# A Generalized Direct-Form Delta Operator-Based IIR Filter with Minimum Noise Gain and Sensitivity

Ngai Wong and Tung-Sang Ng

***Abstract*—This brief presents the derivation of an arbitrary order delta operator-based direct-form IIR filter with minimum roundoff noise gain and sensitivity. It utilizes the concept of different *coupling coefficients* at different branch nodes for better noise gain suppression. Two possible structures for realizing the inverse delta operator are considered and procedures for calculating the optimal filter coefficients are given. By means of state-space representation and matrix manipulation, it is also shown that expressions for sensitivity measures of different filter coefficients and their corresponding roundoff noise gain expressions are the same. This enables the simultaneous minimization of sensitivity and noise power for the proposed generalized filter structure.**

***Index Terms*—Delta operator, direct-form, IIR filter, minimization, roundoff noise, sensitivity.**

## I. Introduction

The advantages of delta operator-based implementation over the conventional shift operator approach have recently gained attention due to the work of Goodwin and Middleton [1], [2]. In addition to the interesting unification of continuous and discrete time models, the numerical benefits of using a delta operator in implementing digital filters have received a lot of attention [3]–[8]. The major potential lies in fast sampling systems using fixed-point arithmetic where filter poles cluster toward unity in the $z$ plane. Such a situation often causes the system to become numerically ill-conditioned. Delta operator (defined as $\delta = (z-1)/\Delta$) based filters alleviate this problem by characterizing the difference between these poles and unity. This usually brings about much less roundoff noise gain and more robust coefficient and frequency sensitivities.

On the practical side, such as in ASIC design, an optimal (generally fully parameterized) state-space delta operator filter design requires a relatively large number of components and computations and is less favorable than the much simpler direct-forms (DFs). Extensive comparative study of different DF delta structures has been carried out in [3]. It was found that, of all the DF structures, the delta DFII transposed ($\delta$DFIIt) structure shows the lowest roundoff noise gain and outperforms both the conventional shift operator DFs, as well as narrow-band state-space structures. However, to the knowledge of the authors, only cascaded second-order $\delta$DFIIt sections have been considered so far. In this brief, generalization to an arbitrary order $\delta$DFIIt structure is presented. An immediate advantage is that higher order filters generally provide more savings in hardware than second-order cascades. Two structures for realizing an inverse delta operator, $\delta^{-1} = \Delta z^{-1}/(1-z^{-1})$, are considered and their respective influences on the filter's noise and sensitivity characteristics are compared. State-space representation and matrix manipulation similar to that of [8]–[14] ease tedious transfer function derivation and make the resultant expressions scalable for any order. Moreover, by extending the idea in a previous work [6] on a second-order $\delta$DFIIt section, the multiplicative constant $\Delta$ (called *coupling coefficient* in this brief) in $\delta^{-1}$ operators are separated from the traditional definition into a novel diagonal *coupling matrix*, which features various coupling coefficients at different nodes. This enables the utilization of the dynamic range at filter nodes where scaling is necessary and further decreases the output noise gain.

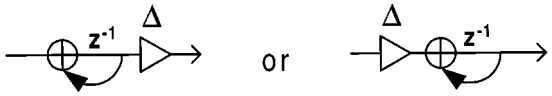


Fig. 1. Two realizations of a $\delta^{-1}$ operator.

Another important topic in filter design is filter coefficient sensitivity. Studies in the past focused on optimal state-space realization where all elements in every state-space matrix contribute to the sensitivity measure [8], [12]–[14]. The $\delta$DFIIt structure, whose coefficient sensitivity has not been formulated in the literature so far, represents a special form of sparse realization in which only certain elements within the state-space matrices will affect the transfer function. In this brief, specific sensitivity measures are defined and expressions for sensitivity due to different parameters are obtained.

Techniques described in this brief may also find application in a fast growing branch of DSP known as sigma-delta ($\Sigma\Delta$) modulation [15]. A widely adopted topology for building high-order bandpass $\Sigma\Delta$ modulators [16] is known as the cascade-of-resonators architecture [17], [18]. This topology can be easily modified from the $\delta$DFIIt structure by adding state variable feedback branches. If zero resonator frequency is considered, such as in baseband $\Sigma\Delta$ modulation, the cascade-of-resonators architecture is equivalent to the $\delta$DFIIt structure. Moreover, motivated by the potential benefits of eliminating the decimation and interpolation filters in the processing of $\Sigma\Delta$ A/D converted one-bit signal, single-bit digital signal processing at oversampled rates can also make use of this $\delta$DFIIt structure for direct bit-stream filtering [19]–[21]. It is therefore valuable to completely characterize an arbitrary order $\delta$DFIIt structure in terms of its roundoff noise gain and coefficient sensitivity and to devise procedures to build an optimal $\delta$DFIIt filter.

The rest of the brief is organized as follows. In Section II, a generalized $\delta$DFIIt structure is presented. Two possible realizations of a $\delta^{-1}$ operator are described and the conversion of shift domain filters into the delta domain is studied. Section III considers the first $\delta^{-1}$ operator realization and derives the roundoff noise gain formulas. Section IV examines dynamic range scaling and its relation with the corresponding noise gain. In Section V, the second $\delta^{-1}$ operator realization is considered. Section VI presents the sensitivity analysis and shows that there is strong correlation between the roundoff noise gain formulae and the sensitivity expressions, and both attain their minima simultaneously. Numerical examples are given in Section VII and concluding remarks are drawn in Section VIII.

## II. Structural Transformation

This section describes the transformation from an initial reference filter into a $\delta$DFIIt structure. To begin with, the hardware implementation of an $\delta^{-1}$ operator can take on two forms, as shown in Fig. 1. Both forms represent the same operator and our analysis will begin with the first form having its coupling coefficient $\Delta$ after the discrete integrator. In our notation, every $\delta^{-1}$ operator is associated with a different $\Delta$ and is represented by $k_i^{-1}$ at the $i$th node, as shown in Fig. 2.

Another hardware consideration relates to dynamic range scaling. When two's complement and less-than-double precision fixed-point arithmetic are used, summation nodes are allowed to overflow

Manuscript received August 28, 2000; revised March 2001. This work was supported by the Hong Kong Research Grants Council and by the University Research Committee of The University of Hong Kong. This paper was recommended by Associate Editor A. Skodras.

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: nwong@eee.hku.hk; tsng@eee.hku.hk).

Publisher Item Identifier S 1057-7130(01)05222-3.

