# ON THE OPTIMAL STRUCTURE of 2-D DIGITAL FILTERS WITH $L_2$-SENSITIVITY MINIMIZATION

*Gang LI*

School of EEE
Nanyang Technological University
Singapore 639798 e-mail: egli@ntu.edu.sg

## ABSTRACT

An expression is derived for the error variance of transfer function of a Two Dimensional system due to FWL errors. The optimal realization problem is then formulated by minimizing this variance with respect to all possible realizations of the system. This problem is shown to be equivalent to the minimization of a pure $L_2$ norm based sensitivity measure $M_{L_2}$ and can be sloved using any standard minimization algorithm with guaranteed convergence.

## 1 INTRODUCTION

The finite word length (FWL) effects have been considered as one of the most serious problems in the actual implementation of a digital system. Due to the FWL errors on the parameters, the actually implemented transfer function may be very different from the desired one. This leads to a class of sensitivity studies for different sensitivity measures such as transfer function sensitivity (see, e.g., [1-2]). Many classical and recent developments on this issue for one-dimensional (1-D) systems can be found in [3]. Traditionally, the transfer function sensitivity measure was defined with a mixture of $L_1/L_2$ norm. The corresponding results were extended to 2-D case by many researchers (see, e.g., [4-6]). Recently, a pure $L_2$ based transfer function sensitivity measure was studied and some properties of this measure were revealed in [3, 7]. The main objective of this paper is to extend the $L_2$ sensitivity minimization problem from 1-D to 2-D.

It should be pointed out that in the traditionally used $L_1/L_2$ sensitivity approach, the sensitivity measure is replaced by an upper bound. For 1-D case, the optimal realizations that minimize this upper bound are exactly the same as those minimizing the $L_1/L_2$ sensitivity measure itself. This is not true for 2-D case, where these two optimal realizations can be very different and hence the upper bound optimal realizations may not yield the performance as expected as in 1-D case. It is true that the solution to the pure $L_2$ minimization problem does not have a closed form and requires more computation. This is the main drawback of using this pure $L_2$ sensitivity measure. The computational complexity is, however, not of concern here since this is in the design stage.

## 2 PROBLEM FORMULATION

In this paper, we consider a 2-D discrete linear time-invariant Single Input Single Output system (SISO) $H(z_h, z_v)$ of order $(n_h, n_v)$. This system can be represented with the following Roesser state-space equations [8]:

$$\begin{aligned}\begin{bmatrix} x^h_{i+1,j} \\ x^v_{i,j+1} \end{bmatrix} &= \begin{bmatrix} A_h & A_{hv} \\ A_{vh} & A_v \end{bmatrix} \begin{bmatrix} x^h_{i,j} \\ x^v_{i,j} \end{bmatrix} + \begin{bmatrix} B_h \\ B_v \end{bmatrix} u_{i,j} \\ &\triangleq Ax_{i,j} + Bu_{i,j} \\ y_{i,j} &= \begin{bmatrix} C_h & C_v \end{bmatrix} \begin{bmatrix} x^h_{i,j} \\ x^v_{i,j} \end{bmatrix} + du_{i,j} \\ &\triangleq Cx_{i,j} + du_{i,j}, \end{aligned} \tag{1}$$

where $x^h \in R^{n_h \times 1}$ and $x^v \in R^{n_v \times 1}$ are called horizontally and vertically propagating local state vector, re-

spectively, and $A_x \in R^{n_x \times n_y}$, $A_{xy} \in R^{n_x \times n_y}$, $B_x, C_x^{\mathcal{T}} \in R^{n_x \times 1}$ for $x, y = h, v$, and $d \in R$.

$(A, B, C, d)$ is called a realization of the 2-D system $H(z_h, z_v)$, satisfying $H(z_h, z_v) = d + C(z_h I_{n_h} \oplus z_v I_{n_v} - A)^{-1} B$, where $\oplus$ denotes the direct sum of matrices. Denote $S_H$ as the set of all the realizations of $H(z_h, z_v)$. It is well known that $S_H$ is an infinite set and that if $(A_0, B_0, C_0, d) \in S_H$, $S_H$ can be characterized by

$$A = T^{-1} A_0 T \quad B = T^{-1} B_0 \quad C = C_0 T \tag{2}$$

with $T = T_h \oplus T_v$, where $T_h \in R^{n_h \times n_h}$ and $T_v \in R^{n_v \times n_v}$ are any non-singular (transformation) matrix.

