Performance Analysis of a New Structure for Digital Filter Implementation

Gang Li, Member, IEEE, Michel Gevers, Fellow, IEEE, and Youxian Sun

Abstract—It is well known that for a digital filter of order \( n \) the number of nontrivial parameters in the classical optimal state-space realizations is proportional to \( n^2 \). In this paper, a new structure is proposed for digital filter implementation. This structure possesses \( 5n - 1 \) nontrivial parameters and is shown to be equivalent to the discrete-time counterpart of the well-known orthonormal ladder filter in \([13]\). A new property of this ladder filter is revealed. The performance of the proposed structure is analyzed. Expressions for evaluating the sensitivity measure and the roundoff noise gain are derived. A numerical example is presented to illustrate the design procedure. Simulations show that the performance of the proposed structure is almost the same as that of the fully parameterized optimal realizations.

Index Terms—Realizations, roundoff noise, sensitivity, structures.

I. INTRODUCTION

Much attention has been paid to the numerical problems caused by the finite word length (FWL) effects in digital filter implementation for more than two decades. The optimal FWL state-space realization design has been considered as one of the most effective methods (see, e.g., [1]–[6]) to minimize the effects of FWL errors on the performance of digital filters and controllers. It is well known that for a digital filter there exist a number of different representations/parameterizations. Theoretically, they are equivalent since they represent the same system transfer function. However, different representations have different numerical properties and for a given application (measure or criterion) one representation may be better than another. A digital filter transfer function can be implemented with anyone of its state-space realizations. In digital filter implementation, the optimal FWL state-space design is to compute those realizations that minimize the degradation of the filter due to the FWL effects. These effects are classified into two categories: parameter perturbation and roundoff noise. During the last fifteen years, many results have been achieved in terms of finding optimal realizations that optimize a certain FWL effect related cost function such as sensitivity measure and roundoff noise gain; see, [1]–[6].

It has been noted that optimal realizations are usually fully parameterized. In practice, it is desired that the filter have a nice performance as well as a very simple structure \(^1\) that possesses many trivial parameters, which can be implemented exactly and produce no rounding errors. Noting this fact, some modified measures were proposed (see, e.g., [7]–[9]). A lot of effort has been made to achieve sparse optimal or quasi-optimal realizations. For minimal sensitivity realizations, even though the number of nontrivial parameters can be much reduced using the degrees of freedom in the optimal realizations, the amount of nontrivial parameters is still proportional to \( n^2 \), where \( n \) is the order of the filter (see, e.g., [10], [11]). Noting that the degrees of freedom in the optimal roundoff noise realizations are very limited, sophisticated numerical algorithms were developed to obtain the so-called quasi-optimal realizations that have a very sparse structure \([7], [12]\). Besides the numerical difficulty involved in those algorithms, the position of each trivial parameter in the obtained structure is not predictable. In fact, the quasi-optimal sparse structure for one digital filter may be very different from the: for another.

Johns et al. \([13]\) introduced a state-space structure for implementing analog filters. This structure, referred as JSS-structure in the sequel, has a very good sensitivity and dynamic range performance comparable to a cascade of biquads. Filters implemented with such structure are called orthonormal ladder filters and ensure optimal dynamic range scaling with an \( L_2 \)-norm. Another interesting property of the JSS-structure is that it is very sparse. Its discrete-time counterpart, obtained with the bilinear transformation, is called DJSS-structure and can be shown to be inherently \( L_2 \)-scaled. The FWL effects can be greatly reduced when a digital filter is implemented with the DJSS-structure. Unfortunately, the DJSS-structure, unlike the JSS-structure, is fully parameterized.

In this paper, we develop a new structure for digital filter implementation. This structure, referred to as LGS-structure, is very sparse. The basic idea in the development is to factorize the transition matrix of the DJSS-structure into a series of very sparse matrices. With such a factorization, the LGS-structure possesses \( 5n - 1 \) nontrivial parameters instead of \( n^2 + 2n + 1 \). For each output, \( 7n - 3 \) multiplications and \( 6n - 3 \) additions are required, compared with \( n^2 + 2n + 1 \) and \( n^2 + n \), respectively. One of the interesting properties of this structure is that the position of each parameter is always fixed. Performance analysis on this structure is conducted by deriving the expressions for the sensitivity measure and the roundoff noise gain. It should be pointed out that the LGS-structure does not belong to the

\(^1\) Throughout this paper, a structure means a way in which the digital filter is implemented.

\(^2\) Here, by trivial parameters we mean those that are 0 and \( \pm 1 \). Other parameters are, therefore, referred to as nontrivial parameters.
state-space realization set though it is theoretically equivalent to the DISS-structure. The sparseness of this structure makes it a very attractive candidate for digital filter implementation. Simulations show that its performance can be even better than that of the fully parameterized DISS-structure and of the optimal roundoff noise realizations.

An outline of this paper is given as follows. Section II provides some preliminaries on the optimal FWL state-space realization design. The JSS-structure is introduced in Section III. As our first contribution in this paper, a new stability property of this structure is also revealed in this section. Based on the DISS-structure, the LGS-structure is developed in Section IV. Section V is devoted to analyzing the performance of the LGS-structure. In this section, the analytical expressions for computing the sensitivity measure and the roundoff noise gain are derived. A design example is given in Section VI to illustrate the design procedure and to compare the performance of the proposed structure with that of five well-known realizations.