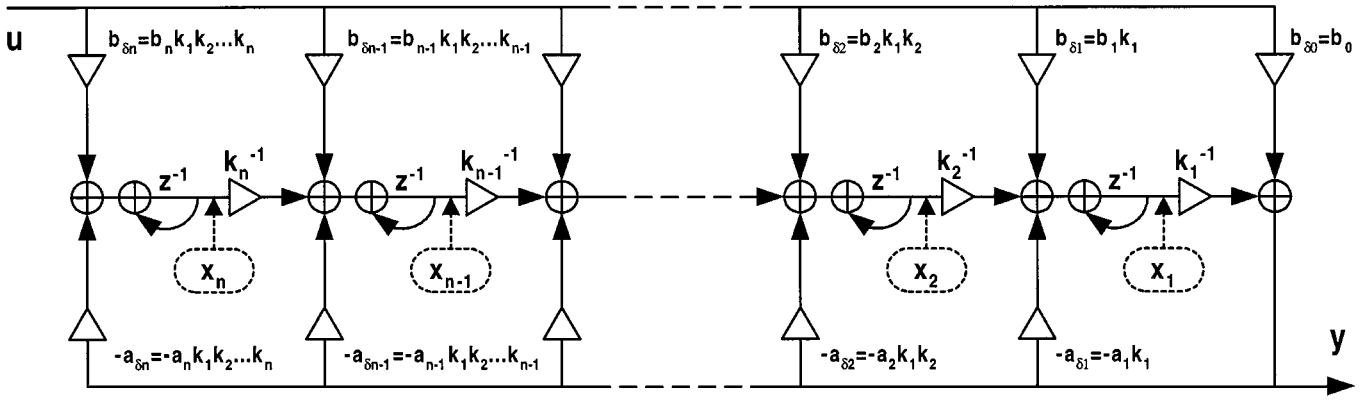


Fig. 2. The generalized $\delta$DFIIt structure.

except before multipliers. These nodes are called branch nodes [22]. Referring to Fig. 2, these nodes correspond to the state variables $x_i$ $(i = 1, 2, \ldots, n)$. To prevent overflow, state variable scaling is required. Further, it is assumed that the output of the initial filter is properly prescaled, say, $L_2$- or $L_\infty$-norm scaled (i.e., $\|H(z)\|_\infty = 1$ or $\|H(z)\|_2 = 1$), to prevent output overflow under a particular input. The $L_p$-norm of a transfer function $H$ is defined as

$$\|H(z)\|_p = \left[\frac{1}{2\pi}\int_0^{2\pi} |H(e^{j\omega})|^p d\omega\right]^{\frac{1}{p}}. \tag{1}$$

Analysis begins with an arbitrary minimal observable and controllable state-space description of an initial filter

$$\begin{cases} z\mathbf{x} = \mathbf{A}_Z\mathbf{x} + \mathbf{B}_Z u \\ y = \mathbf{C}_Z\mathbf{x} + b_0 u \end{cases}. \tag{2}$$

Equation (2) encompasses the most common canonical forms [23]. As stated before, $\mathbf{B}_Z$ and $b_0$ should be properly scaled such that the output will not (or unlikely to) overflow for a particular input $u$. By defining $\rho = (z - 1)$, it follows that

$$\begin{cases} \rho\mathbf{x} = (\mathbf{A}_Z - \mathbf{I})\mathbf{x} + \mathbf{B}_Z u \\ y = \mathbf{C}_Z\mathbf{x} + b_0 u \end{cases}. \tag{3}$$

The definition of $\rho$ here separates the coupling coefficient $\Delta$ from the traditional definition of the $\delta$ operator [2]. It should be stressed that (2) and (3) represent the same system (or transfer function), the only difference being in the use of different operators. For compactness of notation, the system is denoted as sets of the four state-space matrices as follows:

$$\begin{aligned} H(z) &= (\mathbf{A}_Z, \mathbf{B}_Z, \mathbf{C}_Z, \mathbf{D}_Z)_Z = \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z + b_0 \\ &= (\mathbf{A}_\rho, \mathbf{B}_\rho, \mathbf{C}_\rho, \mathbf{D}_\rho)_\rho = \mathbf{C}_\rho(\rho\mathbf{I} - \mathbf{A}_\rho)^{-1}\mathbf{B}_\rho + b_0 \end{aligned} \tag{4}$$

where $\mathbf{D}_Z = \mathbf{D}_\rho = b_0$. The subscripts $Z$ and $\rho$ stand for systems employing the $Z$ and $\rho$ operators, respectively. Obviously

$$\mathbf{A}_\rho = \mathbf{A}_Z - \mathbf{I}, \quad \mathbf{B}_\rho = \mathbf{B}_Z, \quad \mathbf{C}_\rho = \mathbf{C}_Z, \quad \mathbf{D}_\rho = \mathbf{D}_Z. \tag{5}$$

Next, a similarity transform is performed to take those matrices in (3) into observable canonical form [2] such that a $\delta$DFIIt structure as in Fig. 2 can be built from them. Modifying the proof in [24], it is easy to verify that $\mathbf{T}_O$ achieves this task

$$\mathbf{T}_O = \begin{bmatrix} \mathbf{A}_\rho^{n-1}\mathbf{T}_1 & \mathbf{A}_\rho^{n-2}\mathbf{T}_1 & \cdots & \mathbf{A}_\rho\mathbf{T}_1 & \mathbf{T}_1 \end{bmatrix} \tag{6}$$

where

$$\mathbf{T}_1 = \begin{bmatrix} \mathbf{C}_\rho \\ \mathbf{C}_\rho\mathbf{A}_\rho \\ \vdots \\ \mathbf{C}_\rho\mathbf{A}_\rho^{n-1} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \tag{7}$$

This similarity transformation gives another similar set of state-space representation

$$\begin{aligned} &(\hat{\mathbf{A}}_\rho, \hat{\mathbf{B}}_\rho, \hat{\mathbf{C}}_\rho, \hat{\mathbf{D}}_\rho)_\rho \\ &\quad = \left(\mathbf{T}_O^{-1}(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O, \mathbf{T}_O^{-1}\mathbf{B}_Z, \mathbf{C}_Z\mathbf{T}_O, b_0\right)_\rho. \end{aligned} \tag{8}$$

Equation (8) corresponds to Fig. 2 and (11) with all $k_i$s set to 1, i.e., $a_{\delta i} = a_i$ $(i = 1, 2, \ldots, n)$ and $b_{\delta i} = b_i$ $(i = 0, 1, \ldots, n)$. Finally, to incorporate state variable scaling, a similarity transform by a diagonal scaling matrix $\mathbf{T}_S$ is needed where

$$\begin{aligned} \mathbf{T}_S = \operatorname{diag}\big[&k_1^{-1}, (k_1k_2)^{-1}, (k_1k_2k_3)^{-1}, \ldots, \\ &(k_1k_2\ldots k_n)^{-1}\big]. \end{aligned} \tag{9}$$

The function of this matrix is to separately scale the amplitude of each state variable. It generates the set

$$\begin{aligned} &(\hat{\mathbf{A}}'_\rho, \hat{\mathbf{B}}'_\rho, \hat{\mathbf{C}}'_\rho, \hat{\mathbf{D}}'_\rho)_\rho \\ &\quad = \left(\mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\mathbf{T}_S, \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}\mathbf{B}_Z, \right. \\ &\qquad \left. \mathbf{C}_Z\mathbf{T}_O\mathbf{T}_S, b_0\right)_\rho. \end{aligned} \tag{10}$$

