On the impulse response of a discrete-time linear IIR system

Ranjan K. Mallik and Kaluri V. Rangarao

Abstract

An explicit formula for the impulse response of a discrete-time causal linear time-varying infinite-impulse-response (IIR) system in terms of the system parameters is obtained from the solution of a linear difference equation with variable coefficients. The case of a linear time-invariant system is also discussed. The results are used to approximate an IIR system by a finite-impulse-response system, and a relative truncation error criterion is presented to measure the accuracy of such an approximation.

Keywords: Discrete-time system; LTV IIR system; Impulse response; Linear difference equation; Variable coefficients; Explicit solution; FIR approximation

1. Introduction

In electrical engineering, particularly in signal processing, numerous practical systems are represented by discrete-time causal linear infinite-impulse-response (IIR) models [1,2], which serve as useful tools for both analysis and synthesis. The input-output behavior of such a system is viewed as a linear difference equation (also called a linear recurrence) [3,4] with coefficients or parameters which vary with the time-index if the system is time-varying or which are independent of the time-index if the system is time-invariant. Although the parameters completely characterize a system, its input-output behavior is better explained
by the impulse response, which is a function of the parameters. In the case of a linear time-invariant (LTI) IIR system, the impulse response can be obtained by transform techniques. However, for a linear time-varying (LTV) IIR system, transform techniques cannot be applied in general for arbitrarily varying coefficients of the corresponding recurrence. Although one resorts to numerical computation of the impulse response in such cases, it is well known that the impulse response is some function of the parameters, which is in a sum-of-products form. However, the complicated structure of the relation of the parameter indices to the impulse response has not been deciphered in the available open literature [5].

The explicit solution of a linear difference equation with variable coefficients was obtained recently [6]. Using this solution, we derive in this paper an explicit formula for the impulse response of a causal LTV IIR system by viewing the system as a difference equation in which the forcing term is the input to the system, the state is the output of the system, and the initial values are zero. The case of a causal LTI system is also discussed. The results are used to approximate an IIR system by a finite-impulse-response (FIR) system, and a relative truncation error criterion is presented to measure the accuracy of such an approximation.

The method is as follows. The solution of a linear difference equation is a sum of the homogeneous solution and the particular solution [7]. When the initial values are zero, the homogeneous solution goes to zero, and only the particular solution remains. We represent this particular solution as a linear combination of the forcing terms up to the present time index (owing to the causality of the system). The coefficients of this linear combination (which are explicitly in terms of the difference equation coefficients) give the impulse response of the causal LTV system which the difference equation represents. Such an LTV system is autoregressive (AR).

Next, we express the input of this AR LTV system as the output of another causal LTV system which is moving-average (MA). The composite autoregressive moving-average (ARMA) LTV system, which consists of the cascade of an MA system followed by an AR system, is then considered. The output of this ARMA system is again given by a linear combination of its inputs. The coefficients of this linear combination are expressed in terms of the AR and MA coefficients, and this comprises an explicit formula for the impulse response of the ARMA system.

2. A discrete-time causal LTV IIR system

Consider the linear difference equation

$$y_{k+N} = \sum_{j=1}^{N} a_{k,j} y_{k+N-j} + x_{k+N},$$

of order $N$ ($N \geq 1$) with state $y_{k+N}$, variable complex coefficients $a_{k,j}$, $j = 1, \ldots, N$, complex forcing term $x_{k+N}$, and complex initial values $y_1, \ldots, y_N$. It has been proved in Proposition 2 of [6] that the solution of this equation, which is an expression for $y_{k+N}$, for $k > 0$ in terms of only coefficients, initial values, and forcing terms, is given by

$$y_{k+N} = H_{k} N + P_{k+N},$$

where $H_{k} N$ and $P_{k+N}$ are functions of $k$.
where the homogeneous solution $H_{k+N}$ is expressed as
\begin{equation}
H_{k+N} = \sum_{j=1}^{N} d_{k+1,j}y_{N+1-j}. \tag{2b}
\end{equation}

the particular solution $P_{k+N}$ as
\begin{equation}
P_{k+N} = \sum_{j=2-k}^{0} d_{k,j}y_{N+1-j} + x_{k+N}. \tag{2c}
\end{equation}

and the combination coefficients $d_{k,j}$ are expressed in terms of the difference equation coefficients $a_{k,j}$ as
\begin{equation}
* u = \sum_{r=1}^{k+j-l} \sum_{1 \leq \ell_1, \ldots, \ell_r \leq N} \alpha_{k,j} \prod_{m=1}^{r} w_{m} \tag{2d}
\end{equation}