Let $\{p_i\}$ and $\{p_i^*\}$ be the set of the ideal parameters and those actually implemented with FWL of the same realization $R$, respectively, and assume that this realization has $N$ parameters. Denote $\{\Delta p_i \triangleq p_i - p_i^*\}$ as the corresponding parameter perturbations. With a first order approximation, one has

$$\begin{aligned} \Delta H(z_h, z_v) &\triangleq H(z_h, z_v) - H^*(z_h, z_v) \\ &= \sum_{i=1}^{N} \frac{\partial H(z_h, z_v)}{\partial p_i} \Delta p_i. \end{aligned} \tag{3}$$

We adopt a statistical approach where the perturbations of the parameters are considered as independent random variables uniformly distributed within the range $[-\frac{1}{2}2^{-B_c}, \frac{1}{2}2^{-B_c}]$ for a fixed-point implementation of $B_c$ bits (see, e.g., [9]). We now define the transfer function error measure as follows:

$$\sigma_H^2 \triangleq \frac{1}{(2\pi)^2} \int \int_{-\pi}^{\pi} E[|\Delta H(e^{j\omega_h}, e^{j\omega_v})|^2] d\omega_h d\omega_v, \tag{4}$$

where $E(.)$ denotes the ensemble average operation.

Keeping the assumption in mind that $\{\Delta p_i\}$ is an independent and uniformly distributed random variable set with $\sigma_c^2 = E[(\Delta p_i)^2] = \frac{1}{12} 2^{-2B_c}$, one can show that

$$\sigma_H^2 = (\sum_{X=\{A,B,C,d\}} ||S_X(z_h, z_v)||_2^2) \sigma_c^2 \triangleq M_{L_2} \sigma_c^2, \tag{5}$$

where $S_X(z_h, z_v) \triangleq \frac{\partial H(z_h, z_v)}{\partial X}$, called the sensitivity function of $H(z_h, z_v)$ with respect to matrix $X = \{x_{kl}\}$, is a matrix of the same dimension as $X$ with its $(k, l)$th element given by $\frac{\partial H}{\partial x_{kl}}$ and $M_{L_2}$ is called the $L_2$-sensitivity measure:

$$\begin{aligned} M_{L_2} &\triangleq \sum_{X=\{A,B,C,d\}} ||S_X(z_h, z_v)||_2^2 \\ &= tr(Q_A + Q_B + Q_C) + Q_d \end{aligned} \tag{6}$$

with

$$Q_X = \frac{1}{(2\pi j)^2} \oint_{\Gamma} \left(\frac{\partial H)}{\partial X}\right) \left(\frac{\partial H}{\partial X}\right)^{\mathcal{H}} z_h^{-1} z_v^{-1} dz_h dz_v \tag{7}$$

for $X = A, B, C^{\mathcal{T}}, d$, where $\mathcal{T}$ denotes the transpose operation.

It is easy to show that $S_d(z_h, z_v) = 1$ and

$$\begin{aligned} S_A(z_h, z_v) &= G(z_h, z_v) F^{\mathcal{T}}(z_h, z_v) \\ S_B(z_h, z_v) &= G(z_h, z_v), \; S_C(z_h, z_v) = F^{\mathcal{T}}(z_h, z_v), \end{aligned} \tag{8}$$

where

$$\begin{aligned} F(z_h, z_v) &\triangleq (z_h I_{n_h} \oplus z_v I_{n_v} - A)^{-1} B \\ G(z_h, z_v) &\triangleq (z_h I_{n_h} \oplus z_v I_{n_v} - A^{\mathcal{T}})^{-1} C^{\mathcal{T}}. \end{aligned} \tag{9}$$

$Q_X$ defined by (7) are usually called gramian. Now, let see how to compute $Q_X$ for $X = A, B, C^{\mathcal{T}}$. First of all, it follows from (7) that

$$\begin{aligned} Q_B &= \frac{1}{(2\pi j)^2} \oint_{\Gamma} G G^{\mathcal{H}} z_h^{-1} z_v^{-1} dz_h dz_v \triangleq W \\ Q_C &= \frac{1}{(2\pi j)^2} \oint_{\Gamma} F F^{\mathcal{H}} z_h^{-1} z_v^{-1} dz_h dz_v \triangleq K, \end{aligned} \tag{10}$$

where $K$ and $W$ are called the 2-D controllability and observability gramian. In the sequel, it is assumed that the realization $(A, B, C, d)$ is locally *reachable* and *observable*. Therefore, the gramians $K$ and $W$ are always positive-definite.

It can be shown (see the full version of the paper) that $Q_A$ can be computed as a gramian of a 2-D system of higher order.

As far as we know, there are two commonly used algorithms to compute 2-D gramians. The first one was given by Premaratne *et al* in [10], and the second method was proposed by Lu *et al* in [11] where based on the Åström-Jury-Agniel algorithm, an method was developed, which can *accurately* evaluate the 2-D gramians.

It has been noted that the algorithm by Lu *et al* is very efficient only for 2-D gramians of relatively low dimension. In this algorithm, polynomial convolutions and factorizations are involved. The order of the polynomials increases greatly with the order of system. Therefore, its efficiency degrades and undesired numerical problems may occur for high order systems. Combining the two algorithms, we propose a new algorithm for efficiently computing 2-D gramians. For details, we refer to the full version of the paper.

