A Closed Form Solution to $L_2$-Sensitivity Minimization of Second-Order State-Space Digital Filters

Shunsuke Yamaki, Masahide Abe and Masayuki Kawamata
Department of Electrical Engineering, Graduate School of Engineering, Tohoku University
6-6-05, Aoba, Aramaki, Aoba-ku, Sendai, 980-8579 Japan
Email: yamaki@mk.ecei.tohoku.ac.jp

Abstract—This paper proposes a closed form solution to $L_2$-sensitivity minimization of second-order state-space digital filters. Restricting ourselves to the 2-nd order case of state-space digital filters, we can formulate $L_2$-sensitivity minimization problem by hyperbolic functions. As a result, $L_2$-sensitivity minimization problem can be converted into a problem to find the solution to a 4-th degree polynomial equation of constant coefficients, which can be algebraically solved in closed form without iterative calculations.

I. INTRODUCTION

Several functions which evaluate the coefficient quantization effects of state-space digital filters have been proposed. Most of them can be classified into two main categories: $L_1/L_2$-mixture sensitivity [1], [2] and $L_2$-sensitivity [3], [4]. It is already proved that $L_1/L_2$-mixture sensitivity of a filter is minimized when the filter is a balanced realization, which can be derived analytically [1], [2]. On the other hand, it is quite difficult to minimize $L_2$-sensitivity by analytical method because $L_2$-sensitivity minimization problem is a nonlinear problem. To the $L_2$-sensitivity minimization problem, solutions using iterative calculations are proposed in [3] and [4].

In this paper, we propose a closed form solution to the $L_2$-sensitivity minimization problem of 2-nd order state-space digital filters. We show that the $L_2$-sensitivity minimization problem for 2-nd order state-space digital filters can be formulated by hyperbolic functions. We can obtain the minimum $L_2$-sensitivity by solving a 4-th degree polynomial equation of constant coefficients in closed form without iterative calculations.

II. $L_2$-SENSITIVITY OF STATE-SPACE DIGITAL FILTERS

A. State-Space Digital Filters

For a given $N$th-order transfer function $H(z)$, a state-space digital filter can be described by the following state-space equations:

\[ x(n+1) = Ax(n) + bu(n) \]
\[ y(n) = cx(n) + du(n) \]

where $x(n)$ is an $N \times 1$ state-vector, $u(n)$ is a scalar input, $y(n)$ is a scalar output and $A$, $b$, $c$, $d$ are real constant matrices of appropriate dimensions. The transfer function of the digital filter $(A, b, c, d)$ is given by

\[ H(z) = c(zI - A)^{-1}b + d. \]  (3)

B. $L_2$-Sensitivity

The $L_2$-sensitivity is one of the measurements which evaluates coefficient quantization effects of digital filters. In [3], the $L_2$-sensitivity of the filter $H(z)$ with respect to the realization $(A, b, c, d)$ is defined by

\[ S(A, b, c) = \left\| \frac{\partial H(z)}{\partial A} \right\|_2^2 + \left\| \frac{\partial H(z)}{\partial b} \right\|_2^2 + \left\| \frac{\partial H(z)}{\partial c} \right\|_2^2 \]  (4)

where $\| \cdot \|_2$ denotes the $L_2$-norm of function ($\cdot$). Two formulations of the $L_2$-sensitivity are proposed in [3] and [4]. In [3], the $L_2$-sensitivity is formulated by using a complex integral [3]. On the other hand, in [4], the $L_2$-sensitivity is formulated, in the following expression:

\[ S(A, b, c) = \text{tr}(W_0) + \text{tr}(K_0) + \text{tr}(W_0) + \text{tr}(K_0) + 2 \sum_{i=1}^{\infty} \text{tr}(W_i) \]  (5)

where $K_1$ is a generalized controllability gramian and $W_i$ is a generalized observability gramian [4]. From now on, we adopt the $L_2$-sensitivity of Eq. (5).

The generalized controllability gramian $K_i$ and the generalized observability gramian $W_i$ are defined as the solutions of the following Lyapunov equations:

\[ K_i = AK_iA^T + \frac{1}{2} \left( A^T bb^T + bb^T A^T \right) \]  (6)
\[ W_i = A^T W_i A + \frac{1}{2} \left( c^T cA^T + (A^T)^i c^T c \right) \]  (7)

In particular, $K_0$ and $W_0$ are defined as the solutions of the following Lyapunov equations:

\[ K_0 = AK_0 A^T + bb^T \]  (8)
\[ W_0 = A^T W_0 A + c^T c. \]  (9)

$K_0$ and $W_0$ are called the controllability gramian and the observability gramian of the filter $(A, b, c, d)$. In the field of digital signal processing, the controllability gramian is also called the covariance matrix and the observability gramian is
also called the noise matrix of the filter \((A, b, c, d)\) [1], [2], [5], [6].