II. PRELIMINARIES

Consider a single-input-single-output time-invariant linear digital filter \( H(z) \) implemented with its state space equations

\[
x(t + 1) = Ax(t) + Bu(t) \\
y(t) = Cx(t) + du(t)
\]

(1)

where \( u(t) \) and \( y(t) \) are the scalar input and output of the filter, respectively.\(^3\) \( R \cong (A, B, C, d) \) with \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times 1}, C \in \mathbb{R}^{1 \times n} \) and \( d \in \mathbb{R} \) is called a realization of \( H(z) \), satisfying

\[
H(z) = d + C(zI - A)^{-1}B.
\]

(2)

Denote \( S_H \cong \{ R : R \text{ satisfies (2)} \} \). \( S_H \) is characterized by

\[
A = T^{-1}A_0T, \quad B = T^{-1}B_0, \quad C = C_0T
\]

(3)

where \( R_0 \cong (A_0, B_0, C_0, d) \in S_H \) and \( T \in \mathbb{R}^{n \times n} \) is any nonsingular matrix.

A. Sensitivity Measure

For an actual implementation, the ideal parameters in \( (A, B, C, d) \) have to be truncated into FWL coefficients. Therefore, the actually implemented transfer function, denoted by \( \tilde{H}(z) \), is different from \( H(z) \). In the traditional FWL analysis, the parameter errors are modeled as zero mean uniformly distributed independent random variables [6], [18], [19]. Keeping this in mind, one can show (see, e.g., [6] and [14]) that the variance of the degradation \( \Delta H(z) \cong \tilde{H}(z) - H(z) \) is proportional to the following sensitivity measure:

\[
M_{L_2} \cong \sum_{k=1}^{N_p} \left\| \frac{\partial H(z)}{\partial p_k} \right\|_2^2
\]

(4)

where \( \{ p_k \} \) are the nontrivial parameters in the realization \( (A, B, C, d) \) and \( \left\| \cdot \right\|_2 \) is the \( L_2 \)-norm defined below with \( q = 2 \).

Definition 1: Let \( f(z) \in C^{n \times m} \) be any complex matrix-valued function of the complex variable \( z \). The \( L_q \)-norm of \( f(z) \) is defined as

\[
\| f \|_q \cong \left( \frac{1}{2\pi} \int_0^{2\pi} \| f(e^{j\omega}) \|_F^q \ d\omega \right)^{1/q}
\]

(5)

where \( \| f(e^{j\omega}) \|_F \) is the Frobenius norm of the matrix \( f(e^{j\omega}) \)

\[
\| f(e^{j\omega}) \|_F \cong \left( \sum_{k=1}^n \sum_{l=1}^m | f_{kl}(e^{j\omega}) |^2 \right)^{1/2}
\]

(6)

with \( \text{tr}(\cdot) \) and \( T \) denote the trace and transpose operations, respectively.

The parameter sensitivity \( \partial H/\partial p_k \) can be found from

\[
S_A(z) \cong \frac{\partial H(z)}{\partial A} = C(z)F^T(z) \quad S_B(z) \cong \frac{\partial H(z)}{\partial B} = G(z) \quad S_C(z) \cong \frac{\partial H(z)}{\partial C^T} = F(z), \quad S_d(z) \cong \frac{\partial H(z)}{\partial d} = 1
\]

(7)

where

\[
F(z) \cong (zI - A)^{-1}B = [f_1(z) \ldots f_n(z)]^T \quad G^T(z) \cong C(zI - A)^{-1} = [g_1(z) \ldots g_n(z)].
\]

(8)

It is easy to see that different realizations have different sensitivity measures. The optimal sensitivity realization design problem is to identify those realizations that minimize \( M_{L_2} \).

Since the minimization of \( M_{L_2} \) was a hard problem, it was initially replaced by the minimization of the following \( L_1/L_2 \) mixed sensitivity measure [4]:

\[
M_{L_1/L_2} \cong \left\| \frac{\partial H}{\partial A} \right\|_1^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C^T} \right\|_2^2 + \left\| \frac{\partial H}{\partial d} \right\|_2^2
\]

(9)

It can be shown [4] that

\[
M_{L_1/L_2} \leq \text{tr}(W_c) \text{tr}(W_o) + \text{tr}(W_c) + \text{tr}(W_o) + 1 \cong \bar{M}_{L_1/L_2}
\]

(10)

and

\[
\min_{R \in S_H} M_{L_1/L_2} \cong \min_{R \in S_H} \bar{M}_{L_1/L_2}
\]

(11)

where \( W_c, W_o \) are the controllability and observability gramians of \( (A, B, C, d) \), respectively, which are the solutions of the Lyapunov equations

\[
W_c = AW_cA^T + BB^T, \quad W_o = A^TW_oA + C^TC.
\]

(12)

The optimal realizations, denoted by \( R_{L_1/L_2} \), are characterized by the following identity:

\[
W_c = W_o.
\]

(13)
The use of the \( L_1 / L_2 \) mixed sensitivity measure leads to an easy treatment of the optimal realization problem. The pure \( L_2 \)-based sensitivity measure is more logical since it has a clear physical interpretation [6], [8], [14].

Returning to the measure \( M_{L_2} \), for fully parameterized realizations one has

\[
M_{L_2} = \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C_T} \right\|_2^2 + \left\| \frac{\partial H}{\partial d} \right\|_2^2.
\]  

(14)

The problem of minimizing \( M_{L_2} \) within the fully parameterized realization set, denoted with \( S_H^f \), has been solved; see, e.g., [6], [14], and [15]. The corresponding realizations are denoted by \( R_{L_2} \).