We now apply a change of variables on the state, the forcing term, and the coefficients. By substituting
\begin{align*}
w_k &= y_{k+N+1}, & k & \geq N, \\
v_k &= x_{k+N+1}, & k & \geq 0, \\
\alpha_{k,j} &= a_{k+1,j}, & j &= 1, \ldots, N, \ k \geq 0, \\
v_{k,j} &= \begin{cases} d_{k+1,j}, & j = -N, \ldots, k-1, \\
1, & j = k, \ k \geq 0, \end{cases}
\end{align*} \tag{3}

in (1) and (2), we can rewrite the difference equation (1) as an AR system in terms of the new state $w_k$ and new forcing term $v_k$ as
\begin{equation}
w_k = \sum_{j=1}^{N} \alpha_{k,j} w_{k-j} + v_k, \ k \geq 0. \tag{4}
\end{equation}

where $\alpha_{k,j}, j = 1, \ldots, N$ are the new complex coefficients and $w_{-N}, \ldots, w_{-1}$ are the complex initial values. The solution of (4), which is an expression for $w_k$ in terms of the forcing terms $v_0, \ldots, v_k$, initial values $w_{-N}, \ldots, w_{-1}$, and coefficients $\{\alpha_{k,j}\}$, can be written using (2) as
\begin{equation}
w_k = \sum_{j=-N}^{-1} v_{k,j} + \sum_{j=0}^{k} \sum_{|j|=j}^{N} \mu_{k,j} w_{k-j} \tag{5a}
\end{equation}

where $y_{\hat{\cdot}}$ can be expressed in terms of $\{\alpha_{k,j}\}$ as
\[
\gamma_{k,j} = \sum_{r=1}^{k-j} \mathbb{E} \left[ \prod_{m=1}^{r} d_{k-j-m} \right]
\]

\[\text{for } j = -N, \ldots, k-l, k-l. \quad (5b)\]

\[\gamma_{k,k} = 1 \quad \text{for } \epsilon > 0. \quad (5c)\]

Let the state $w_k$ of difference equation (4) represent the output of a discrete-time causal LTVIR system with time index $k$ whose input $u_k$ is related to the forcing term $v_k$ by an MA system given by the equation

\[v_k = \sum_{i=0}^{M} \beta_{k,i} u_{k-i}, \quad k \geq 0. \quad (6)\]

where $M > 0$, and whose initial input and output conditions are zero, that is,

\[u_{k-1} = 0, \quad w_{k-1} = 0. \quad (7a)\]

or simply

\[u_k = 0, \quad w_k = 0, \quad \text{for } k < 0. \quad (7b)\]

Combining the AR(N) system (4) with the MA(M) system (6), we have a time-varying ARMA(M, N) system given by the recurrence

\[u_k = \sum_{j=1}^{N} \alpha_{k,j} u_k+j + \sum_{i=0}^{M} \beta_{k,i} u_{k-i}, \quad k \geq 0. \quad (8)\]

with initial conditions (7). The AR parameters are $\alpha_k, 1, \ldots, \alpha_k, N$, while the MA parameters are $\beta_{k,0}, \ldots, \beta_{k,M}$.

3. **The impulse response**

If $h_{k,i}$ denotes the impulse response of the LTV system (8), then the input $u_k$ and output $w_k$ are governed by the relation

\[w_i = f^2 k \sum_{j=0}^{k} h_{k,j} u_i, \quad k \geq 0. \quad (9)\]

Due to the initial conditions being zero (see (7)), we have from (5a)

\[w_i = f^2 k \sum_{j=0}^{k} h_{k,j} u_i, \quad k \geq 0. \quad (10)\]

where, as in (5b) and (5c),
\[ \gamma_{j} = \sum_{i=\lceil (k-j)/N \rceil}^{k-j} \sum_{l_1, \ldots, l_r \leq N} \prod_{m=1}^{r} a_{k - \sum_{l=1}^{m-1} l_n} h, l_m \]

\[ \gamma_{k, k} = 1 \text{ for } k > 0, \quad (11b) \]

\( \lceil \cdot \rceil \) denoting the ceiling function. Note that the condition \( l_1 < j \) in the composite summation (over \( l_1, \ldots, l_r \)) of (5b) has been omitted in (11a) because now the index \( j \) is nonnegative and hence this condition is automatically satisfied. Moreover, the minimum value of \( r \) can be only \( \lceil (k-j)/N \rceil \), due to the conditions \( 1 < l_1, \ldots, h < N \), \( 1 + l_2 - l_1 \), \( \ldots \), \( k_j - h \). When \( k_j - h < N \), \( \lceil (k-j)/N \rceil = 1 \).

Combining (10) and (6), we get the output \( w_k \) in terms of the input \( u_k \) resulting in the expression

\[ w_k = \sum_{j=0}^{k} \gamma_{k, j} \sum_{i=0}^{M} p_{j} u_{j-i}, \quad (12) \]

Owing to the zero initial conditions (7), we can rewrite (12) as

\[ w_k = \sum_{j=0}^{k} \gamma_{k, j} \sum_{i=0}^{\min(j, M)} p_{j} \beta_{j, i} u_{j-i}. \quad (13) \]

Replacing the variable \( j - i \) by \( i \) in the inner summation of (13) yields

\[ w_k = \sum_{j=0}^{k} \gamma_{k, j} \sum_{i=0}^{\max(0, j-M)} \beta_{j, i} u_{j-i}. \quad (14) \]

When \( k \leq M \), we have \( \max(0, j-M) = 0 \) for \( j = 0, \ldots, k \) in (14), and therefore (14) can be written as

\[ w_k = \sum_{j=0}^{k} \gamma_{k, j} \sum_{i=0}^{j-M} \beta_{j, i} u_{j-i} = \sum_{i=0}^{k} u_{i} \sum_{j=1}^{k} \gamma_{k, j} \beta_{j, i} u_{j-i}. \quad (15) \]