Here $\hat{(\ )}$ and $(\ )'$ indicate the quantities after structural change and scaling, respectively. To reflect the actual computational process, (10) is expressed in (11), shown at the bottom of the page, as the "unfolded" observable canonical form corresponding to Fig. 2. As in the figure, we have $b_{\delta 0} = b_0$, $b_{\delta i} = b_i k_1 k_2 \ldots k_i$ and $a_{\delta i} = a_i k_1 k_2 \ldots k_i$ $(i = 1, 2 \ldots, n)$, etc. Thus, we can calculate all filter coefficients once all

$$\begin{cases} \rho\mathbf{x} = \begin{bmatrix} -a_{\delta 1} & 1 & 0 & \\ -a_{\delta 2} & 0 & \ddots & \\ \vdots & \vdots & & 1 \\ -a_{\delta n} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} k_1^{-1} & 0 & \cdots & 0 \\ 0 & k_2^{-1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & k_n^{-1} \end{bmatrix} \mathbf{x} + \left( \begin{bmatrix} b_{\delta 1} \\ b_{\delta 2} \\ \vdots \\ b_{\delta n} \end{bmatrix} + b_{\delta 0} \begin{bmatrix} -a_{\delta 1} \\ -a_{\delta 2} \\ \vdots \\ -a_{\delta n} \end{bmatrix} \right) u \\ \mathbf{y} = \begin{bmatrix} k_1^{-1} & 0 & \cdots & 0 \end{bmatrix} \mathbf{x} + b_{\delta 0} u \end{cases} \tag{11}$$

$k_i$ $(i = 1, 2, \ldots, n)$ are obtained, as will be discussed in later sections. With reference to the first equation in (11), we denote $\hat{\mathbf{A}}'_\rho$ by

$$\hat{\mathbf{A}}'_\rho = \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\mathbf{T}_S = \mathbf{A}_O\mathbf{K}^{-1}$$

where $\mathbf{A}_O$ and $\mathbf{K}^{-1}$ are the two matrices preceding the state variable $\mathbf{x}$ on the right side of the first equation in (11). It then becomes clear that the analysis here differs from traditional ones [7], [8] in that the coupling constant $\Delta$ is separated from the $\delta$ operator and translated into a diagonal *coupling matrix* $\mathbf{K}^{-1} = \mathrm{diag}[k_1^{-1}, k_2^{-1}, \ldots, k_n^{-1}]$. We shall see in later sections that this representation will enable the overall roundoff noise gain as well as filter sensitivity measures/bounds to be minimized simultaneously.

## III. Roundoff Noise Gain

Roundoff quantization noise occurs in coefficient multiplication if less-than-double precision fixed-point arithmetic and rounding are used. Assuming rounding occurs *after multiplication*, expressions for roundoff noise gain generated by coefficient multiplication, namely zero-pole ($b_{\delta i}$s and $a_{\delta i}$s) and coupling coefficients ($k_i^{-1}$s), are derived in this section. The common assumption of additive uncorrelated white noise is made. Noise analysis here does not restrict to the unit noise assumption (i.e., single noise source for each state equation) generally adopted [7], [10] and all possible noise sources are taken into account. In what follows, the noise gain due to the single parameter $q_i$ is denoted by $\mathrm{NG}_{q_i}$ and the total noise gain due to muliple parameters $q_i$ $(i = 1, 2, \ldots, n)$ is denoted by $\mathrm{NG}_{q_i\,(i=1,2,\ldots,n)}$.

### A. Noise Gain Due to Zero-Pole Coefficients

The following will first deal with noise gain due to zero-coefficient multiplication except $b_{\delta 0}$ (the direct transmission). With reference to the first equation in (11) and Fig. 2, the noise transfer function $g_i(z)$ from the noise source after $b_{\delta i}$ $(i = 1, 2, \ldots, n)$ to the output can be found by setting $b_{\delta i} = 1$ and others (including $b_{\delta 0}$) to zero. Thus we have the $i$th element of $\hat{\mathbf{B}}'_\rho$ being 1 and others zeros and obtain the following column vector:

$$\begin{aligned}\mathbf{g}(z) &= [g_1(z) \quad \cdots \quad g_n(z)]^T \\ &= [\hat{\mathbf{C}}'_\rho(\rho\mathbf{I} - \hat{\mathbf{A}}'_\rho)^{-1}]^T \\ &= \mathbf{T}_S^T\mathbf{T}_O^T\left(z\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{C}_Z^T. \end{aligned} \tag{13}$$

Total noise gain arising from these zero coefficients $b_{\delta i}$ $(i = 1, 2, \ldots, n)$ can be evaluated, by noting $\delta[n] = (2\pi)^{-1}\int_0^{2\pi} e^{j\omega n}\,d\omega$ and $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, as

$$\begin{aligned}&\mathrm{NG}_{b_{\delta i}(i=1,2,\ldots,n)} \\ &= \sum_{i=1}^{n}\|g_i(z)\|_2^2 \\ &= \mathrm{tr}\left(\frac{1}{2\pi j}\oint \mathbf{g}(z)\mathbf{g}^T(z^{-1})z^{-1}\,dz\right) \\ &= \mathrm{tr}\left(\frac{1}{2\pi}\int_0^{2\pi}\mathbf{T}_S^T\mathbf{T}_O^T\left(e^{j\omega}\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\right. \\ &\quad\left.\times\,\mathbf{C}_Z^T\mathbf{C}_Z(e^{-j\omega}\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\,d\omega\right) \\ &= \mathrm{tr}\left(\mathbf{T}_S^T\mathbf{T}_O^T\left(\sum_{i=0}^{\infty}\left(\mathbf{A}_Z^T\right)^i\mathbf{C}_Z^T\mathbf{C}_Z\mathbf{A}_Z^i\right)\mathbf{T}_O\mathbf{T}_S\right) \\ &= \mathrm{tr}\left(\mathbf{T}_S^2\mathbf{T}_O^T\mathbf{W}_O\mathbf{T}_O\right)\end{aligned} \tag{14}$$

where the symmetric, positive definite matrix

$$\mathbf{W}_O = \sum_{i=0}^{\infty}\left(\mathbf{A}_Z^T\right)^i\mathbf{C}_Z^T\mathbf{C}_Z\mathbf{A}_Z^i \tag{15}$$

is the observability gramian of (2) and $\mathrm{tr}(\circ)$ denotes the trace of a matrix. Similarly, as can be seen from Fig. 2, the roundoff noise sources after the multiplication of the pole coefficients $a_{\delta i}$ $(i = 1, 2, \ldots, n)$ see the same summation nodes as in the case of $b_{\delta i}$ $(i = 1, 2, \ldots, n)$. Therefore, their transfer functions to the output should be the same as (13) and the noise gain for pole coefficients, denoted by $\mathrm{NG}_P$, is

$$\mathrm{NG}_P = \mathrm{NG}_{a_{\delta i}(i=1,2,\ldots,n)} = \mathrm{NG}_{b_{\delta i}(i=1,2,\ldots,n)}. \tag{16}$$