For sparse realizations, it can be shown [8] that \( M_{L_2} \) can be evaluated with

\[
M_{L_2} = \sum_{k=1}^{n} \sum_{k=1}^{n} \phi_{lk}(C, 0) R_{lk}(C, 0) R_{lk}(C, 0)^\top
\]

\[
+ \sum_{k=1}^{n} \phi_k W_o(k, k) + \psi_k W_c(k, k) + \nu
\]

(15)

where \( W_o, W_c \) are defined in (12), \( R_{lk} \) is the solution of the following Lyapunov equation:

\[
R_{lk} - \begin{pmatrix} A & e_k e_k^\top \\ 0 & A \end{pmatrix} R_{lk} \begin{pmatrix} A & e_k e_k^\top \\ 0 & A \end{pmatrix}^\top = \begin{pmatrix} 0 \\ B \end{pmatrix} \begin{pmatrix} 0 \\ B \end{pmatrix}^\top, \quad \forall(l, k)
\]

(16)

with \( e_i \) denoting the \( i \)th elementary (column) vector, and

\[
\phi_{lk} = \begin{cases} 0, & \text{if } A(l, k) \text{ is trivial} \\
1, & \text{if } A(l, k) \text{ is nontrivial} 
\end{cases}
\]

\[
\varphi_k = \begin{cases} 0, & \text{if } B(k) \text{ is trivial} \\
1, & \text{if } B(k) \text{ is nontrivial} 
\end{cases}
\]

\[
\psi_k = \begin{cases} 0, & \text{if } C(k) \text{ is trivial} \\
1, & \text{if } C(k) \text{ is nontrivial} 
\end{cases}
\]

\[
\nu = \begin{cases} 0, & \text{if } d \text{ is trivial} \\
1, & \text{if } d \text{ is nontrivial} 
\end{cases}
\]

(17)

Though for a given realization one can compute this sensitivity measure, how to minimize \( M_{L_2} \) within the realization space of sparse realizations, denoted with \( S_H^f \), is very difficult, and as far as we know this is still an open problem.

**B. Roundoff Noise Gain**

Another FWL effect is due to the rounding of the states in (1). Consider the situation where the rounding operation is performed after a multiplication (RAM). It can be shown [1], [2], [6] that the variance of the output error due to the roundoff is given by

\[
\sigma_{\Delta y}^2 = \sigma_0^2 + m_n + 1
\]

(18)

where \( \sigma_0^2 \) is a constant, determined by the word length used for representing the states, \( Q = \text{diag}(m_1, m_2, \ldots, m_n) \) with \( m_k \) the number of nontrivial parameters in the row vector \( k \)th row of \( (A, B) \) for \( k = 1, \ldots, n \), and \( m_n + 1 \) is the number of nontrivial parameters in the row vector \( (C, d) \).

In the actual implementation, all states in (1) have to be maintained in the same dynamic range, which is achieved with the following \( \ell_2 \)-scaling [1], [2]:

\[
W_c(i, i) = 1, \quad \forall i.
\]

(19)

Equation (19) defines a subset, denoted \( S_H^f \), of the realization set \( S_H \) defined in (3). The optimal roundoff noise realization problem is to find those realizations in \( S_H^f \) that minimize \( \sigma_{\Delta y}^2 \) or, equivalently, the roundoff noise gain defined as

\[
G = \frac{\sigma_{\Delta y}^2}{\sigma_0^2} = \text{tr}(W_o Q) + m_n + 1.
\]

(20)

If the realization is fully parameterized, that is \( m_k = n + 1 \), \( \forall k \), one has

\[
G = \text{tr}(W_o) + 1)(n + 1).
\]

(21)

The following optimal realization problem:

\[
\min_{R \in S_H^f} G = \text{tr}(W_o) + 1)(n + 1)
\]

(22)

was solved in [1] and [2] independently. The corresponding realizations are denoted by \( R_G \).

The optimal realization problem defined over the set of sparse realizations

\[
\min_{R \in S_H^f} G
\]

(23)

just as the optimal sensitivity realization problem, seems very difficult and is still an open problem. In [7], a numerical algorithm was proposed to solve a suboptimal problem.

It is interesting to note that for all \( \ell_2 \)-scaled realizations in \( S_H^f \), one has

\[
\tilde{M}_{L_1/L_2} = G
\]

(24)

where \( \tilde{M}_{L_1/L_2} \) and \( G \) are defined in (10) and (21), respectively.

Equation (24) implies that any minimal roundoff realization \( R_G \) is also a minimum sensitivity realization in this subset. Therefore, \( R_G \) should yield a very good performance in terms of reducing FWL effects. \( R_G \), however, possesses \( n^2 + 2n + 1 \) nontrivial parameters, which limits its applications in digital filter implementation. In what follows, based on a well-known continuous-time realization, we will develop a new structure which is sparse and yields a performance comparable to the fully parameterized \( R_G \).

**III. The JSS Structure**

In this section, we introduce the JSS-structure used in [13] for analog filter implementation. A new stability property of this structure is revealed. Based on the JSS-structure, a new structure will be developed for digital filter implementation in the next section.