C. Coordinate Transformation

Consider a coordinate transformation defined by \(\pi(u) = T^{-1}u\). Under the coordinate transformation, the coefficient matrices are transformed as

\[
(\bar{A}, \bar{b}, \bar{c}, \bar{d}) = (T^{-1}AT, T^{-1}b, cT, d)
\]

and the generalized gramians are transformed as

\[
(K_i, W_i) = (T^{-1}K_iT^{-T}, T^{-T}W_iT).
\]

Substituting Eq. (11) into Eq. (5), we have the \(L_2\)-sensitivity of a filter \((T^{-1}AT, T^{-1}b, cT, d)\) as follows:

\[
S(P) = \text{tr}(W_0P)\text{tr}(K_0P^{-1}) + \text{tr}(W_0P) + \text{tr}(K_0P^{-1}) + 2\sum_{i=1}^{\infty} \text{tr}(W_iP)\text{tr}(K_iP^{-1})
\]

where \(P\) is a positive definite symmetric matrix defined by \(P = TT^T\).

The \(L_2\)-sensitivity minimization problem is to find a similarity transformation \(T\) so that \(S(T^{-1}AT, T^{-1}b, cT)\) is minimized. The \(L_2\)-sensitivity of a filter \((T^{-1}AT, T^{-1}b, cT, d)\) is a function of \(P = TT^T\), as shown in Eq. (12). Therefore, we will derive the positive definite matrix \(P\) instead of matrix \(T\) as a solution to the \(L_2\)-sensitivity minimization problem. The \(L_2\)-sensitivity minimization problem is formulated as follows:

\[
\begin{align*}
\text{For a given digital filter } (A, b, c, d), \\
\text{minimize the } L_2\text{-sensitivity } S(P) \text{ w.r.t. } P \\
\text{where } P \text{ is an arbitrary positive definite matrix.}
\end{align*}
\]

It is proved in [7] that \(L_2\)-sensitivity \(S(P)\) has the unique global minimum, which is achieved by \(P_o\) satisfying

\[
\left. \frac{\partial S(P)}{\partial P} \right|_{P = P_o} = 0.
\]

It is difficult to solve the above equation with respect to matrix \(P\) analytically because Eq. (14) is a nonlinear equation of matrix \(P\). To this problem, the approaches which derive the optimal solution with iterative calculations are proposed in [3] and [4].

III. \(L_2\)-Sensitivity Minimization in Closed Form

A. Novel Expressions of the Generalized Gramians

It seems to be necessary to solve Lyapunov equations given by Eqs. (6) and (7) when we calculate the generalized gramians \(K_i\) and \(W_i\). However, after simple algebraic manipulations, we can derive a novel expression of generalized gramians as follows:

\[
K_i = \frac{1}{2} (A^iK_0 + K_0(A^i)^T)
\]

\[
W_i = \frac{1}{2} (W_0A^i + (A^i)^TW_0).
\]

Thus, we calculate the generalized gramians \(K_i\) and \(W_i\) using the above simple formulas rather than solving Lyapunov equations.

B. Balanced Realization as an initial realization

We adopt a balanced realization as an initial realization to synthesize the \(L_2\)-sensitivity minimization problem. The balanced realization is the filter structure of which controllability and observability gramians are diagonal and equal as follows:

\[
K_0 = W_0 = \Theta, \quad \Theta = \text{diag}(\theta_1, \cdots, \theta_N)
\]

where the parameters \(\theta_i (i = 1, \cdots, N)\) are the second order modes of the filter \(H(z)\) [1], [2]. Substituting Eq. (17) into Eqs. (15) and (16), we can express the generalized gramians of the balanced realization \((A_o, b_o, c_o, d_o)\) as

\[
K_i = \frac{1}{2} \left( A_o^i\Theta + \Theta(A_o^i)^T \right)
\]

\[
W_i = \frac{1}{2} \left( \Theta A_o^i + (A_o^i)^T\Theta \right).
\]

In addition, the balanced realization has the following symmetric properties for the coefficient matrices

\[
(A_o^i, b_o^T, c_o^T) = (\Sigma A_o\Sigma, c_o\Sigma, \Sigma b_o)
\]

and for the generalized gramians

\[
(K_i, W_i) = (\Sigma W_i\Sigma, \Sigma K_i\Sigma)
\]

where \(\Sigma\) is the signature matrix defined by

\[
\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_N), \quad \sigma_i = \pm 1 (i = 1, \cdots, N).
\]