On the other hand, for \( k > M + 1 \), we can express (14) as

\[ w_k = \sum_{j=0}^{M} \gamma_{k, j} \sum_{i=0}^{j-M} \beta_{j, i} u_{j-i} + \sum_{j=M+1}^{k} \gamma_{k, j} \sum_{i=0}^{j-M} \beta_{j, i} u_{j-i}, \quad k > M + 1. \quad (16) \]

Interchanging the summations over \( i \) and \( j \) in (16), we get

\[ w_k = \sum_{i=0}^{k-M} \gamma_{i} \sum_{j=i}^{i+M} \gamma_{k} \beta_{j, i} u_{j-i} + \sum_{i=k-M+1}^{k} \gamma_{i} \sum_{j=i}^{k} \gamma_{k} \beta_{j, i} u_{j-i}, \quad k > M + 1. \quad (17) \]

We can now combine (15) and (17) to yield a single expression

\[ w_k = \sum_{j=0}^{k} \gamma_{k, j} \sum_{i=0}^{\min(i+M, k)} \beta_{j, i} u_{j-i}, \quad k > 0. \quad (18) \]
Comparing (18) with (9), we conclude that the impulse response \( h_{k,i} \) of the system (8) with input \( u_k \) and output \( w_k \) is given by

\[
\min(\{M,k\})
\]
\[
h_{k,i} = \sum_{j=0}^{\min(\{M,k\})} y_{k,j} \beta_{j,i-j}, \quad i = 0, \ldots, k, \ k \geq 0,
\]

where \( y_{ij} \) is expressed in (11) in terms of the AR parameters.

Alternatively, we can write (19) as

\[
h_{k,i} = \sum_{j=\max(i-M,0)} \gamma_{k,k-j}^P_{k-j,i-j}, \quad i = 0, \ldots, k, \ k \geq 0.
\]

Thus (19) combined with (11), or, alternatively, (20), provides an explicit formula for the impulse response in terms of the AR and MA parameters of the causal LTV IIR system (8).

When \( M = 0 \), we get \( h_{k,i} = \gamma_{k,i}^P \beta_i \), which corresponds to an AR(N) system.

### 3.1. An example

For example, consider an ARMA(1, 2) system

\[
u_k = \alpha_{1,1} u_{k-1} + \alpha_{1,2} u_{k-2} + \beta_{k,0} u_k + \beta_{k,1} u_{k-1}, \quad k \geq 0.
\]

with initial conditions

\[
U-2 = W_j = 0, \quad W-2 = W_{-i} = 0,
\]

where we need to find \( h^{*k-i} \) for \( i = 0, 1, 2, 3, 4 \).

We first obtain \( Y_{k,j} \) for \( j = 0, 1, 2, 3, 4 \). Now \( \gamma k, k \) = 1. To get \( Y_{k,j} \) for \( j = 1, 2, 3, 4 \), we consider the enumerations of \((h, ..., l)\) where \( 1 < l_1, ..., l_r < l \), \( 1 \leq l_1 + l_2 + \cdots + l_r = j \), \( j = 0, 1, 2, ..., j \).
Therefore, from (20b),
\[ \begin{align*}
Y_{k, k-1} &= \alpha_{k, 1}, \\
Y_{k, k-2} &= \alpha_{k, 2} + \alpha_{k, 1}^2, \\
Y_{k, k-3} &= \alpha_{k, 1} \alpha_{k-1, 1}^2 + \alpha_{k, 1} \alpha_{k-2, 1}^2 + \alpha_{k-1, 1}^2, \\
Y_{k, k-4} &= \alpha_{k, 1} \alpha_{k-1, 1} \alpha_{k-2, 1} + \alpha_{k, 1} \alpha_{k-2, 1}^2 + \alpha_{k-1, 1}^2 \alpha_{k-2, 1}^2 + \alpha_{k-1, 1} \alpha_{k-2, 1} \alpha_{k-3, 1} + \alpha_{k, 1} \alpha_{k-2, 1} \alpha_{k-3, 1}
\end{align*} \]
and, using (20a), we obtain
\[ \begin{align*}
h_{k, k} &= \beta_k, \quad \gamma_{k, k} = \beta_k, \\
h_{k, k-1} &= \beta_k, \gamma_{k, k-1} = \beta_k, \\
h_{k, k-2} &= \beta_k, \gamma_{k, k-2} = \beta_k, \\
h_{k, k-3} &= \beta_k, \gamma_{k, k-3} = \beta_k, \\
h_{k, k-4} &= \beta_k, \gamma_{k, k-4} = \beta_k.
\end{align*} \]

4. A discrete-time LTI system

In the case of a discrete-time LTI system, the AR and MA parameters in (8) do not vary with the time index \( k \); that is,
\[ \alpha_{k, j} = \alpha_j, \quad j = 1, \ldots, N, \quad \beta_k, i = \beta_i, \quad i = 0, \ldots, M. \]
This results in the LTI system
\[ w_k = \sum_{i=1}^{N} \alpha_j w_{k-j} + \sum_{i=0}^{M} \beta_i w_{k-i}, \quad k \geq 0. \quad (21) \]
We can then express \( \gamma_{k, j} \) in (11) as
\[ \begin{align*}
\gamma_{k, j} &= Y_{k, j} = \sum_{r=0}^{k-j} \sum_{(j_1, \ldots, j_r) \leq N} \sum_{l_1+\cdots+l_r = j} \sum_{l_1+\cdots+l_r = k-j} \prod_{m=1}^{r} \alpha_{l_m}
\end{align*} \]
for \( k-j = 1, \ldots, k \), \( k \geq 1 \). \quad (22a)
\[ \gamma_{k, 0} = 1 \quad \text{for } k \geq 0. \quad (22b) \]
Thus \( y_{tj} \) depends only on the difference \( k - j \) and is denoted as \( Y_{k-j} \). The impulse response \( h_{k,i} \) in (19), which now depends only on \( k - i \) and is denoted as \( h_{k-i} \), becomes