It is well known that the continuous-time counterpart of \( H(z) \), denoted with \( F(s) \), can be obtained with the bilinear
transformation \( s = (z - 1/z + 1) \), or \( z = (1 + s/1 - s) \), such that
\[
F(s) = H(z)_{z = \frac{1+s}{1-s}}. \quad (25)
\]
Clearly, \( H(z) \) has all its poles in \( |z| < 1 \) if and only if \( F(s) \) has all its poles in \( \text{Re}(s) < 0 \), where \( \text{Re}(z) \) denotes the real part of any complex number \( z \).

Let \( (A, B, C, d) \in S_H \) and let \( F(s) \) be defined from \( H(z) \) as in (25). It can be shown that there exists a realization of \( F(s) \) given by
\[
\Phi = (I + A)^{-1}(A - I), \quad K = \sqrt{2}(I + A)^{-1}B
\]
\[
L = \sqrt{2}C(I + A)^{-1}, \quad D = d - C(I + A)^{-1}B \quad (26)
\]
such that
\[
F(s) = D + L(sI - \Phi)^{-1}K. \quad (27)
\]
The continuous-time gramian pair \( (P, Q) \) of \( \Phi, K, L, D \) satisfies the following equations:
\[
\Phi P + P \Phi^T = -KK^T, \quad \Phi^T Q + Q \Phi = -L^T L \quad (28)
\]
It is interesting to note that
\[
P = W_o, \quad Q = W_o \quad (29)
\]
where \( (W_o, W_o) \), as defined before, is the gramian pair of \( (A, B, C, d) \).

Like \( H(z) \), \( F(s) \) also has an infinite number of realizations. Consider the following continuous-time realization \((\Phi_{in}, K_{in}, L_{in}, D)\) of \( F(s) \):
\[
\Phi_{in} = \begin{pmatrix}
0 & \alpha_1 & 0 & 0 & \cdots & 0 & 0 \\
-\alpha_1 & 0 & \alpha_2 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 0 & \alpha_{n-1} \\
0 & 0 & 0 & 0 & \cdots & -\alpha_{n-1} & -\alpha_n
\end{pmatrix} \quad (30)
\]
\[
K_{in} = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 \\
\sqrt{2\alpha_n}
\end{pmatrix}
\]
where \( \alpha_k \) are real \( \forall k \) and \( L_{in} \) has no special structure.

This state-space structure, proposed by Johns et al.,\(^4\) for implementing a given continuous-time transfer function \( F(s) \), was shown to have a very good behavior against FWL effects [13] and leads to the so-called orthonormal ladder filters. We note that \( P = I \) if \( \Phi, K \) is replaced with \( \Phi_{in}, K_{in} \) in (28), and with (29) one can see that the JSS-structure is inherently \( L_2 \)-scaled [13] and its discrete-time counterpart \((A_{in}, B_{in}, C_{in}, d)\), called \textit{DSS-structure}, is automatically \( L_2 \)-scaled. For other nice properties of the JSS-structure, we refer to [13].

Before going to the issue of computing the JSS-structure, let us reveal a new property of this structure.

A. A New Property of the JSS-Structure

Let \( \lambda \) be any eigenvalue of \( \Phi_{in} \). We show that \( \Phi_{in} \) is (strictly) stable, that is the real part of \( \lambda \) is negative, iff \( \alpha_k \neq 0, k = 1, 2, \ldots, n - 1 \) with \( \alpha_n > 0 \). To prove this claim, one needs the following lemma.

Lemma 1: Let \( M \in \mathbb{R}^{n \times n} \) and \( M = M_s + M_{sh} \) with \( M_s = M_s^T \) and \( M_{sh} = -M_{sh}^T \). Denote \( \{\lambda_k\} = \lambda(M) \) and as the eigenvalue set of \( M \). Then
\[
\text{Re}(\lambda_k) = -\frac{x_k^T M_s x_k}{x_k^T M_{sh} x_k} \quad \forall k \quad (31)
\]
where \( x_k \) is an eigenvector corresponding to \( \lambda_k \) and \( M \) denotes the transpose-conjugate operator.

\textbf{Proof:} Let \( x_k \) be an eigenvector of \( \lambda_k \), that is \( M x_k = \lambda_k x_k \). Clearly, \( x_k^T M_s x_k = \lambda_k x_k^T x_k \). With the decomposition \( M = M_s + M_{sh} \), one has \( x_k^T M_s x_k = \text{Re}(\lambda_k) x_k^T x_k \), which leads to (31).

Theorem 1: Let \( \Phi_{in} \) be given by (30). Denote \( \{\lambda_k\} = \lambda(\Phi_{in}) \), then \( \Phi_{in} \) is strictly stable iff \( \alpha_k \neq 0 \) for \( k = 1, 2, \ldots, n - 1 \) and \( \alpha_n > 0 \).

\textbf{Proof:} First of all, let \( M = \Phi_{in} \) in Lemma 1. Noting all the elements of the symmetric matrix \( M_s \) are zero except the last diagonal element, which is \( -\alpha_n \), it follows from (31) that
\[
\text{Re}(\lambda) = -\frac{\alpha_n |x_n|^2}{x_k^T x_k} \quad (32)
\]
where \( x = (x_1, x_2, \ldots, x_n)^T \) is an eigenvector corresponding to \( \lambda \in \lambda(\Phi_{in}) \). Clearly, the sufficiency part of the theorem can be proved if one can show that \( \alpha_k \neq 0 \) for \( k = 1, 2, \ldots, n - 1 \) and \( \alpha_n > 0 \) implies \( x_n \neq 0 \). In fact, it follows from \( \Phi_{in} x = \lambda x \) that
\[
\alpha_1 x_2 + \alpha_2 x_3 + \cdots + \alpha_{n-1} x_n + \alpha_n x_n = \lambda x_n \quad (33)
\]
It is easy to see that with all \( \alpha_k \neq 0 \), \( x_n = 0 \) means \( x = 0 \), which contradicts the assumption that \( x \) is an eigenvector corresponding to \( \lambda \in \lambda(\Phi_{in}) \).