It is proved in [3] that

\[
\left. \frac{\partial S(P)}{\partial P} \right|_{P = P_o} = 0 \quad \left. \frac{\partial S(P)}{\partial P} \right|_{P = \Sigma P_o^{-1}\Sigma} = 0.
\]

Proposition (23) and the uniqueness of the global minimum of \(S(P)\) proved in [7] can conclude that the transformation matrix \(P\) which gives the minimum \(L_2\)-sensitivity structure satisfies the following symmetric property:

\[
P_o = \Sigma P_o^{-1}\Sigma
\]

where \(\Sigma\) is a signature matrix defined in Eq. (22). We will thus search the optimal solution \(P_o\) among the positive definite matrices \(P\) which satisfy

\[
P = \Sigma P^{-1}\Sigma
\]

Substituting Eqs. (18), (19) and (25) into Eq. (12), we can simplify the \(L_2\)-sensitivity \(S(P)\) as follows:

\[
S(P) = \text{tr}(\Theta P) (2 - \text{tr}(\Theta P)) + 2\sum_{i=0}^{\infty} (\text{tr}(\Theta A_o^i P))^2.
\]

C. \(L_2\)-Sensitivity Minimization for 2-nd order digital filters

In this chapter, we restrict ourselves to the 2-nd order case of state-space digital filters, and propose a new approach to the \(L_2\)-sensitivity minimization problem.
The form of the signature matrix $\Sigma$ is classified into the following two cases:

\[
\begin{align*}
(i) \Sigma &= \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \pm I \\
(ii) \Sigma &= \pm \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.
\end{align*}
\]

(27)

The method for deriving the optimal solution depends on the above case of the signature matrix $\Sigma$.

In the case of (i), Eq. (25) follows $P = P^{-1}$, that is, $P = I$. It means that $T = U$, that is, an orthogonal transformation. In this case, the minimum $L_2$-sensitivity structure is equal to a rotated balanced realization since the orthogonal transformation is just a rotation of coordinate axes which does not change the value of the $L_2$-sensitivity. Thus, we need no more discussion in this case since the minimum $L_2$-sensitivity structure is already achieved.

In the case of (ii), Eq. (25) can be written as

\[
\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \frac{1}{p_{11}p_{22} - p_{12}p_{21}} \begin{bmatrix} p_{22} & p_{12} \\ p_{21} & p_{11} \end{bmatrix}
\]

(28)

where $p_{ij}$ is the $(i, j)$-element of matrix $P$. From Eq. (28), we have

\[
p_{11}p_{22} - p_{12}p_{21} = 1
\]

(29)

\[
p_{11} = p_{22}.
\]

(30)

Moreover, since $P$ is a symmetric matrix, it satisfies

\[
p_{12} = p_{21}.
\]

(31)

From Eq. (29), (30) and (31), we can easily derive

\[
p_{11}^2 - p_{12}^2 = 1.
\]

(32)

Now we can express $p_{ij}$ as

\[
p_{11} = p_{22} = \pm \cosh(\alpha)
\]

(33)

\[
p_{12} = p_{21} = \pm \sinh(\alpha)
\]

(34)

recalling the following formula for the hyperbolic functions

\[
\cosh^2(\alpha) - \sinh^2(\alpha) = 1
\]

(35)

for real $\alpha$. Since $P$ is a positive definite matrix, $p_{11}$ and $p_{22}$ must be positive. Thus, Eqs. (33) and (34) are rewritten as follows:

\[
p_{11} = p_{22} = \cosh(\alpha)
\]

(36)

\[
p_{12} = p_{21} = \sinh(\alpha)
\]

(37)

without loss of generality. We newly propose the general form of a positive definite matrix $P$ which satisfies Eq. (25) as follows:

\[
P = \begin{bmatrix} \cosh(\alpha) & \sinh(\alpha) \\ \sinh(\alpha) & \cosh(\alpha) \end{bmatrix}.
\]

(38)

For the balanced realization $(A_b, b_b, c_b, d_b)$, the coefficient matrix $A_b$ can be decomposed as

\[
A_b = V \Lambda V^{-1}
\]

