971 \\ -0.7401 \\ 1.1527 \\ -1.0041 \\ 0.4857 \\ -0.0912 \end{pmatrix}, \\
\mathbf{C}_c &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ d = 0.0460.
\end{aligned}
$$

With $\mathbf{R}_c$ as the initial digital controller realization and the plant $\mathbf{R}_z$, we compute the corresponding $\bar{\mathbf{K}}_0$ and $\bar{\mathbf{W}}_0$, based on which $\mathbf{K}_0$ and $\mathbf{W}_0$ can be obtained. Using the same procedure as given in [9], an optimal realization in the shift operator, denoted as $\mathbf{R}_z^{opt}$, can be found by solving (35). The corresponding minimum roundoff noise gain is 6.0351×10^6.

With $\mathbf{R}_c$, one can obtain the controller transfer function coefficients $\mathbf{V}_a$ and $\mathbf{V}_b$. For a given set of $\{\gamma_j\}$, one can compute the $\mathbf{V}_\alpha$ and $\mathbf{V}_\beta$ for $\Delta_j = 1, \forall j$, using (43), and hence the corresponding controller realization $(\tilde{\mathbf{A}}^0, \tilde{\mathbf{B}}^0, \tilde{\mathbf{C}}^0)$. The corresponding controllability and observability gramians defined by (10) and (21), denoted as $\tilde{\mathbf{K}}^0$ and $\tilde{\mathbf{W}}^0$, respectively, can then be computed with the MATLAB command $dgram.m$. With $\tilde{\mathbf{K}}^0$, the coupling coefficients $\{\Delta_j\}$ can be obtained with (53). With the obtained $\{\Delta_j\}$, one can compute the corresponding $\mathbf{V}_\alpha$ and $\mathbf{V}_\beta$ using (43) hence the corresponding controller realization $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ as well as the observability gramian $\tilde{\mathbf{W}}$. The roundoff noise gain G is computed with (23) with $\bar{\mathbf{W}}$ replaced by $\tilde{\mathbf{W}}$ and $\mathbf{B}_{cl}$ given by (17), where $\mathbf{A}_f = \tilde{\mathbf{A}}_f$ as defined in (48), $\mathbf{B}_f = \tilde{\mathbf{B}}$ and $\mathbf{C}_f = \tilde{\mathbf{C}}$ as defined in (49). The problem of finding optimal polynomial operators (55) can be solved with an exhaustive research (a total of $3^6 = 729$ combinations). Table 1 shows the roundoff noise gains for four different sets of polynomial operators, in which $(1\ \ 1\ \ 0\ \ 0\ \ 1\ \ 1)$ is known to yield the optimal polynomial operators.

Comment 4.1: The combination $(0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0)$ for $\{\gamma_l\}$ corresponds to the (diagonally) l_2-scaled version of the canonical controller realization $\mathbf{R}_c$ in the shift operator. The corresponding roundoff noise gain is much higher than that of the optimal realization $\mathbf{R}_z^{opt}$ in the shift operator. While the combination $(1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1)$ for $\{\gamma_l\}$ is the (diagonally) l_2-scaled version of the canonical controller realization in the delta operator. The corresponding roundoff noise gain is much smaller than that of the optimal fully parametrized realization in the shift operator. For the controller realization corresponding to the operator polynomial operators $(1\ \ 1\ \ 0\ \ 0\ \ 1\ \ 1)$, it yields a roundoff noise which is just half of that produced by the delta-operator based realization mentioned above. In fact, it is also simpler than this delta-operator based realization for implementation in which two additions can be saved.

5 Conclusions

In this paper, the effect of rounding errors that occur in the controller of a digital feedback control system has been investigated. Our contribution is three-fold. Firstly, we have proposed a new implementation model of a state-space controller realization, where each coefficient matrix of the realization is separated into a trivial part, which only contains elements from $\{-1, 0, 1\}$ and hence causes no rounding error, and a non-trivial part. Secondly, based on this proposed model, we have analyzed output deviation of the closed-loop system due to the roundoff noise in the digital controller. An analytical expression for the roundoff noise gain has been obtained. The problem of identifying the optimal realizations (in the shift operator) has also been solved. Our third contribution, which is the most interesting one, is to have derived a new sparse controller realization based on a polynomial operator approach, which is a generalization of the direct forms in the classical shift operator and the prevailing delta operator. The problem of finding optimal polynomial operators has been investigated. It is shown that, with the proposed sparse realization, the optimal polynomial operators can outperform not only the shift- and delta-operators but also the fully parametrized optimal controller realizations (in shift operator).