Now, assume that \( \Phi_{in} \) is strictly stable. The necessity part of the theorem is proved if one can show that no \( \alpha_k \) is zero and that \( \alpha_n > 0 \).

It is easy to see from (32) that \( \alpha_n \) is necessarily positive and from (30) that \( \alpha_1 \neq 0 \) if \( \Phi_{in} \) is stable. Assume \( \alpha_k = 0 \) for all \( k = 2, \ldots, n - 1 \). Then \( \Phi_{in} \in \mathbb{R}^{(n-1) \times (n-1)} \) and \( \Phi_k \in \mathbb{R}^{(n-k-1) \times (n-k-1)} \) and \( k \) is any number between \( 2 \) and \( n - 1 \). Noting the fact that \( \Phi_1 \) is a skew-symmetric matrix, it follows from (31) that all eigenvalues of \( \Phi_1 \) are imaginary. This means that \( \Phi_{in} \) is marginally stable, which is against the assumption.

B. Computing the JSS-Structure

Now, let us consider the problem of computing the JSS-structure for a given \( H(z) \). First of all, suppose \( F(s) = N(s)/D(s) \)
is obtained by (25). Let the denominator be decomposed as
\[ D(s) = E(s) + O(s), \]
where \( E(s) \) and \( O(s) \) are the even- and odd-power terms of \( D(s) \), respectively. Define \( Z(s) = O(s)/E(s) \), if the order of \( D(s) \) is odd, or \( Z(s) = E(s)/O(s) \), if the order of \( D(s) \) is even. Then we have the following continued fraction form:
\[
Z(s) = r_n s + \frac{1}{r_{n-1}s + \frac{1}{r_{n-2}s + \ldots + \frac{1}{r_1s}}}.
\]
(34)

The Hurwitz test (see, e.g., [16] and [17]) says that \( D(s) \) is (strictly) stable iff \( r_k > 0, \forall k \). It was shown in [13] that (30) can be obtained by diagonally transforming the realization of an orthonormal ladder filter derived from a singly terminated LC ladder filter with a unit valued resistor, where the states correspond to the capacitor voltages and inductor currents, and \( r_k \) defined in (34) is either the capacitor or inductor value for all \( k \). The corresponding diagonal similarity transformation is
\[
T = \text{Diag}(\sqrt{1/2\gamma_1}, \sqrt{1/2\gamma_2}, \ldots, \sqrt{1/2\gamma_n}),
\]
which leads to
\[
\begin{align*}
\alpha_k &= \sqrt{\frac{1}{r_k\gamma_{k+1}}}, & k &= 1, 2, \ldots, n-1 \\
\alpha_n &= \frac{1}{r_n}.
\end{align*}
\]
(35)

\( \Phi_{in} \) and \( K_{in} \) can directly be defined from the \( \alpha_k \) obtained above.

Let \( (\Phi_c, K_c, L_c, d) \) be any realization of \( F(s) \), say the controllable realization. With \( \Phi_{in} \) and \( K_{in} \), one can find the similarity transformation \( T_{in} \) that transforms \( (\Phi_c, K_c, L_c, d) \) into the JSS-structure and hence also \( L_{in} \) with \( L_{in} = L_cT_{in} \).

IV. The LGS Structure

It follows from (26) that the DJSS-structure is given by
\[
\begin{align*}
A_{in} &= (I + \Phi_{in})(I - \Phi_{in})^{-1} \\
B_{in} &= \frac{\sqrt{2}}{2} (I + A_{in}) = K_{in} \\
C_{in} &= \frac{\sqrt{2}}{2} L_{in}(I + A_{in}) \\
d &= D + C_{in}(I + A_{in})^{-3}B_{in}.
\end{align*}
\]
(36)

It can be shown that the expression for computing the sensitivity measure for analog filters has the same form as that given in (10) [20]. Let \( R_k^* \) be a continuous-time realization and \( R_k \) the corresponding digital realization, for \( k = 1, 2 \). Equation (29) implies that the JSS-structure and the DJSS-structure have exactly the same behavior in the sense that if \( R_k^* \) has a smaller sensitivity measure than \( R_k^* \), then the same holds for \( R_1 \) vis-à-vis \( R_2 \). Noting the fact that the JSS-structure has a very small sensitivity measure [13], it follows from (24) that the the DJSS-structure has a very small sensitivity as well as a very small roundoff noise gain.

One would therefore suggest using this DJSS-structure for digital filter implementation. However, it has been noted that, unlike \( \Phi_{in} \), \( A_{in} \) obtained with (36) is fully parameterized due to the matrix inversion involved in (36). A direct implementation of \( A_{in} \) leads to \( n^2 \) multiplications for computing \( A_{in}x(t) \) in (1).
Note that $x^{(N)}(t) = A_{in}x(t)$. Simple calculations show that computing $x^{(N)}(t)$ with (43) requires only $5n - 4$ multiplications and $4n - 3$ additions, rather than $n^2$ and $n(n - 1)$, respectively, as required in the DJISS-structure. This is a significant reduction of computational complexity.