(39)

in order to simplify the expression of $S(P)$. In the above decomposition, $V$ is a nonsingular matrix which diagonalizes the matrix $A_b$. $\Lambda$ is a matrix of which diagonal elements are equal to eigenvalues of the matrix $A_b$ (equivalently poles of the filter). Eq. (39) means diagonalization of $A_b$ or transformation of $A_b$ to the Jordan form. In this paper, we consider the case that the two poles $(\lambda, \lambda^*)$ of the filter $(A_b, b_b, c_b, d_b)$ are conjugate, but the other cases are omitted due to the restriction of the manuscript. The decomposition of $2 \times 2$ coefficient matrix $A_b$ shown in Eq. (39) can be represented as

\[
A_b = \begin{bmatrix} 1 & v \\ v & 1 \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \lambda^* \end{bmatrix} \begin{bmatrix} 1 & v \\ v & 1 \end{bmatrix}^{-1}
\]

(40)

The $L_2$-sensitivity $S(P)$ can be converted into a function of the single scalar variable $S(\alpha)$ by substituting Eqs. (38), (40) and $\Theta = \text{diag}(\theta_1, \theta_2)$ into Eq. (26). Consequently, $L_2$-sensitivity $S(\alpha)$ is expressed as Eq. (41), which can be simplified as

\[
S(P) = S(\alpha) = \sum_{n=-2}^{2} C_n(\Theta, \Lambda, V)e^{n\alpha}
\]

(42)

which does not contain infinite summations. These coefficients $C_n(\Theta, \Lambda, V)$ are easily computed directly from the coefficient matrices $(A_b, b_b, c_b, d_b)$. The parameter $\alpha$ which minimizes $S(\alpha)$ can be derived by solving the following equation with respect to $\alpha$:

\[
\frac{\partial S(\alpha)}{\partial \alpha} = \sum_{n=-2}^{2} nC_n(\Theta, \Lambda, V)e^{n\alpha} = 0.
\]

(43)

Letting $\beta = e^\alpha$ gives

\[
\sum_{n=-2}^{2} nC_n(\Theta, \Lambda, V)\beta^n = 0
\]

(44)

which is a 4-th degree polynomial equation with respect to $\beta$ of constant coefficients. The above equation can be solved

\[
S(\alpha) = 2(\theta_1 + \theta_2) \cosh(\alpha) - (\theta_1 + \theta_2)^2 \cosh^2(\alpha) + \left( \frac{\theta_2}{1 - \lambda^2} + \frac{2\theta_1 \theta_2}{1 - |\lambda|^2} + \frac{\theta_1^2}{1 - (\lambda^*)^2} \right) \cosh^2(\alpha)
\]

\[
+ 2(\theta_1 - \theta_2) \left( \frac{\theta_2}{1 - \lambda^2} - \frac{\theta_1}{1 - (\lambda^*)^2} \right) \cosh^2(\alpha) - v \sinh(\alpha) \cosh(\alpha)
\]

\[
+ (\theta_1 - \theta_2)^2 \left( \frac{2}{1 - |\lambda|^2} - \frac{2}{1 - (\lambda^*)^2} \right) \cosh^2(\alpha) - 2v \sinh(\alpha) \cosh(\alpha) + v^2 \sinh^2(\alpha)
\]

(41)
analytically because there exists the formula of solutions for 4-th degree polynomial equations. Eq. (44) has four solutions, in which the positive real solution $\beta_o$ is adopted to derive the optimal solution as

$$
\begin{align*}
P_o &= \begin{bmatrix}
\cosh(\log(\beta_o)) & \sinh(\log(\beta_o)) \\
\sinh(\log(\beta_o)) & \cosh(\log(\beta_o))
\end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix}
\beta_o + \beta_o^{-1} & \beta_o - \beta_o^{-1} \\
\beta_o - \beta_o^{-1} & \beta_o + \beta_o^{-1}
\end{bmatrix}.
\end{align*}
\tag{45}
$$

Considering the decomposition of the optimal solution $P_o = T_o^*T_o$ as

$$
T_o = P_o^{\frac{1}{2}} = \frac{1}{2} \begin{bmatrix}
\beta_o^{1/2} + \beta_o^{-1/2} & \beta_o^{1/2} - \beta_o^{-1/2} \\
\beta_o^{-1/2} - \beta_o^{1/2} & \beta_o^{1/2} + \beta_o^{-1/2}
\end{bmatrix},
\tag{46}
$$

we finally obtain the optimal coordinate transformation matrix $T_o$:

$$
T_o = P_o^{\frac{1}{2}} = \frac{1}{2} \begin{bmatrix}
\beta_o^{1/2} + \beta_o^{-1/2} & \beta_o^{1/2} - \beta_o^{-1/2} \\
\beta_o^{-1/2} - \beta_o^{1/2} & \beta_o^{1/2} + \beta_o^{-1/2}
\end{bmatrix}. \tag{47}
$$