\[
h_{k,i} = h_{k-i} = \sum_{j-i=0}^{\min(M,k-i)} Y_{k-j} \beta_{j-i}, \quad k-i = 0, \ldots, k, \quad k \geq 0, \tag{23}
\]

where \( y_{t,j} \) is given by (22).

Note that in (22a), the sum of \( \alpha_{t_1} \cdots \alpha_{t_r} \) over all \( l_1, \ldots, l_r \), satisfying

\[
i \leq l_1, \ldots, l_r \leq N, \quad l_1 + l_2 + \cdots + l_r = k - j
\]

when \( r = \lfloor (k - j)/N \rfloor, \ldots, k - j \) is the same as the sum of

\[
\left( t_1 + t_2 + \cdots + t_N \right) \alpha_{t_1}^0 \cdots \alpha_{t_N}^0 \frac{(t_1 + t_2 + \cdots + t_N) \beta_{t_1}^0 \cdots \beta_{t_N}^0}{h[t_1+\ldots+t_N]} \sum_{1 \leq i_1 \leq \ldots \leq N}
\]

over all \( t_1, \ldots, t_N \) satisfying

\[
t_1 + t_2 + \cdots + t_N = k - j.
\]

Therefore, we can rewrite (23) as

\[
h_i = \sum_{j=m \leq (i-M,0)} Y_{k-j} \beta_{i-j}, \tag{24a}
\]

where, from (22),

\[
Y_{k,k-j} = y_j = \sum_{i=1}^{n_1=n_{k-j}} (t_1 + t_2 + \cdots + t_N) \beta_{t_1}^0 \cdots \beta_{t_N}^0 \prod_{m=1}^{n} \alpha_{t_m}^m, \quad j \geq 0. \tag{24b}
\]

Thus (23) combined with (22), or, alternatively, (24), is an explicit formula for the impulse response of the LTI system (21).

4.1. An example

Consider the ARMA(1,2) system of Section 3.1 with the LTI condition, that is,

\[
w_k = \sigma_1 u_{k-1} + \sigma_2 u_{k-2} + \beta_0 u_k + \beta_1 u_{k-1}, \quad k \geq 0,
\]

with initial conditions

\[
u_{-2} = u_{-1} = 0, \quad w_{-2} = w_{-1} = 0,
\]

where we need to find \( h_i \) for \( i = 0, 1, 2, 3, 4 \).

From the example in Section 3.1, we get \( \gamma_0 = 1 \), and

\[
\gamma_1 = \sigma_1, \quad \gamma_2 = \sigma_2 + \sigma_1^2, \quad \gamma_3 = 2\sigma_1 \sigma_2 + \sigma_1^3, \quad \gamma_4 = 2\sigma_2^2 + 3\sigma_1^2 \sigma_2 + \sigma_1^5,
\]

which satisfies (24b) and (22a). Therefore, from (24a), we obtain
\[ h_0 = \beta_0 \gamma_0 = \beta_0, \quad h = AK_0 + \beta_0 \gamma_1 = A + \beta_0 \alpha_1, \]
\[ h_2 = \beta_1 \gamma_1 + \beta_0 \gamma_2 = \beta_1 \gamma_1 + \beta_0 (\alpha_2 + \alpha_1), \]
\[ h_3 = \beta_1 \gamma_2 + \beta_0 \gamma_3 = A (\alpha_2 + 3^1) + \beta_0 (2 \alpha_1^2 + 3^1), \]
\[ h_4 = AK_3 + AK_4 = A (2 \alpha_1 Q^2 + 3^1) + \alpha_2 (2^2 + 3^1), \]

5. FIR approximation

It is often desirable to approximate an IIR system by an FIR one, since an FIR system is stable and easily implementable. For this purpose, we consider the following IIR AR(N) system with input \( u_k \) and output \( w_k \):

\[ u_k = \sum_{i=0}^{N} a_k u_{k-i} + b_k u_k, \quad k \geq 0. \]  

(25)

Suppose we wish to approximate the above system by an MA(M) system which is FIR. The approximate system can be written as

\[ \hat{w}_k = \sum_{i=1}^{M} \beta_{k,i} u_{k-i} + \beta_{k,0} u_k, \quad k \geq 0. \]  

(26)

where \( \hat{w}_k \) is the approximate output at time index \( k \). The coefficients of the approximate system are \( \beta_{k,i}, i = 1, \ldots, M \). Now, from (9) and (20a), the actual system has the representation

\[ w_k = \sum_{i=1}^{k} \gamma_{k,k-i} u_{k-i} = \gamma_{k,k}, Q u_k. \]  

(27)

Since \( \gamma_{k,k} = 1 \), (27) can also be written as

\[ w_k = \gamma_{k,k} Q u_k + \beta_{k,0} u_k. \]  

(28)

Comparing (26) and (28), it is observed that if we choose

\[ a_{k,i} = h_{k-k-i} = n, k-i P_k - i b, \quad i = 1, \ldots, M, \]  

(29)

then the approximate FIR system is a truncated version of the IIR one. We need to know the value of \( \hat{M} \) we can choose for such an approximation to be good. Consider the absolute sum of the impulse response \( h^* \) for time index \( k \) when \( i \) goes from \( \hat{M} + 1 \) to \( k \), which we call the truncation error and denote as \( E(\hat{M}, k) \). A reasonable criterion for measuring the goodness of such an approximation is an upper bound on this truncation error.