The remaining noise gain term to be considered is the one due to $b_{\delta 0}$. The transfer function seen by the noise after multiplication by $b_{\delta 0}$ can be found by setting $b_{\delta 0}$ to 1 and $b_{\delta i} = 0$ $(i = 1, 2, \ldots, n)$. It follows that

$$\begin{aligned}g_0(z) &= (\hat{\mathbf{A}}'_\rho, \mathbf{A}_O[1 \quad 0 \quad \cdots \quad 0]^T, \hat{\mathbf{C}}'_\rho, 1)_\rho \\ &= \left(\mathbf{T}_S^{-1}\mathbf{T}_O^{-1}\mathbf{A}_Z\mathbf{T}_O\mathbf{T}_S, \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\right. \\ &\quad\left.\times\,[1 \quad 0 \quad \cdots \quad 0]^T, \mathbf{C}_Z\mathbf{T}_O\mathbf{T}_S, 1\right)_Z \\ &= \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}(\mathbf{A}_Z - \mathbf{I}) \\ &\quad\times\,\mathbf{T}_O[1 \quad 0 \quad \cdots \quad 0]^T + 1\end{aligned} \tag{17}$$

and

$$\begin{aligned}\mathrm{NG}_{b_{\delta 0}} &= \|g_0(z)\|_2^2 \\ &= \frac{1}{2\pi}\int_0^{2\pi} g_0^T(e^{j\omega})g_0(e^{-j\omega})\,d\omega \\ &= 1 + \mathrm{tr}\left(\mathrm{diag}[1, 0, \ldots, 0]\mathbf{T}_O^T\left(\mathbf{A}_Z^T - \mathbf{I}\right)\right. \\ &\quad\left.\times\,\mathbf{W}_O(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\right).\end{aligned} \tag{18}$$

Note that $\mathrm{NG}_{b_{\delta 0}}$ is independent of the choice of $\mathbf{T}_S$. The total noise gain due to all zero coefficients, denoted by $\mathrm{NG}_Z$, is then

$$\mathrm{NG}_Z = \mathrm{NG}_{b_{\delta 0}} + \mathrm{NG}_{b_{\delta i}(i=1,2,\ldots,n)} \tag{19}$$

and the aggregate noise gain due to zero-pole coefficients, denoted by $\mathrm{NG}_{ZP}$, is given by

$$\mathrm{NG}_{ZP} = \mathrm{NG}_Z + \mathrm{NG}_P = \mathrm{NG}_{b_{\delta 0}} + 2\mathrm{NG}_P. \tag{20}$$

In case of strictly causal systems, $b_{\delta 0} = 0$, therefore $\mathrm{NG}_{b_{\delta 0}}$ is absent and should be put to 0 in (19) and (20).

### B. Noise Gain Due to Coupling Coefficients

Now the roundoff noise gain due to multiplication by the coupling coefficients $k_i^{-1}$ $(i = 1, 2, \ldots, n)$ is considered. Again, roundoff quantization noise is assumed to occur after coefficient multiplication. First, since noise source after $k_1^{-1}$ sees the same summation node as that after $b_{\delta 0}$, it follows that

$$\mathrm{NG}_{k_1^{-1}} = \mathrm{NG}_{b_{\delta 0}}. \tag{21}$$

Like $\mathrm{NG}_{b_{\delta 0}}$, $\mathrm{NG}_{k_1^{-1}}$ is independent of the scaling matrix $\mathbf{T}_S$ and is fixed for a given system. Since noise sources after $k_i^{-1}$ $(i = 2, 3, \ldots, n)$ see, respectively, the same summation nodes as those after $b_{\delta i}$ $(i = 1, 2, \ldots, n-1)$, it can be proved by similar procedures as in the previous section that

$$\begin{aligned}&\mathrm{NG}_{k_i^{-1}(i=2,3,\ldots,n)} \\ &= \mathrm{tr}\left(\mathrm{diag}\left[k_1^{-2}, \ldots, (k_1\ldots k_{n-1})^{-2}, 0\right]\mathbf{T}_O^T\mathbf{W}_O\mathbf{T}_O\right) \\ &= \mathrm{tr}\left(\mathrm{diag}\left[0, k_1^{-2}, \ldots, (k_1\ldots k_{n-1})^{-2}\right]\mathbf{T}_O^T\right. \\ &\quad\left.\times\left(\mathbf{A}_Z^T - \mathbf{I}\right)\mathbf{W}_O(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\right).\end{aligned} \tag{22}$$

Subsequently, by noting (18), (21), and (22), the aggregate noise gain due to all coupling coefficients, denoted by $\mathrm{NG}_C$, is

$$\begin{aligned}\mathrm{NG}_c &= \mathrm{NG}_{k_1^{-1}} + \mathrm{NG}_{k_i^{-1}(i=2,3,\ldots,n)} \\ &= 1 + \mathrm{tr}\left((\mathbf{T}_S\mathbf{K})^2\mathbf{T}_O^T\left(\mathbf{A}_Z^T - \mathbf{I}\right)\mathbf{W}_O(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\right).\end{aligned} \tag{23}$$

By noting (14) and (23), noise gain sum is dependent on, and in fact directly proportional to, $\mathbf{T}_S^2$. Therefore, minimum total noise gain is achieved when all the diagonal elements of $\mathbf{T}_S$ are minimized. This is equivalent to maximizing all the diagonal elements of $\mathbf{T}_S^{-1}$, and it will be shown in the next section that their maximum values are constrained by the available dynamic range and can be computed in a simple manner due to the way $\mathbf{T}_S$ is defined.

## IV. Dynamic Range Scaling by $L_2$- and $L_\infty$-Norms

As noted in Section II, the state variables (or branch nodes) in Fig. 2 need to be scaled to prevent overflow. Therefore, we need a set of transfer functions from the input to the state variables. The signal transfer function $f_i(z)$ from the input to the $i$th state variable can be found by setting the output in the second equation of (11) to be that particular state variable. Thus, we have the counterpart of $\mathbf{g}(z)$

$$\begin{aligned}\mathbf{f}(z) &= [f_1(z) \quad \cdots \quad f_n(z)]^T \\ &= (\rho\mathbf{I} - \hat{\mathbf{A}}'_\rho)^{-1}\hat{\mathbf{B}}'_\rho \\ &= \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z. \end{aligned} \tag{24}$$

As stated in [10], [11], the $L_2$-norm is practically the most convenient choice. To fully utilize the available dynamic range of the state variables, $k_i$ $(i = 1, 2, \ldots, n)$ are required to satisfy

$$\begin{aligned}&\frac{1}{2\pi j}\oint \mathbf{f}(z)\mathbf{f}^T(z^{-1})z^{-1}\,dz \\ &= \frac{1}{2\pi}\int_0^{2\pi} \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(e^{j\omega}\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z\mathbf{B}_Z^T \\ &\quad\times \left(e^{-j\omega}\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{T}_O^{-T}\mathbf{T}_S^{-T}\,d\omega \\ &= \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}\mathbf{W}_C\mathbf{T}_O^{-T}\mathbf{T}_S^{-T} \\ &= \begin{bmatrix} k_1^2\alpha_1^2 & \times & \cdots & \times \\ \times & (k_1k_2)^2\alpha_2^2 & & \vdots \\ \vdots & & \ddots & \times \\ \times & \cdots & \times & (k_1\ldots k_n)^2\alpha_n^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \times & \cdots & \times \\ \times & 1 & & \vdots \\ \vdots & & \ddots & \times \\ \times & \cdots & \times & 1 \end{bmatrix} \end{aligned} \tag{25}$$