References

[1] G. Li and M. Gevers, “Optimal finite precision implementation of a state-estimate feedback controller,” *IEEE Trans. Circuits and Systems*, Vol.CAS-38, No.12, pp.1487–1499, Dec. 1990.

[2] M. Gevers and G. Li, *Parametrisations in Control, Estimation and Filtering Problems: Accuracy Aspects*. Springer Verlag, London, Communication and Control Engineering Series, 1993.

[3] I.J. Fialho and T.T. Georgiou, “On stability and performance of sampled-data systems subject to wordlength constraint,” *IEEE Trans. Automatic Control*, Vol.39, pp.2476–2481, Dec. 1994.

[4] I.J. Fialho and T.T. Georgiou, “Optimal finite wordlength digital controller realizations,” in *Proc. American Control Conf.* (San Diego, USA), pp.4326–4327, June 2-4, 1999.

[5] G. Li, “On the structure of digital controllers with finite word length consideration,” *IEEE Trans. Automatic Control*, Vol.43, No.5, pp.689–693, May 1998.

[6] 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 Applications*, Vol.145, No.5, pp.472–478, 1998.

[7] D. Williamson and K. Kadiman, "Optimal finite wordlength linear quadratic regulation," *IEEE Trans. Automatic Control*, Vol.34, No.12, pp.1218–1228, 1989.

[8] K. Liu, R. Skelton and K. Grigoriadis, "Optimal controllers for finite wordlength implementation," *IEEE Trans. Automatic Control*, Vol.37, pp.1294–1304, 1992.

[9] G. Li, J. Wu, S. Chen, and K.Y. Zhao, "Optimum structures of digital controllers in sampled-data systems: a roundoff noise analysis," *IEE Proc. Control Theory and Applications*, Vol.149, No.3, pp.247–255, May 2002.

[10] S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," *IEEE Trans. Automatic Control*, Vol.44, No.11, pp.2149–2153, 1999.

[11] J. Wu, S. Chen, G. Li and J. Chu, "Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems," *IEEE Trans. Automatic Control*, Vol.45, No.8, pp.1550–1554, Aug. 2000.

[12] S. Chen, J. Wu, R.H. Istepanian, J. Chu, and J.F. Widborne, "Optimising stability bounds of finite-precision controller structures for sampled-data systems in the delta operator domain," *IEE Proc. Control Theory and Applications*, Vol.146, No.6, pp.517-526, 1999.

[13] S. Chen, R.H. Istepanian, J. Wu and J. Chu, "Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators," *Systems and Control Letters*, Vol.40, No.3, pp.153–163, 2000.

[14] J. Wu, S. Chen, G. Li, R.H. Istepanian, and J. Chu, "Shift and delta operator realizations for digital controllers with finite word length considerations," *IEE Proc. Control Theory and Applications*, Vol.147, No.6, pp.664-672, Nov. 2000.

[15] G. Ami and U. Shaked, "Small roundoff realization of fixed-point digital filters and controllers," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.ASSP-36, No.6, pp.880–891, June 1988.

[16] J. Wu, S. Chen, G. Li and J. Chu, "Constructing sparse realizations of finite-precision digital controllers based on a closed-loop stability related measure," *IEE Proc. Control Theory and Applications*, Vol.150, No.1, pp.61–68, 2003.

[17] V. Peterka, "Control of uncertain processes: applied theory and algorithms," *Kybernetika*, Vol. 22, pp.1-102, 1986.

[18] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach*, Prentice Hall, Englewood Cliffs, New Jersey, 1990.

[19] R.A. Roberts and C.T. Mullis, *Digital Signal Processing*. Addison Wesley, 1987.

[20] C.T. Mullis and R.A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits and Systems*, Vol.CAS-23, pp.551–562, Sept. 1976.

[21] S.Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.ASSP-25, No.4, pp.273–281, Aug. 1977.

[22] K.J. Åström, *Introduction to Stochastic Control Theory*. Academic Press, New York, 1970.