From (29), the truncation error is given by

\[ E(\hat{M}, k) = \sum_{i=\hat{M}+1}^{k} |h_{k-k-i}| = J_2 \sum_{i=\hat{M}+1}^{k} |\gamma_{k,k-i} P_{k-i}|. \]  

(30)
Let the complex coefficients $\beta_k, 0$ and $a^i, \cdots, o^w$ of the original AR(N) system be bounded by

$$|\beta_{k,0}| \leq B, \quad |a_{j,k}| \leq f_j(k), \quad j = 1, \ldots, N, \quad k \geq 0,$$

where $B$ is a finite positive constant and $f_1(k), \ldots, f_N(k)$ are nonnegative monotonically nondecreasing functions of the time index $k$. Then, from (30), we obtain the following upper bound for $E(\hat{M}, k)$:

$$\mathcal{E}(\hat{M}, k) \leq B \sum_{i=\hat{M}+1}^{k} |n_{k-i}|.$$  \hspace{1cm} (32)

Now from (20b), we get

$$r = \sum_{i=1}^{M} E \left[ \prod_{n=-1}^{r} n_{i}(k) \right] \text{ for } i = 1, \ldots, k, \quad k \geq 1.$$ \hspace{1cm} (34)

The bound on the right-hand side of (33) is similar to the expression (22a) for $yt-j$. Therefore, it can also be written in a form similar to (24b), which results in the inequality

$$|y_{k-i}| \leq \sum_{(t_1, \ldots, t_N) \in U, t_1 + \cdots + t_N = 0} \frac{(h+t_i-1)}{h+t_i-1} \left[ \prod_{m=1}^{N} \{f_m(k)\}^{t_m} \right].$$ \hspace{1cm} (34)

In the composite summation on the right-hand side of (34), we have

$$t_1 + t_2 + \cdots + t_N = 0.$$ \hspace{1cm} (40)

When $t_1 + t_2 + \cdots + t_N = 0$, the maximum value of $t_1 + t_2 + \cdots + t_N$ occurs when $t_1 = i, \quad t_2 = \cdots = t_N = 0$, while the minimum occurs when $t_N = \left\lfloor \frac{i}{N} \right\rfloor$.  \hspace{1cm} (41)

Therefore

$$\frac{i}{N} \leq t_1 + t_2 + \cdots + t_N \leq 1.$$ \hspace{1cm} (42)

when $i$ is a multiple of $N$,

$$\left\lfloor \frac{i}{N} \right\rfloor + 1 < t_1 + t_2 + \cdots + t_N < i$$ \hspace{1cm} (43)

when $i$ is not a multiple of $N$.

implying

$$\left\lfloor \frac{i}{N} \right\rfloor \in t_1 + t_2 + \cdots + t_N \leq t.$$ \hspace{1cm} (44)
Putting $t_1 + t_2 + h t N = q$, we can have a further upper bound on the right-hand side of (34) to yield

$$\gamma_k, k = 1 \leq \sum_{q=\frac{i}{N}}^{i} E \left[ \frac{q^l}{t_k^{d-k_0}} \prod_{m=1}^{N} \left( f_m(k) f_n \right) \right]$$

Using the result in (35), we get

$$i + 1 - N i = N - 1 \text{ if } i \neq 1, m_1 \leq m_2,$$

in (35), we get

$$\gamma_k, k = 1 \leq \left\{ \begin{array}{ll}
\left[ \sum_{m=1}^{N} f_m(k) \right]^{i/N} \sum_{m=1}^{N} f_m(k) \prod_{m=1}^{N} \left( f_m(k) f_n \right) \right]^{i/N} & \text{if } \sum_{m=1}^{N} f_m(k) \neq 1, \\
i + 1 - N i & \text{if } \sum_{m=1}^{N} f_m(k) = 1.
\end{array} \right.$$  (37)

Since $\sqrt{N}/N > i/N$, the right-hand side of (37) can again be bounded by

$$\gamma_k, k = 1 \leq \left\{ \begin{array}{ll}
\left[ \sum_{m=1}^{N} f_m(k) \right]^{i/N} \sum_{m=1}^{N} f_m(k) \prod_{m=1}^{N} \left( f_m(k) f_n \right) \right]^{i/N} & \text{if } \sum_{m=1}^{N} f_m(k) \neq 1, \\
i + 1 - N i & \text{if } \sum_{m=1}^{N} f_m(k) = 1.
\end{array} \right.$$  (38)