For convenience, (43) is referred to as the LGS-structure. Before turning to the next section, we point out that the LGS-structure, though equivalent to the state-space DJISS-structure, does not belong to the state-space realization set. Equation (43) yields a different class of implementation structures.

V. PERFORMANCE ANALYSIS OF THE LGS STRUCTURE

In this section, we analyze the performance of the structure proposed in the previous section in terms of sensitivity measure and roundoff noise.

A. Sensitivity Analysis

First of all, it follows from (41) and (42) that

$$A^{(2k)} A^{(2k-1)} = \begin{pmatrix} 1 & \gamma_k & 0 \\ 0 & -\alpha_k & 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$A^{(N)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for $k = 1, \ldots, n$, and $A^{(N)} = I + \Phi_0$. Therefore, the LGS-structure (43) is parameterized with $B_{in}, C_{in}, d$, and three parameter sets $\{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}$. Noting that $B_{in}$ and $C_{in}$ are generally fully parameterized with nontrivial parameters, one can see that the LGS-structure has $5n - 1$ parameters to be implemented with $\beta_0 = \alpha_0$. The transfer function sensitivities $\partial H/\partial B_{in}, \partial H/\partial C_{in}$, and $\partial H/\partial d$ can be obtained with (7)-(8).

Now, let us consider $\partial H/\partial p, \forall p \in S_p$, where $S_p$ denotes the parameter set containing the parameters that define $A_{in}$: $\alpha_k$ for $k = 1, \ldots, n, \beta_k$ for $k = 2, \ldots, n - 1$, and $\gamma_k$ for $k = 1, \ldots, n - 1$. Denote

$$A_{in} = A^{(N)} \prod_{i=N-1}^{N-1} A^{(i)} \prod_{i=1}^{2(n-1)} A^{(i)} \prod_{i=1}^{2(n-1)} A^{(i)}$$

It follows from $H(z) = d + C_{in}(z I A_{in})^{-1} B_{in}$ that

$$\frac{\partial H}{\partial p} = C_{in}(z I - A_{in})^{-1} \frac{\partial A_{in}}{\partial p}(z I - A_{in})^{-1} B_{in}$$

where

$$\frac{\partial A_{in}}{\partial p} = \frac{\partial A^{(N)}}{\partial p} A_{in} A_{in} + A^{(N)} \frac{\partial A_{in}}{\partial p} A_{in} + A^{(N)} A_{in} \frac{\partial A_{in}}{\partial p}.$$  

(46)

Noting $\beta_0 = \alpha_0$, one can show (47), as shown at the bottom of the page. Similarly, we have (48) and (49), shown at the bottom of the next page. Noting that $B_{in}$ and $C_{in}$ are generally fully parameterized, and assuming that $d$ is nontrivial, the $L_2$-sensitivity measure for the proposed structure is

$$M_{LGS} = \sum_{p \in s_p} \left( \frac{\partial H}{\partial p} \right)_{1,2}^2 + \left( \frac{\partial H}{\partial B_{in}} \right)_{1,2}^2 + \left( \frac{\partial H}{\partial C_{in}} \right)_{1,2}^2 + \left( \frac{\partial H}{\partial d} \right)_{1,2}^2$$

$$= \sum_{p \in s_p} \left( \frac{\partial H}{\partial p} \right)_{1,2}^2 + \text{tr} \left( W_{in}^T \right) + (n + 1)$$

(50)

where $W_{in}^T$ is the observability gramian of $(A_{in}, B_{in}, C_{in})$ for which $W_{in}^T = I$.

Noting the following equality:

$$M_{LGS} = (I - M_1) (M_1 - M_2)$$

Note from (45) that

$$\frac{\partial H}{\partial p} = C_{in}(I - 0) \left( \begin{array}{c} z I_{2n} - (A_{in} \frac{\partial A_{in}}{\partial p}) \end{array} \right)^{-1} \left( \begin{array}{c} 0 \\ I_n \end{array} \right) B_{in}$$

$$= C_{in}(z I_{2n} - A_{in})^{-1} B_{in}$$

(51)

where $I_{2n}$ denotes the unit matrix of order $2n$ and $0$s are zero matrices of proper dimension. It is easy to show that

$$\left( \frac{\partial H}{\partial p} \right)_{1,2}^2 = C_{in} W_c C_{in}^T = B_{in}^T W_c B_{in}$$

(52)

where $(W_c, W_c)$ is the gramian pair of the realization $(A_{in}, B_{in}, C_{in})$ defined in (31).
Therefore, \( \sum_{p \in S_p} \| (\partial H/\partial \rho) \|_{L_2}^2 \) and hence \( M_{LGS} \) can be computed using (52) with (46)–(50).

### B. Roundoff Noise Analysis

In an actual implementation, all intermediate variables should be maintained to a certain wordlength, say \( B_s \) bits. Therefore, the product of any nontrivial parameter and an intermediate variable has to be rounded to \( B_s \) bits. So, the actual model of the LGS-structure (43) with all intermediate variables rounded is

\[
x^*(t + 1) = Q \left[ A^{(N)} Q \left[ \cdots \left[ A^{(2)} Q \left[ A^{(1)} x^*(t) \right] \cdots \right] \right] + Q[B_{in} u(t)] \right] y^*(t) = Q[C_{in} x^*(t)] + Q[du(t)]
\]

where \( Q[Mv] \) is the quantizer that rounds all products occurring in the multiplication \( Mv \) into \( B_s \) bits.