where "$\times$" denotes "don't care." The symmetric, positive definite matrix

$$\mathbf{W}_C = \sum_{i=0}^{\infty}\mathbf{A}_Z^i\mathbf{B}_Z\mathbf{B}_Z^T\left(\mathbf{A}_Z^T\right)^i \tag{26}$$

is the controllability gramian of (2) and $\alpha_i^2$ stands for the $i$th diagonal element of $\mathbf{T}_O^{-1}\mathbf{W}_C\mathbf{T}_O^{-T}$. If $L_\infty$-norm is to be used, the following condition should be satisfied, namely:

$$\begin{aligned}\|\mathbf{f}(z)\|_\infty &= \mathbf{T}_S^{-1}\left\|\mathbf{T}_O^{-1}(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z\right\|_\infty \\ &= [1 \quad \cdots \quad 1]^T. \end{aligned} \tag{27}$$

Here $\|\mathbf{f}(z)\|_\infty$ denotes the $L_\infty$-norm of each element in the column vector, usually obtained by analytical methods.

Now it is evident that, contrary to the noise gain terms, the state variable amplitudes are inversely proportional to $\mathbf{T}_S^2$ in (25) (or $\mathbf{T}_S$ in (27)). As noted before, the total noise gain is minimized by maximizing the diagonal elements of $\mathbf{T}_S^{-1}$. This is achieved in (25) and (27) by setting the state variable norms ($L_2$- or $L_\infty$-norms) to be 1, i.e., to utilize the dynamic range. The values of $k_i$ $(i = 1, 2, \ldots, n)$ under such conditions can then be determined easily starting from $k_1$ down to $k_n$, and all coefficients for this noise optimal filter can now be computed according to (9)–(11).

It is now apparent that the proposed modified delta operator filter has its total noise gain minimized as its state variable amplitudes are maximized. This is consistent with results in state-space filters using the shift operator [10]. It is also clear from the procedure of determining $k_i$ $(i = 1, 2, \ldots, n)$ that it is unlikely, if not impossible, to utilize the dynamic range fully at all nodes by using only a single coupling coefficient, i.e., $k_i^{-1} = \Delta$ for all $i$s, as in the traditional delta structures [3], [4].

## V. Coupling Coefficients before Integrators

The second form of $\delta^{-1}$ operator in Fig. 1 with coupling coefficient before the integrator will now be investigated. This is the realization adopted in previous work [3]–[6]. Referring to Fig. 2, the coupling coefficients (as well as state variables/branch nodes just before them) are now moved before the integrators. This structural change requires state variables to be scaled by a different set of coupling coefficients, which in turn alter the noise gain terms due to them. Using our convention, but with a slight abuse of notation by putting the operator $\rho$ alongside matrices, this new architecture can be described by

$$H'(z) = (\hat{\mathbf{A}}'_\rho, \rho\hat{\mathbf{B}}'_\rho, \rho^{-1}\hat{\mathbf{C}}'_\rho, \hat{\mathbf{D}}'_\rho)_\rho. \tag{28}$$

Note that for the same initial system (2), $H'(z)$ in (28) is equivalent to $H(z)$ in (4).

The noise gain formulae due to zero-pole coefficients are the same as before. This is because the change of operator realization does not affect the noise transfer functions as seen by these noise sources. Using primes to distinguish formulae for this new $\delta^{-1}$ operator realization (note that usage of primes here differs from the meaning of matrix scaling in (10) and should be clear from the context), we have

$$\mathrm{NG}'_Z = \mathrm{NG}_Z, \quad \mathrm{NG}'_P = \mathrm{NG}_P, \quad \mathrm{NG}'_{ZP} = \mathrm{NG}_{ZP}. \tag{29}$$

It should be stressed that, though the formulae are the same, their quantities are different due to a new set of coupling coefficients, and thus $\mathbf{T}_S$, under the new scaling conditions. Following procedures in Section III, the noise transfer functions as seen by roundoff quantization noise sources after coupling coefficient multiplication can be shown to be

$$\begin{aligned}\mathbf{g}'(z) &= [g'_1(z) \quad \cdots \quad g'_n(z)]^T \\ &= \mathbf{K}^T\mathbf{T}_S^T\mathbf{T}_O^T\left(z\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{C}_Z^T. \end{aligned} \tag{30}$$

Similar to (14), the total noise gain due to coupling coefficients is obtained as

$$\begin{aligned}\mathrm{NG}'_C &= \mathrm{NG}'_{k_i^{-1}(i=1,2,\ldots,n)} = \sum_{i=1}^{n}\|g'_i(z)\|_2^2 \\ &= \mathrm{tr}\left((\mathbf{T}_S\mathbf{K})^2\mathbf{T}_O^T\mathbf{W}_O\mathbf{T}_O\right). \end{aligned} \tag{31}$$

For dynamic range scaling, let the transfer function from the input to the $i$th state variable be $f'_i(z)$, then

$$\begin{aligned}\mathbf{f}'(z) &= \rho\mathbf{f}(z) = [f'_1(z) \quad \cdots \quad f'_n(z)]^T \\ &= (\rho\mathbf{I} - \hat{\mathbf{A}}'_\rho)^{-1}\rho\hat{\mathbf{B}}'_\rho \\ &= \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}[(\mathbf{A}_Z - \mathbf{I})(z\mathbf{I} - \mathbf{A}_Z)^{-1} + \mathbf{I}]\mathbf{B}_Z. \end{aligned} \tag{32}$$

To fully utilize the dynamic range of the state variables, $L_2$-norm scaling requires $k_i$ $(i = 1, 2, \ldots, n)$ to satisfy

$$\begin{aligned}&\mathbf{T}_S^{-1}\mathbf{T}_O^{-1}\Big[(\mathbf{A}_Z - \mathbf{I})\mathbf{W}_C\left(\mathbf{A}_Z^T - \mathbf{I}\right) \\ &\quad + \mathbf{B}_Z\mathbf{B}_Z^T\Big]\mathbf{T}_O^{-T}\mathbf{T}_S^{-T} = \begin{bmatrix} 1 & \times & \cdots & \times \\ \times & 1 & & \vdots \\ \vdots & & \ddots & \times \\ \times & \cdots & \times & 1 \end{bmatrix} \end{aligned} \tag{33}$$

and $L_\infty$-norm scaling requires $k_i$ $(i = 1, 2, \ldots, n)$ to satisfy

$$\|\mathbf{f}'(z)\|_\infty = \mathbf{T}_S^{-1}\left\|\mathbf{T}_O^{-1}[(\mathbf{A}_Z - \mathbf{I})(z\mathbf{I} - \mathbf{A}_Z)^{-1} + \mathbf{I}]\mathbf{B}_Z\right\|_\infty = [1 \quad \cdots \quad 1]^T. \tag{34}$$

From these constraints, the coupling coefficients in $\mathbf{T}_S$ can be determined as before. Again, the total noise gain is minimized when the diagonal elements of $\mathbf{T}_S^{-1}$ are maximized.