Substituting (38) in (32), and applying (36), we obtain

$$\mathcal{E}(\hat{M}, k) \leq \left\{ \begin{array}{ll}
B & \text{if } \sum_{m=1}^{N} f_m(k) \neq 1, \\
\left[ \sum_{m=1}^{N} f_m(k) \right]^{i+1} \left[ \sum_{m=1}^{N} f_m(k)^{i+1} \right] & \text{if } \sum_{m=1}^{N} f_m(k) = 1.
\end{array} \right.$$  (39)

It is clear from (39) that $E(\hat{M}, k)$ is bounded for all $k$ and all choices of $\hat{M}$ provided

$$\sum_{m=1}^{N} f_m(k) X < l, \quad k^O,$$

where $A$ is a finite positive constant. We shall call $A$ the bound on the absolute sum of the AR coefficients. Now both $E(\hat{M}, k)$ and its bound on the right-hand side of (39) increase with an increase of $k$. So when (40) is satisfied, the maximum truncation error, which we denote as $E_{\text{max}}(\hat{M})$, is bounded by
\[ E_{\text{max}}(\hat{M}) = \lim_{k \to \infty} E(\hat{M}, k) \]

\[ = \frac{B}{1 - \sum_{n=1}^{N} f_n(k)} \left( \left[ \sum_{m=1}^{N} f_m(k) \right]^{1/N} \right)^{\hat{M} + 1} - \frac{B}{1 - \sum_{n=1}^{N} f_n(k)} \left( \left[ \sum_{m=1}^{N} f_m(k) \right]^{1/N} \right)^{\hat{M} + 2} \]

(41)

It can be easily shown that for a given \( M \) and \( N \), the right-hand side of (41) monotonically increases with an increase in \( J2^{\hat{M}+1}f_n(k) \) when \( 0 < J2^{\hat{M}+1}f_n(k) \leq A_s \) and \( 0 < A < 1 \) from (40). Therefore, under condition (40), \( E_{\text{max}}(\hat{M}) \) is bounded by

\[ E_{\text{max}}(\hat{M}) \leq \frac{B}{(1 - A)} \left( A^{1/N} \right)^{\hat{M} + 1} \left( 1 - AV^N - \frac{1}{1 - A} \right) \]

(42)

where \( B \) is the upper bound on \( |\beta_k,0| \), and \( A \) is the upper bound on the absolute sum of the AR coefficients of the AR(N) system represented by (25).

To find an estimate of the relative error due to truncation, we consider the absolute sum of the impulse response, expressed as

\[ S(k) = \sum_{i=1}^{k} |h_k, k-i| = \sum_{i=1}^{k} |n, k-i \tilde{p} \\tilde{p}_i, o| \]

(43)

Note that

\[ S(k) = E(\hat{M}, k) \big|_{\hat{M}=0} = E(0, k) \]

where \( E(\hat{M}, k) \) is given by (30). Denoting \( S_{\text{max}} \) as the maximum absolute impulse response sum, we obtain, by putting \( \hat{M} = 0 \) in (41),

\[ S_{\text{max}} = \lim_{k \to \infty} S(k) \]

\[ \leq \frac{B}{1 - \sum_{n=1}^{N} f_n(k)} \left( \left[ \sum_{m=1}^{N} f_m(k) \right]^{1/N} \right)^{\hat{M} + 1} - \frac{B}{1 - \sum_{n=1}^{N} f_n(k)} \left( \left[ \sum_{m=1}^{N} f_m(k) \right]^{1/N} \right)^{\hat{M} + 2} \]

which implies, from condition (40), that

\[ S_{\text{max}} = \frac{B}{(1 - A)} \left( A^{1/N} \right)^{\hat{M} + 1} \left( 1 - AV^N - \frac{1}{1 - A} \right) \]

(45)

An estimate of the relative truncation error can be expressed as the ratio of the upper bounds of \( E_{\text{max}}(\hat{M}) \) and \( S_{\text{max}} \), given by

\[ R(\hat{M}, N, A) = \frac{E_{\text{max}}(\hat{M}, k)}{S_{\text{max}}} \]

(46)

The quantity \( R(\hat{M}, N, A) \) decreases with an increase of \( \hat{M} \) and increases with an increase of \( N \) or an increase of \( A \).

For a given IIR system of order \( N \), and bound on the absolute sum of AR coefficients \( A \), we can fix a maximum value \( R_{\text{max}} \) for \( R(\hat{M}, N, A) \) and find the minimum value of \( \hat{M} \), say \( \hat{M}_{\text{min}} \), which satisfies

\[ U(M_{\text{min}}, N, A) < R_{\text{max}}, \quad U(M_{\text{min}} - 1, N, A) > TU^* \]
Once the minimum order of the approximate FIR system is estimated, we choose the MA coefficients of the approximate FIR system as

\[ \hat{p}_{kj} = \hat{h}_{kj} \cdot \hat{y}_{k-j} = \sum_{i=1}^{\hat{M}_{\text{min}}} \hat{c}_{k,i} \cdot \hat{y}_{k-j} \cdot \hat{c}_{i,0}, \quad i = 1, \ldots, \hat{M}_{\text{min}}, \]

where \( \hat{y}_{k-j} \) can be computed in terms of the AR parameters as in (20b). Note that the enumerations of \( 1, \ldots, i \) in the composite summation of (20b) can be prestored as lookup tables.