Denoting

\[
Z_1(t) \triangleq Q \left[ A^{(N-1)} \cdots A^{(2)} x^*(t) \cdots \right] \\
Z_2(t) \triangleq Q \left[ A^{(N-1)} \cdots A^{(N-(n-1))} Z_1(t) \cdots \right]
\]

it then follows from (53) that

\[
x^*(t + 1) = A^{(N)} Z_2(t) + B_{in} u(t) + e_3(t) \\
y^*(t) = C_{in} x^*(t) + du(t) + \eta(t)
\]

where \( e_3(t) \) and \( \eta(t) \) are the roundoff noises in \( Q[A^{(N)} Z_2(t)] + Q[B_{in} u(t)] \) and \( Q[C_{in} x^*(t)] + Q[du(t)] \). Their covariance is given by

\[
R_{e3} = E[e_3(t)e_3^T(t)] = \text{diag}(2, 3, \ldots, n) \sigma^2_e \triangleq D_3 \sigma^2_0 \\
R_{\eta} = E[\eta(t)\eta^T(t)] = (n - 1) \sigma^2_0
\]

with \( \sigma^2_0 \) a constant depending on \( B_s \).

It is easy to see that

\[
Z_2(t) = A_2 Z_1(t) + e_2(t)
\]

where \( e_2(t) \) is the quantization error of \( Z_2(t) \) given by (54). Noting the structure specified in (44), one can show that

\[
R_{e2} = E[e_2(t)e_2^T(t)] = \text{diag}(1, 1, \ldots, 1, 0) \sigma^2_e \triangleq D_2 \sigma^2_0.
\]

Similarly

\[
Z_1(t) = A_1 x^*(t) + e_1(t)
\]

where \( e_1(t) \) is the roundoff noise, satisfying

\[
R_{e1} = E[e_1(t)e_1^T(t)] = \text{diag}(0, 1 + \gamma_1^2, 1 + \gamma_2^2, \ldots, 1 + \gamma_{n-1}^2) \sigma^2_0 \triangleq D_1 \sigma^2_0.
\]

Combining (55), (57) and (59), one has

\[
x^*(t + 1) = A_in x^*(t) + B_{in} u(t) + e(t) \\
y^*(t) = C_{in} x^*(t) + du(t) + \eta(t)
\]

where

\[
e(t) = A^{(N)} A_2 e_1(t) + A^{(N)} e_2(t) + e_3(t).
\]

Denoting \( \Delta y(t) = y^*(t) - y(t) \) and \( E(t) = x^*(t) - x(t) \), it follows from (43) and (61) that

\[
E(t + 1) = A_{in} E(t) + e(t) \\
\Delta y(t) = C_{in} E(t) + \eta(t).
\]

Keeping in mind the assumption that all roundoff noises are independent and white, it turns out from (63) that

\[
\sigma^2_y \triangleq E[(\Delta y(t))^2] = C_{in} P C_{in}^T + R_\eta
\]

where

\[
P = A_{in} P A_{in}^T + R_e \\
R_e \triangleq E[e(t)e(t)^T] = A^{(N)} \left[ A_2 R_{e1} A_2^T + R_{e2} \right] \left( A^{(N)} \right)^T + R_{e3} \\
= \left( A^{(N)} \right) \left[ A_2 D_1 A_2^T + D_2 \right] \left( A^{(N)} \right)^T + D_3 \sigma^2_0 \triangleq R_0 \sigma^2_0.
\]

It can then be shown that

\[
\sigma^2_y = \text{tr} \left( W^{(in)}_e R_e + R_\eta \right)
\]

\[
\frac{\partial A^{(N)}}{\partial \beta_k} = 0, \quad \frac{\partial A_1}{\partial \beta_k} = 0 \]

\[
\frac{\partial A_2}{\partial \beta_k} = \left\{ \begin{array}{ll}
- \prod_{i=2}^{(N-1)} A^{(i)} e_k e_{i+1}^T & 2 \leq k \leq n - 2 \\
- \prod_{i=2(n-k)+2}^{(N-1)} A^{(i)} e_{n-i} e_n^T & k = n - 1
\end{array} \right.
\]

\[
\frac{\partial A^{(N)}}{\partial \gamma_k} = 0, \quad \frac{\partial A_1}{\partial \gamma_k} = 0 \]

\[
\frac{\partial A_2}{\partial \gamma_k} = \left\{ \begin{array}{ll}
\prod_{i=2}^{(n-1)} A^{(i)} e_k e_{i+1}^T \prod_{i=1}^{2k-1} A^{(i)} & 1 \leq k \leq n - 2 \\
e_n e_n^T \prod_{i=1}^{2(n-1)-1} A^{(i)} & k = n - 1
\end{array} \right.
\]
The roundoff noise gain for the proposed structure (43) is therefore

\[ G = \sigma_{x}^{2} \sigma_{\delta}^{-2} = \text{tr} \left( W^{(\text{in})} R_{0} \right) + (n + 1) \]  

(67)

where \( R_{0} \) is defined in (65).