## VI. Sensitivity Analysis

In practice, finite word-length prevents the exact implementation of the desired filter. Roundoff or truncation of filter coefficients causes deviation from the ideal transfer function and thus coefficient sensitivity is another important aspect of filter design. Sensitivity measures specific for the $\delta$DFIIt structure are defined and the attainable minimum bounds are derived in this section.

For simplicity, $L_2$-norm scaling of both filter norm and state variables is assumed, consequently $\|H(z)\|_2^2 = 1$. By noting (11), the transfer function is first rewritten in terms of matrices containing the actually implemented filter coefficients.

$$H(z) = \hat{\mathbf{C}}_\rho'(\rho\mathbf{I} - \mathbf{A}_O\mathbf{K}^{-1})^{-1}\hat{\mathbf{B}}_\rho' + b_{\delta 0}. \tag{35}$$

Next, let $\|\circ\|_F$ be the Frobenius norm of an $m \times n$ matrix $\mathbf{F}(e^{j\omega})$, i.e.,

$$\|\mathbf{F}(e^{j\omega})\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|\mathbf{F}_{ij}(e^{j\omega})|^2} = \sqrt{\mathrm{tr}\left(\mathbf{F}_{ij}(e^{j\omega})\mathbf{F}_{ij}^T(e^{-j\omega})\right)} \tag{36}$$

and let the "matrix $L_p$-norm" of $\mathbf{F}(z)$ be

$$\|\mathbf{F}\|_p = \left[\frac{1}{2\pi}\int_0^{2\pi}\|\mathbf{F}(e^{j\omega})\|_F^p d\omega\right]^{\frac{1}{p}}. \tag{37}$$

It is seen that, for a single element matrix, (37) reduces to (1). Since sensitivity involves differentiating transfer function with respect to some parameter matrix, say, $\mathbf{M}$, we define

$$\frac{\partial H}{\partial \mathbf{M}} = \mathbf{S} \tag{38}$$

where the $(i, j)$th element of $\mathbf{S}$ is $s_{ij} = \partial H/\partial m_{ij}$. It should be noted that the sensitivity analysis in the following differs from traditional ones (e.g., [12]) where fully parameterized state-space realizations are considered and every matrix element contributes to the sensitivity measure. In the proposed $\delta$DFIIt filter, since 1s and 0s within matrices $\mathbf{A}_O, \mathbf{K}^{-1}$ and $\hat{\mathbf{C}}_\rho'$ in (11) are implemented exactly, only a subset of coefficients within each matrix (e.g., only the first column of $\mathbf{A}_O$) will affect the transfer function.

### A. Sensitivity Due to Pole Coefficients

By observing (11) and (35), "jitters" in the pole coefficients affect the implementation of $a_{\delta i}$ $(i = 1, 2, \ldots, n)$ in $\mathbf{A}_O$ and $\hat{\mathbf{B}}_\rho'$. With first-order approximation, it can be verified that slight variation of $a_{\delta i}$ affects $H(z)$ as follows:

$$\begin{aligned}\Delta H(z) &= \hat{\mathbf{C}}_\rho(\rho\mathbf{I} - \mathbf{A}_O\mathbf{K}^{-1})^{-1}(\Delta\mathbf{A}_O)\mathbf{K}^{-1}(\rho\mathbf{I} - \mathbf{A}_O\mathbf{K}^{-1})^{-1} \\ &\quad \times \hat{\mathbf{B}}_\rho' + \hat{\mathbf{C}}_\rho(\rho\mathbf{I} - \mathbf{A}_O\mathbf{K}^{-1})^{-1}b_{\delta 0}\Delta a_{\delta i}\mathbf{I}_i \\ &= \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{I}_i\Delta a_{\delta i}\mathbf{I}_1^T\mathbf{K}^{-1}\mathbf{T}_S^{-1}\mathbf{T}_O^{-1} \\ &\quad \times (z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z + \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S b_{\delta 0} \\ &\quad \times \Delta a_{\delta i}\mathbf{I}_i \\ &= \Delta a_{\delta i}(\mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z + b_{\delta 0}) \\ &\quad \times \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{I}_i \\ &= \Delta a_{\delta i}H(z)\mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{I}_i\end{aligned} \tag{39}$$

where $\mathbf{I}_i$ represents the $i$th column of an identity matrix. The third line of the equation is due to the fact that $\mathbf{I}_1^T\mathbf{K}^{-1} = \hat{\mathbf{C}}_\rho' = \mathbf{C}_Z\mathbf{T}_O\mathbf{T}_S$. It follows that

$$\frac{\partial H}{\partial \mathbf{a}_\delta} = \left[\frac{\partial H}{\partial a_{\delta 1}} \quad \cdots \quad \frac{\partial H}{\partial a_{\delta n}}\right]^T = H(z)\mathbf{T}_S^T\mathbf{T}_O^T\left(z\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{C}_Z^T. \tag{40}$$

Defining the pole-coefficient sensitivity measure to be $\|\partial H/\partial \mathbf{a}_\delta\|_1^2$, its bound can be obtained by Cauchy–Schwarz inequality, (13) and (14)

$$\left\|\frac{\partial H}{\partial \mathbf{a}_\delta}\right\|_1^2 = \left\|H(z)\mathbf{T}_S^T\mathbf{T}_O^T\left(z\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{C}_Z^T\right\|_1^2 \le \|H(z)\|_2^2\|\mathbf{g}(z)\|_2^2 = \mathrm{NG}_P = \mathrm{NG}_P'. \tag{41}$$

This simple result is the reason for taking the matrix $L_1$-norm instead of using the otherwise desirable but intractable matrix $L_2$-norm. Equation (41) shows that the pole-coefficient sensitivity, for either $\delta^{-1}$ operator realization, is bounded by its corresponding noise gain. Again, though formulae for $\mathrm{NG}_P$ and $\mathrm{NG}_P'$ are the same, in practice they would give different values due to different $\mathbf{T}_S$s. It can be inferred that the worst case pole-coefficient sensitivity is minimized when its noise gain is also minimized.

### B. Sensitivity Due to Zero Coefficients

First, by ignoring $b_{\delta 0}$, it is easy to show that

$$\frac{\partial H}{\partial \mathbf{b}_\delta} = \left[\frac{\partial H}{\partial b_{\delta 1}} \quad \cdots \quad \frac{\partial H}{\partial b_{\delta n}}\right]^T = \mathbf{T}_S^T\mathbf{T}_O^T\left(z\mathbf{I} - \mathbf{A}_Z^T\right)^{-1}\mathbf{C}_Z^T. \tag{42}$$

Then, by differentiating $H$ with respect to $b_{\delta 0}$ (which appears as standalone term and in $\hat{\mathbf{B}}_\rho'$) and using the following sensitivity measures, highly correlated results are obtained as

$$\left\|\frac{\partial H}{\partial \mathbf{b}_\delta}\right\|_2^2 = \mathrm{NG}_{b_{\delta i}(i=1,2,\ldots,n)}, \quad \left\|\frac{\partial H}{\partial b_{\delta 0}}\right\|_2^2 = \mathrm{NG}_{b_{\delta 0}}. \tag{43}$$

The zero-coefficient sensitivity measure is defined as the sum of these two terms in (43) and, not surprisingly, it is equal to its own noise gain, i.e., $\mathrm{NG}_Z$ or $\mathrm{NG}_Z'$ in (19) and (29). Again, when the zero-coefficient noise gain is minimized, its sensitivity is also minimized.