Table 1 gives values of \( \hat{M}_{\text{min}} \) for \( N = 1, \ldots, 10 \) when \( A = 0.2, 0.4, 0.6, 0.8 \) and \( R(\hat{M}, N, A) < T \max = 0.01, 0.1 \). We find that for a given \( A \), \( \hat{M}_{\text{min}} \) increases with an increase of \( N \), and, for a given \( N \), it increases with an increase of \( A \). Moreover, for a given \( N \) and \( A \), there is a tradeoff between \( \hat{M} \) and the relative truncation error. When \( A = 0.01 \), we see that \( \hat{M}_{\text{min}} \sim 3N, 5N, 9N, 21N \) for \( A = 0.2, 0.4, 0.6, 0.8 \), respectively. In comparison, when \( R_{\text{max}} = 0.1 \), which is an increase by a factor of 10, \( \hat{M}_{\text{min}} \sim 1.5N, 2.5N, 4.5N, 10.5N \) for \( A = 0.2, 0.4, 0.6, 0.8 \), respectively, implying a reduction by a factor of 1/2. Thus a lower relative truncation error calls for a higher value of \( \hat{M} \).
5.1. Example of an LTV system

Consider a weighted averager with exponentially varying weights, whose output \( w_k \) is a weighted mean of the inputs \( u_0, \ldots, u_K \) and is given by

\[
    w_k = \frac{1}{(1 + c)(1 - c) k^2} \sum_{i=0}^{k} u_i, \quad 0 < c < 1. \tag{47}
\]

We can express the averager (47) in the form of an AR(1) system having input \( u_k \) and output \( w_k \) governed by the relation

\[
    w_k = \left( \frac{c - c^k + 1}{1 - c^k + 1} \right) a_{k-1} + \left( \frac{1 - c}{1 - c^k + 1} \right) u_k, \quad k \geq 0, \quad 0 < c < 1, \tag{48}
\]

with initial value \( w_0 = 0 \). This corresponds to a time-varying first order lowpass filter. By comparing (48) with (25), the AR parameter \( a^1 \) and the MA parameter \( \beta_k, \theta \) are expressed as

\[
    a^1_k = \frac{c - c^k + 1}{1 - c^k + 1} = c, \quad \theta = \frac{1 - c}{1 - c^k + 1}. \tag{49}
\]

Consider the case when \( c = 0.4 \). Suppose we want to approximate the system (48) by an FIR system with \( R_{\text{max}} = 0.01 \). Here \( N = 1 \), and \( A \) is an upper bound on the AR parameter \( a^1 \) given by (49). It is easily seen that when \( 0 < c < 1 \),

\[
    a^1 = \frac{1 - c^k}{1 - c^k + 1} \leq c.
\]

and therefore \( A = c = 0.4 \). From Table 1, we find that \( N = 1, R_{\text{max}} = 0.01, A = 0.4 \) gives \( M_{\text{min}} = 6 \).

The coefficients of the approximate MA(6) system are given by

\[
    \hat{P}_{k,i} = h_{k,i} = Y_{k,i} / \left( k - m + 1 \right), \quad i = 1, \ldots, 6, \tag{50a}
\]

where, from (20b),

\[
    Y_{k,i} = \sum_{n=1}^{\infty} Y_n k^{n-m+1}, \tag{50b}
\]

Substituting (49) in (50), we obtain

\[
    \hat{P}_{k,i} = h_{k,i} = c \left( \frac{1 - c^{k+1}}{1 - c^{k+1}} \right) \left( \frac{1 - c}{1 - c^{k+1}} \right) = \frac{c^i - c^{i+1}}{1 - c^{k+1}}, \quad i = 1, \ldots, 6.
\]

Therefore, the approximate FIR system is given by
\[ \hat{w}_k = \hat{b}_{k,0} u_k + \sum_{i=1}^{6} \hat{b}_{k,i} u_{k-i} \]
\[ = \left( \frac{1-0.4}{1-0.4^{k+1}} \right) u_k + \left( \frac{1-0.4}{1-0.4^{k+1}} \right) \sum_{i=1}^{6} 0.4^i u_{k-i} \]
\[ = \frac{1}{1-0.4^{k+1}} \left[ 0.6u_k + 0.24\mathcal{E} \cdot i + 0.096u_{k-2} + 0.0384u_{k-3} \right. \]
\[ \left. + 0.01536\mathcal{E} \cdot -5 + 0.006144\mathcal{E} \cdot -1 + 0.0024576\mathcal{E} \cdot -1 \right] . \]

5.2. Example of an LTI system

Consider the LTI AR(2) system
\[ w_k = 0.16 w_{k-1} - 0.64 w_{k-2} + u_k, \]
with input \( u_k \) and output \( w_k \), which corresponds to a second order bandpass filter whose radial pole position is \( 0.64 = 0.8 \). Suppose we want to approximate this system by an FIR system with \( R_{\text{max}} = 0.1 \).