IV. A Numerical Example

We present a numerical example to demonstrate the effectiveness of the proposed method. Consider a 2-nd order digital filter $H(z)$ given by

$$
H(z) = \frac{0.20657 + 0.41314z^{-1} + 0.20657z^{-2}}{1 - 0.36953z^{-1} + 0.19582z^{-2}}
\tag{48}
$$

which has conjugate poles. The balanced realization, as an initial realization, of the digital filter $H(z)$ is derived as follows:

$$
\begin{bmatrix}
A_o & b_o \\
C_o & d_o
\end{bmatrix} = \begin{bmatrix}
0.41691 & 0.46430 & -0.72446 \\
-0.46430 & 0.04738 & -0.88805 \\
-0.72446 & 0.18805 & 0.20657
\end{bmatrix}
\tag{49}
$$

which satisfy the symmetric property in Eq. (20) where $\Sigma = \text{diag}(1, -1)$. The controllability and observability gramians of the filter $(A_o, b_o, c_o, d_o)$ are derived as

$$
K_0 = W_0 = \text{diag}(0.68301, 0.18301). \tag{50}
$$

Using the coefficient matrices $(A_o, b_o, c_o, d_o)$ and the gramians $(K_0, W_0)$, the coefficients $C_o$’s in Eq. (42) can be computed as follows:

$$(C_{-2}, C_{-1}, C_0, C_1, C_2) = (0.20249, 0.86603, 0.40047, 0.86603, 0.32189). \tag{51}$$

We solve the 4-th degree polynomial equation (44) and obtain the following four solutions:

$$
\beta = 0.93919, -0.57760, -0.85341 \pm j0.65677. \tag{52}
$$

We adopt $\beta_o = 0.93919$, which is the positive real number, to derive the optimal solution. Substituting $\beta_o = 0.93919$ into Eq. (45) gives the optimal solution $P_o$, and substituting $\beta_o = 0.93919$ into Eq. (47) gives the optimal coordinate transformation matrix $T_o$:

$$
T_o = \begin{bmatrix}
1.00049 & -0.03137 \\
-0.03137 & 1.00049
\end{bmatrix}. \tag{53}
$$

A minimum $L_2$-sensitivity realization

$$(A_o, b_o, c_o, d_o) = (T_o^{-1}AT_o, T_o^{-1}b_o, cT_o, d) \tag{54}$$

is given by

$$
\begin{bmatrix}
A_o & b_o \\
cT_o & d
\end{bmatrix} = \begin{bmatrix}
0.38822 & 0.45064 & -0.73072 \\
-0.45064 & -0.01869 & -0.21088 \\
-0.73072 & 0.21088 & 0.20657
\end{bmatrix}
\tag{55}
$$

of which $L_2$-sensitivity is

$$
S_{\text{min}} = 2.649423679. \tag{56}
$$

Table I shows the comparison of our proposed method with the iterative methods reported in [3], [4]. Our proposed method achieves the minimum $L_2$-sensitivity by only solving a 4-th degree polynomial equation without iterative calculations, but the method in [3] and the method in [4] require many iterative calculations to achieve the minimum $L_2$-sensitivity.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.6509011178</td>
<td>2.6509011178</td>
</tr>
<tr>
<td>1</td>
<td>2.650887687</td>
<td>2.649454402</td>
</tr>
<tr>
<td>2</td>
<td>2.649711582</td>
<td>2.649423803</td>
</tr>
<tr>
<td>3</td>
<td>2.649804110</td>
<td>2.649423680</td>
</tr>
<tr>
<td>4</td>
<td>2.649434868</td>
<td>2.649423679</td>
</tr>
<tr>
<td>5</td>
<td>2.649425887</td>
<td>2.649423679</td>
</tr>
<tr>
<td>6</td>
<td>2.649424115</td>
<td>2.649423679</td>
</tr>
<tr>
<td>7</td>
<td>2.649423765</td>
<td>2.649423679</td>
</tr>
<tr>
<td>8</td>
<td>2.649423666</td>
<td>2.649423679</td>
</tr>
<tr>
<td>9</td>
<td>2.649423683</td>
<td>2.649423679</td>
</tr>
</tbody>
</table>

V. Conclusion

This paper has discussed minimization of $L_2$-sensitivity in case of 2-nd order digital filters. We have shown that the optimal solution to the minimization of $L_2$-sensitivity can be described using hyperbolic functions, and can be derived by only solving a 4-th degree polynomial equation of constant-coefficients. A numerical example has been given to illustrate the effectiveness of our proposed method.

REFERENCES