### C. Sensitivity Due to Coupling Coefficients

Differentiating $H$ with respect to $k_i^{-1}$ and noting that $k_1^{-1}$ appears in both $\hat{\mathbf{C}}_\rho'$ and $\mathbf{K}^{-1}$,

$$\begin{aligned}\frac{\partial H}{\partial k_i^{-1}} &= \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{A}_O\mathbf{I}_i\mathbf{I}_i^T \\ &\quad \times \mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z \\ &\quad + \mathbf{I}_1^T\mathbf{I}_i\mathbf{I}_i^T\mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z \\ &= \left[\mathbf{T}_S^{-1}\mathbf{T}_O^{-1}(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{B}_Z \right. \\ &\quad \left. \times \left(\mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{A}_O + \mathbf{I}_1^T\right)\right]_{(i,i)}\end{aligned} \tag{44}$$

where the subscript $(i, i)$ represents the $i$th diagonal element of a matrix. For convenience, the following row matrix is defined:

$$\begin{aligned}\mathbf{R}(z) &= \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{A}_O + \mathbf{I}_1^T \\ &= \mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}(\mathbf{A}_Z - \mathbf{I})\mathbf{T}_O\mathbf{T}_S\mathbf{K} + \mathbf{C}_Z\mathbf{T}_O\mathbf{T}_S\mathbf{K} \\ &= \rho\mathbf{C}_Z(z\mathbf{I} - \mathbf{A}_Z)^{-1}\mathbf{T}_O\mathbf{T}_S\mathbf{K}.\end{aligned} \tag{45}$$

TABLE I
NUMERICAL EXAMPLES FOR DIFFERENT $\delta^{-1}$ OPERATOR REALIZATIONS

| ***z*-domain Polynomial** $(1+b_1z^{-1}+z^{-2})/(1+a_1z^{-1}+a_2z^{-2})$ | Section A1 | Section A2 | Section A3 | 6th Order Section A1*A2*A3 | Example 6th Order filter from [7] |
|---|---|---|---|---|---|
| ***Denominator* $a_1$, $a_2$** | -1.93504729 | -1.86611453 | -1.80612859 | | |
| | 0.96471582 | 0.88788503 | 0.81824041 | | |
| ***Numerator* $b_1$** | -1.25901348 | -1.87112896 | -1.92379959 | | |
| **1st realization of delta operator: coupling coefficient after integrator** | | | | | |
| **L2 norm scaling** | | | | | |
| $NG_{ZP}$ /dB | 17.7547 | 12.7618 | 12.1382 | 31.7170 | 37.0433 |
| $NG_C$ /dB | 12.0228 | 6.5410 | 4.2474 | 28.4790 | 33.8996 |
| Total gain /dB | 18.7831 | 13.6916 | 12.7922 | 33.4033 | 38.7602 |
| **L-infinity norm scaling** | | | | | |
| $NG_{ZP}$ /dB | 17.7216 | 13.3824 | 13.1159 | 32.2413 | 36.8108 |
| $NG_C$ /dB | 12.0323 | 7.2957 | 5.1693 | 28.9632 | 33.6682 |
| Total gain /dB | 18.7590 | 14.3383 | 13.7622 | 33.9148 | 38.5281 |
| **2nd realization of delta operator: coupling coefficient before integrator** | | | | | |
| **L2 norm scaling** | | | | | |
| $NG'_{ZP}$ /dB | 7.1322 | 4.5732 | 3.4660 | 13.2131 | 19.2134 |
| $NG'_C$ /dB | 14.6183 | 9.1002 | 8.5938 | 25.1044 | 30.4016 |
| Total gain /dB | 15.3313 | 10.4119 | 9.7567 | 25.3766 | 30.7200 |
| **L-infinity norm scaling** | | | | | |
| $NG'_{ZP}$ /dB | 5.0501 | 2.0894 | 0.9538 | 14.9953 | 21.4079 |
| $NG'_C$ /dB | 14.5846 | 9.1966 | 5.8030 | 27.7205 | 32.8381 |
| Total gain /dB | 15.0430 | 9.9691 | 7.0330 | 27.9464 | 33.1398 |

Similar to previous cases, let $\|\partial H/\partial k_i^{-1}\|_1^2$ be the sensitivity measure for that particular coupling coefficient $k_i^{-1}$. Then for the case of coupling coefficients after integrators and noting (24) and (25), the application of Cauchy–Schwarz inequality yields

$$\begin{aligned}\left\|\frac{\partial H}{\partial k_i^{-1}}\right\|_1^2 &= \left\{\frac{1}{2\pi}\int_0^{2\pi}[\mathbf{f}(e^{j\omega})\mathbf{f}^T(e^{-j\omega})]_{(i,i)}^{\frac{1}{2}} \times [\mathbf{R}^T(e^{j\omega})\mathbf{R}(e^{-j\omega})]_{(i,i)}^{\frac{1}{2}}\,d\omega\right\}^2 \\ &\le \left[\frac{1}{2\pi}\int_0^{2\pi}\mathbf{f}(e^{j\omega})\mathbf{f}^T(e^{-j\omega})\,d\omega\right]_{(i,i)} \times \left[\frac{1}{2\pi}\int_0^{2\pi}\mathbf{R}^T(e^{j\omega})\mathbf{R}(e^{-j\omega})\,d\omega\right]_{(i,i)} \\ &= \left[\mathrm{diag}[1,0,\ldots,0] + \mathbf{K}^T\mathbf{T}_S^T\mathbf{T}_O^T\left(\mathbf{A}_Z^T-\mathbf{I}\right)\mathbf{W}_O \times (\mathbf{A}_Z-\mathbf{I})\mathbf{T}_O\mathbf{T}_S\mathbf{K}\right]_{(i,i)}. \end{aligned} \tag{46}$$

The diagonal elements of the first matrix integration on the right of the inequality are all ones due to the scaling constraint in (25). The sensitivity measure due to all coupling coefficients is defined as the sum of all individual sensitivity terms and therefore from (46)

$$\sum_{i=1}^{n}\left\|\frac{\partial H}{\partial k_i^{-1}}\right\|_1^2 \le 1 + \mathrm{tr}\left((\mathbf{T}_S\mathbf{K})^2\mathbf{T}_O^T\left(\mathbf{A}_Z^T-\mathbf{I}\right) \times \mathbf{W}_O(\mathbf{A}_Z-\mathbf{I})\mathbf{T}_O\right) = \mathrm{NG}_C. \tag{47}$$