Table 1: Four different sets of polynomial operators with corresponding roundoff noise gains

γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	G
-1	-1	-1	-1	-1	-1	5.4318×10^{16}
0	0	0	0	0	0	5.9417×10^{13}
1	1	0	0	1	1	5.1128×10^{4}
1	1	1	1	1	1	1.2759×10^{5}

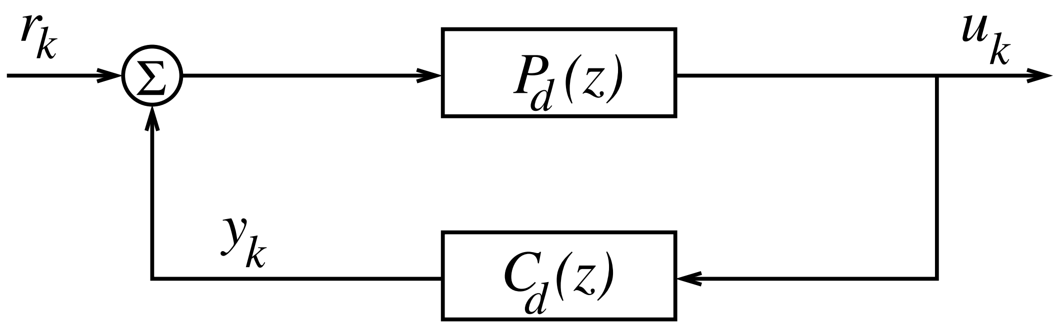

Figure 1: Block-diagram of a discrete-time feedback control system. Here $P_d(z)$ denotes the discrete-time plant model, $C_d(z)$ the digital controller, u_k is the digital controller input, y_k the digital controller output and r_k the input signal of the closed-loop system.

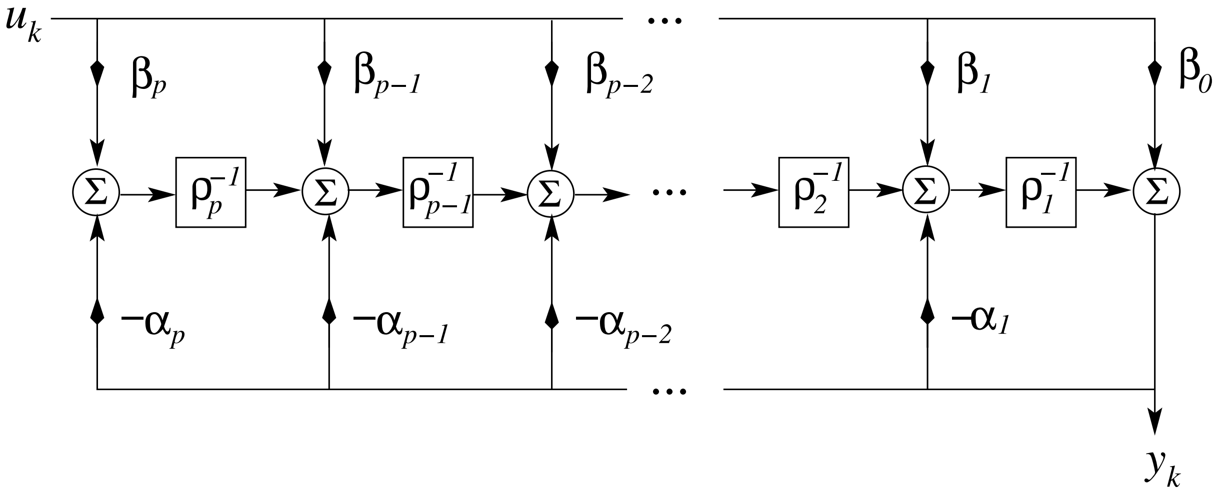

Figure 2: The structure of input-output realization in polynomial operator.

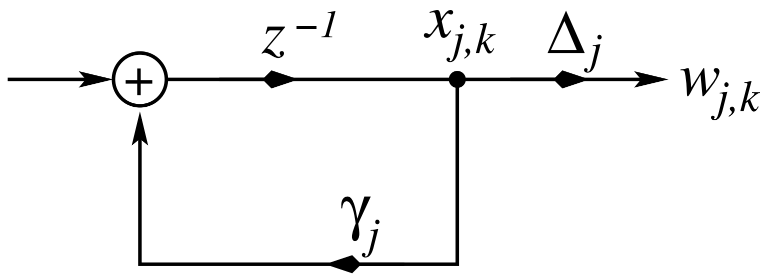

Figure 3: A realization of operator ρ_j^{-1}.