Here \( N = 2, A = 0.16 + 0.64 = 0.8 \), and from Table 1, we find that \( M_{\text{min}} = 25 \). Since \( S_0 = 1 \), we have \( \tilde{h} = h_k \), and owing to the fact that \( N = 2 \), the MA coefficients of the approximate system can be expressed using (22a) as
\[ \hat{b}_{2m+1} = h_{2m+1} = \sum_{i=1}^{m+1} \left( \frac{m+i}{2i-1} \right) \alpha_1^{2i-1} \alpha_2^{m+1-i}, \quad m = 0, 1, 2, \ldots, \]
\[ \hat{b}_{2m} = h_{2m} = \sum_{i=0}^{m} \left( \frac{m+i}{2i} \right) \alpha_1^{2i} \alpha_2^{m-i}, \quad m = 1, 2, 3, \ldots, \]
where \( \alpha_1 = 0.16 \) and \( \alpha_2 = -0.64 \). The first six MA coefficients of the approximate system can be computed as follows:
\[ \hat{h}_1 = h_1 = 0.16 , \]
\[ \hat{h}_3 = h_3 = 2\alpha_2 \alpha_3 + \alpha_1 = 0.2007 , \]
\[ \hat{h}_5 = h_5 = 3\alpha_1 \alpha_2^2 + 4\alpha_1^2 \alpha_3 + \alpha_1^3 = 0.3611 , \]
\[ \hat{h}_7 = h_7 = 3\alpha_1 \alpha_2^2 + 4\alpha_1^2 \alpha_3 + \alpha_1^3 = 0.1862 , \]
\[ \hat{h}_9 = h_9 = 3\alpha_1 \alpha_2^2 + 4\alpha_1^2 \alpha_3 + \alpha_1^3 = -0.2013 . \]

We need to perform the computation up to \( \tilde{e} \cdot \). Therefore, the approximate FIR system is given by
\[ \hat{w}_k = u_k + \sum_{m=0}^{12} \left[ \sum_{i=1}^{m+1} \left( \frac{m+i}{2i-1} \right) \mathcal{E}^{2i-1} (0.64)^{m+1-i} \right] u_{k-2m-i} . \]
\[ \sum_{m=1}^{12} \left( \sum_{i=0}^{m} \binom{m+i}{2i} (0.16)^i (-0.64)^{m-i} \right) u_{k-2m} \]

\[ = u_k + 0.6 \delta u_{k-1} - 0.6U4u_{k-2} - 0.2007w_k \epsilon_3 + 0.3611w_k \epsilon_4 + 0.1862M_{-5} - 0.2013u_{k-6} - \cdots + h_{24}u_{k-24} + h_{25}u_{k-25}. \]

6. Conclusions

Starting with the solution of a linear nonhomogeneous difference equation with variable coefficients, we have obtained an explicit formula for the impulse response of a discrete-time causal LTV IIR system in terms of the system coefficients. The formula utilizes the combinatorial properties of the coefficient indices. It can be easily simplified to the LTI case. We make use of the impulse response terms to approximate an IIR system by an FIR one and specify the accuracy of such an approximation by a relative truncation error criterion.

References


Ranjan K. Mallik was born in Calcutta, India, on November 15, 1964. He received the B.Tech. degree from the Indian Institute of Technology, Kanpur, in 1987 and the M. S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1988 and 1992, respectively, all in electrical engineering. From August 1992 to November 1994 he was a scientist at the Defence Electronics Research Laboratory, Hyderabad, India, working on missile and EW projects. From November 1994 to January 1996 he was a faculty member of the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur. In January 1996 he joined the faculty of the Department of Electronics and Communication engineering, Indian Institute of Technology, Guwahati, where he worked until December 1998. Since December 1998 he has been with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, where he is an associate professor.

His research interests are in communication theory and systems, difference equations, and linear algebra.

Dr. Mallik is a member of Eta Kappa Nu. He is also a senior member of the Institute of Electrical and Electronics Engineers, Inc. (IEEE), a fellow of the Institution of Electronics and Telecommunication Engineers, a member of the American Mathematical Society and the International Linear
Kaluri V. Rangarao was born in Vizianagaram, India, on November 27, 1952. He received the B.E. degree in electrical engineering from Andhra University, Visakhapatnam, in 1974, the M.E. degree in Automation and Computer Science from the Indian Institute of Science, Bangalore, in 1977, the M.S. degree in systems engineering (Electronic Warfare) from the US Naval Postgraduate School, Monterey, in 1991, and the Ph.D. degree in electrical engineering from the Indian Institute of Technology, Madras, in 1994.

From 1977 to 1989, he was a scientist and deputy director at the Naval Systems Laboratory, Visakhapatnam. From 1989 to 1991, he was an International Military Education Scholar at the US Naval Postgraduate School, Monterey. He was the head of the Communication Intelligence Division, Defence Electronics Research Laboratory, Hyderabad, India, from 1991 to 1995 and the head of the Seeker System Analysis Division, Research Centre Imarat, Hyderabad, India, from 1995 to 1999. From 1999 to 2000, he was the president and CEO of GMR Vasavi Infotech., Hyderabad, India.

Since 2000, he has been with Satyam Computer Services Ltd., Hyderabad, India, where he is a vice president, heading a team of 250 software professionals in the areas of telecom and embedded software development. He has been a visiting faculty at JNTU, Hyderabad, India, and a visiting industrial researcher at Oklahoma State University.

He has expertise in business development, software delivery and team building, full life cycle product development and field trials, end-to-end solution development from concept to reality, and design and development of underwater systems, electronic surveillance systems, and missile systems. His research interests are in digital signal processing, wireless communications, and real-time embedded systems.

Dr. Rangarao is a senior member of the Institute of Electrical and Electronics Engineers, Inc. (IEEE); the chairman, Computer Society of India, Vizag Chapter; and a board of studies member, JNTU.