By similar arguments, for the case of coupling coefficients before integrators and noting (32),

$$\begin{aligned}\left\|\frac{\partial H}{\partial k_i^{-1}}\right\|_1^2 &= \left\{\frac{1}{2\pi}\int_0^{2\pi}[\mathbf{f}'(e^{j\omega})\mathbf{f}'^T(e^{-j\omega})]_{(i,i)}^{\frac{1}{2}} \times [\rho^{-1}\mathbf{R}^T(e^{j\omega})\rho^{-1}\mathbf{R}(e^{-j\omega})]_{(i,i)}^{\frac{1}{2}}d\omega\right\}^2 \\ &\le \left[\mathbf{K}^T\mathbf{T}_S^T\mathbf{T}_O^T\mathbf{W}_O\mathbf{T}_O\mathbf{T}_S\mathbf{K}\right]_{(i,i)} \end{aligned} \tag{48}$$

and similarly

$$\sum_{i=1}^{n}\left\|\frac{\partial H}{\partial k_i^{-1}}\right\|_1^2 \le \mathrm{tr}\left((\mathbf{T}_S\mathbf{K})^2\mathbf{T}_O^T\mathbf{W}_O\mathbf{T}_O\right) = \mathrm{NG}'_C. \tag{49}$$

These results indicate again that when the coupling-coefficient noise gain is minimized, the worst-case bound for the aggregate sensitivity of coupling coefficients is also minimized.

Interestingly, for this generalized $\delta$DFIIt structure, all sensitivity measures/bounds for respective filter coefficients are equal to their corresponding noise gain terms. So the filter achieves its minimum sensitivity measures/bounds simultaneously as its total roundoff noise gain is minimized. This result is consistent with that of optimal shift operator–based state-space filters [13].

## VII. NUMERICAL EXAMPLES

Several second-order sections from [3] and a strictly causal sixth-order narrow-band filter from [7] are taken for noise gain and sensitivity calculation. Prescaling of the filter norms is embedded into the

numerator coefficients and the filter norms are scaled in the same way (either $L_2$- or $L_\infty$-norm scaling) as the state variables. The results are given in Table I.

It can be seen that $L_2$- and $L_\infty$-norm scaling offer similar noise performance. Also, in accordance with Fig. 1, the $\delta^{-1}$ operator realization with coupling coefficient before the integrator generally offers a much lower total noise gain. It has a relatively higher noise gain due to the coupling coefficient multiplication but significantly lower gain due to that of the zero-pole coefficients. On the other hand, the $\delta^{-1}$ operator realization with coupling coefficient after the integrator has its major source of noise from the zero-pole coefficient multiplication. It has less noise contribution from coupling coefficients but the difference is not as significant. Making use of the results from Section VI, the noise gain terms under $L_2$-norm scaling are also equal to the corresponding coefficient sensitivity measures/bounds. In hardware implementation, the structure with coupling coefficients before integrators is recommended because roundoff noise can be more effectively suppressed by using higher precision in the coupling coefficient multiplication. To lower hardware complexity, though at the sacrifice of some roundoff noise gain, the coupling coefficients can be rounded to the nearest powers of 2 such that multiplication can be done by simple bit-shifting.

## VIII. Conclusion

This brief has presented a way to obtain an optimal arbitrary order $\delta$DFIIt filter in terms of its roundoff noise gain and coefficient sensitivity. Procedures for deriving optimal filter coefficients have been given. The incorporation of a coupling matrix into the delta domain state-space description provides more flexibility for utilizing the available dynamic range at filter nodes where scaling is necessary. Expressions for noise gain and coefficient sensitivity are defined and derived. Two structures for realizing an $\delta^{-1}$ operator have been compared through numerical examples. Although the technique presented in this brief is dedicated for the $\delta$DFIIt structure characterization, it can be applied similarly to analyze other canonical forms or topologies expressible in state-space format.

## References


[1] G. C. Goodwin, R. H. Middleton, and H. V. Poor, "High speed digital signal processing and control," *Proc. IEEE*, vol. 80, pp. 240–259, Feb. 1992.

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

[3] J. Kauraniemi, T. I. Laakso, I. Hartimo, and S. J. Ovaska, "Delta operator realizations of direct-form IIR filters," *IEEE Trans. Circuits Syst. II.*, vol. 45, pp. 41–52, Jan. 1998.

[4] ——, "Roundoff noise minimization in a direct form delta operator structure," presented at the Proc. 1996 Int. Conf. Acoust., Speech, and Signal Process., Atlanta, GA, May 1996.

[5] J. Kauraniemi and T. I. Laakso, "Roundoff noise analysis of modified delta operator direct form structures," in *IEEE Int. Symp. Circuits Syst.*, 1997.

[6] N. Wong and T. S. Ng, "Roundoff noise minimization in a modified direct form delta operator IIR structure," *IEEE Trans. Circuits Syst. II.*, vol. 47, pp. 1533–1536, Dec. 2000.

[7] G. Li and M. Gevers, "Roundoff noise minimization using delta operator realizations," *IEEE Trans. Signal Process.*, vol. 41, pp. 629–637, Feb. 1993.

[8] ——, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 803–807, May 1993.

[9] C. T. Mullis and R. A. Roberts, "Synthesis of minimum round-off noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551–562, 1976.

[10] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.

[11] ——, "Dynamic range constraint in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 591–593, Dec. 1975.

[12] L. Thiele, "Design of sensitivity and round-off noise optimal state-space discrete systems," *Int. J. Circuit Theory Applicat.*, vol. 12, pp. 39–46, 1984.

[13] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 884–888, Oct. 1984.

[14] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1114–1122, Sept. 1988.

[15] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta-Sigma Converters*, J. C. Candy and G. C. Temes, Eds. New York: IEEE Press, 1992.

[16] R. Schreier and M. Snelgrove, "Bandpass sigma-delta modulation," *Electron. Lett.*, vol. 25, no. 23, pp. 1560–1561, Nov. 1989.

[17] S. Jantzi, W. Snelgrove, and P. F. Ferguson, "A fourth-order bandpass sigma-delta modulator," *IEEE J. Solid-State Circuits*, vol. 28, pp. 282–291, Mar. 1993.

[18] Y. Botteron and B. Nowrouzian, "An investigation of bandpass sigma-delta A/D converters," in *Proc. 40th Midwest Symp. Circuits Syst.*, 1997, pp. 293–296.

[19] D. A. Johns and D. M. Lewis, "Design and analysis of delta-sigma based IIR filters," *IEEE Trans. Circuits Syst.*, vol. 40, pp. 233–240, Apr. 1993.

[20] S. M. Kershaw and M. B. Sandler, "Digital signal processing on a sigma-delta bitstream," *IEE Colloq. Oversampling Tech. Sigma-Delta Modulation*, pp. 9/1–9/8, 1994.

[21] J. A. S. Angus, "One bit digital filtering," *IEE Colloq. Digital Filters: An Enabling Technology*, pp. 8/1–8/6, 1998.

[22] L. B. Jackson, *Digital Filters and Signal Processing*, 2nd ed. Boston, MA: Kluwer, 1989.

[23] S. Y. Hwang, "Realization of canonical digital networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 27–39, Feb. 1974.

[24] B. C. Kuo, *Digital Control Systems*. New York: Holt, Rinehart and Winston, 1990.