VI. A NUMERICAL EXAMPLE AND SIMULATION RESULTS

In this section, we present a design example to illustrate the performance of the proposed structure and compare it with five other structures: the controllable realization \( R_{c} \), \( R_{\text{DISS}} \), and fully parameterized optimal realizations \( R_{L_{1}/L_{2}} \), \( R_{L_{2}} \), and \( R_{Q} \).

Example: This is a sixth-order narrow band low-pass filter used in [6]. The normalized passband frequency \( f_{p} \) is 0.03125, the stopband frequency \( f_{s} \) is 0.0390625, and the passband ripple is 1 dB. The attenuation in the stopband is greater than 46.68 dB. The corresponding magnitude response is depicted in Fig. 1 with solid line.

Table I shows the statistics on the sensitivity measure \( M_{L_{2}} \), the roundoff noise gain \( G \), and the number of nontrivial parameters \( N_{p} \) involved in implementation for the six structures.

In this example, one can see that \( R_{e} \) is very sensitive to the parameter errors. This is due to the very narrow bandwidth of the filter. The other five structures have a much smaller sensitivity measure. The proposed \( R_{L_{2}} \) yields a slightly larger \( M_{L_{2}} \) value than \( R_{C} \) and \( R_{\text{DISS}} \). In fact, all five structures yield almost the same frequency response for a given parameter perturbation. Let us truncate all nontrivial parameters of a structure into \( B_{e} \) bits in their fractional part. The magnitude responses of \( R_{G} \) and \( R_{L_{2}} \), both truncated with \( B_{e} = 10 \), are depicted in Fig. 1 with the dotted and dashdot line, respectively. It clearly shows that both structures yield almost the same response, which is very close to the desired one, especially in the passband. To achieve the same response, \( R_{e} \) requires at least \( B_{e} = 22 \) bits. In terms of roundoff noise gain, \( R_{L_{2}} \) is quite close to \( R_{G} \) and better than \( R_{\text{DISS}} \).

Comment 6.1: The relationship between the \( G \) values of \( R_{G} \) and \( R_{L_{2}} \) depends on examples. In fact, \( R_{L_{2}} \) may yield a smaller roundoff noise gain than \( R_{G} \). This is due to the fact that \( R_{G} \) is optimal in the set of the fully parameterized state-space realizations, while \( R_{L_{2}} \) does not belong to the state-space realizations \( S_{H} \) and in \( R_{L_{2}} \) there is a smaller number of rounding operations. Other examples also show that the \( M_{L_{2}} \) value for \( R_{L_{2}} \) can be smaller than that for \( R_{\text{DISS}} \).

VII. CONCLUSION

In this paper, the digital filter structure problem in FWL implementation has been discussed. Our contribution is three-fold. Firstly, a new stability property of the ISS-structure has been revealed. Secondly, based on the DISS-structure a new structure has been developed, which is sparse and yields a very nice performance. The performance of this proposed structure has been analyzed by deriving the corresponding expressions for sensitivity measure and roundoff noise gain. A design example has been given, with which it is shown that the proposed structure is not only simpler than the DISS-structure but also generally yields a better performance which is very close to that of the fully parameterized optimal roundoff noise realizations.
REFERENCES


Gang Li (M'93) received the B. Engr. degree in electrical engineering from Beijing Institute of Technology, Beijing, China, in 1982, and the M.Eng. and the Ph.D. degrees both from Louvain University, Belgium, in 1988 and 1990, respectively.

He was with the Control Group at Louvain University as a Postdoctoral Researcher until April 1992. Since May 1992, he has been with the School of EEE, Nanyang Technological University. His research interests include digital filter and controller design, numerical problems in estimation and control applications, adaptive signal processing, signal transforms, and speech modeling/coding. Professor Li is a coauthor (with M. Gevers) of the Parametrizations in Control, Estimation, and Adaptive Filtering Problems: Accuracy Aspects (London, U.K.: Springer Verlag, 1993).

Michel Gevers (M'66-SM'86-F'90) received the B.E. degree from the Université Catholique de Louvain, Belgium, in 1968, and the Ph.D. degree from Stanford University, Stanford, CA, in 1972.

He is now Professor and President of CESAME (Center of Engineering Systems and Applied Mechanics) at the Université Catholique de Louvain in Louvain La Neuve, Belgium. He is the coordinator of the Belgian Inter University Pole on Modeling, Identification, Simulation and Control of Complex Systems, funded by the Federal Ministry of Science. His present research interests include system identification and its interconnection with robust control design. He has published 170 papers and conference papers. He is a coauthor (with R. R. Bitmead and V. Wertz) of Adaptive Optimal Control—The Thinking Man's GPC (Englewood Cliffs, NJ: Prentice Hall, 1990) and (with G. Li) of Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects (New York: Springer-Verlag, 1993).

Dr. Gevers is a distinguished member of the IEEE Control Systems Society. He is President of the European Union Control Association (EUCA), a member of the Board of Governors of the IEEE Control Systems Society, and Chairman of the International Committee of this Society, and Vice-President of the IEEE Control Systems Society.

Youssouf Sun received the Diploma from the Department of Chemical Engineering, Zhejiang University, China, in 1964.

He joined the Department of Chemical Engineering, Zhejiang University, in 1964. He is now Professor and Dean of the Institute of Control Science and Engineering, Zhejiang University. His research interests include robust control, nonlinear control, optimal control, and system identification.

Professor Sun is a Fellow of the Chinese Academy of Engineering.