Let $R_0$ be a realization in $S_H$ and $R$ be obtained by transforming $R_0$ with $T$, and $F_0(z_h, z_v)$ and $G_0(z_h, z_v)$ be given by (9) for the realization $R_0$, then one has the following for $R$

$$\begin{aligned} F(z_h, z_v) &= T^{-1} F_0(z_h, z_h) \\ G(z_h, z_v) &= T^{\mathcal{T}} G_0(z_h, z_v) \end{aligned} \tag{11}$$

from which one can see that different realizations have different $L_2$ sensitivity measure. Therefore, it is interesting to find out those realizations that minimize $\sigma_H^2$ over $S_H$, which is equivalent to

$$\min_{(A,B,C,d) \in S_H} M_{L_2}. \tag{12}$$

We note that $Q_d = 1$. Therefore, its value does not affect the optimal realization problem defined above. In the sequel, we consider $Q_d = 0$ (even it should not be), that is $M_{L_2} = tr(Q_A + Q_B + Q_C)$.

# 3 OPTIMAL REALIZATIONS

The main objective of this section is to solve the optimal realization problem, that is to find out the solutions to (12).

Denote

$$P \triangleq TT^{\mathcal{T}} = \begin{pmatrix} T_h & 0 \\ 0 & T_v \end{pmatrix} \begin{pmatrix} T_h & 0 \\ 0 & T_v \end{pmatrix}^{\mathcal{T}} = \begin{pmatrix} T_h T_h^{\mathcal{T}} & 0 \\ 0 & T_v T_v^{\mathcal{T}} \end{pmatrix}. \tag{13}$$

one can see that $M_{L_2}$ is a function of $P$, that is $M_{L_2} = f(P)$. Therefore, (12) is equivalent to

$$\min_{P>0: subject\ to (13)} M_{L_2}. \tag{14}$$

Our main result in this section is summarized as below:

**Theorem 1** : *For a stable 2-D system $H(z_h, z_v)$, the minimum of $f(P)$ with $P$ defined in (13) exists and can be achieved* only *by the* unique *stationary point $P > 0$ of $f(P)$.*

Since $f(P)$ is unimodal, any local minimum is a global minimum. We, therfore, argue that (14) can efficiently be solved using any standard minimization algorithm such as golden section search and *Gauss-Newton* methods with guaranteed convergence to the global minimum.

The proof of the above theorem is very long and is omitted. For details, we refere to the long version of the paper.

In the next section, we will illustrate the optimal realization procedure with an example. The unique solution to (14) $P_{opt} = P_h^{opt} \oplus P_v^{opt}$ is computed with the following *classical* gradient based algorithm:

$$P_{k+1} = P_k - \mu_k \begin{pmatrix} \frac{df(P)}{dP_h} & 0 \\ 0 & \frac{df(P)}{dP_v} \end{pmatrix} |_{P=P_k}, \tag{15}$$

where $\frac{df(P)}{dP_x}$ for $x = h, v$, and $\mu_k$ is a small time-variant positive step size. Clearly, this algorithm always converges to $P_{opt}$.

# 4 NUMERICAL EXAMPLE

We now illustrate our optimal design procedures with a stable digital filter of order $(n_h, n_v) = (2, 2)$, which was used in [6]. This filter is given by the following realization $R_0$:

$$A_0 = \begin{pmatrix} 1.888990 & -0.912190 & -1 & 0 \\ 1 & 0 & 0 & 0 \\ 0.027710 & -0.025800 & 1.888990 & 1 \\ -0.025800 & 0.024310 & -0.912190 & 0 \end{pmatrix}$$

$$B_0 = \begin{pmatrix} 0.219089 \\ 0 \\ -0.028889 \\ 0.091219 \end{pmatrix}, C_0^{\mathcal{T}} = \begin{pmatrix} 0.288890 \\ -0.091219 \\ -0.219089 \\ 0 \end{pmatrix}.$$

The 2-D balanced realization $R_b$ can be obtained with the initial realization $R_0$. This $R_b$ was shown to be $\bar{M}_{L_1/L_2}$ optimal in [5] and is given below:

$$A_b = \begin{pmatrix} 0.9664 & 0.1279 & -0.4909 & -0.1945 \\ -0.1611 & 0.9226 & -0.1823 & -0.0723 \\ 0.0463 & 0.0088 & 0.9778 & -0.1747 \\ 0.0105 & 0.0187 & 0.1215 & 0.9112 \end{pmatrix}$$

$$B_b = \begin{pmatrix} 0.2678 \\ 0.0995 \\ 0.2252 \\ -0.7498 \end{pmatrix}, C_b^{\mathcal{T}} = \begin{pmatrix} 0.4881 \\ -0.6778 \\ -0.0880 \\ -0.0349 \end{pmatrix}.$$

Denote $\eta(P) \triangleq max_{x,(i,j)} |\frac{df(P)}{dP_x(i,j)}|$ for $x = h, v$. We now take $R_b$ as the initial realization. Starting with

$P_0 = I$ as the initial condition, we run algorithm (15) with $\mu_k = 10^{-3}\eta^{-1}(P_k)$. Computation shows $\eta(P_0) = 4.9657 \times 10^3$ and $M_{L_2}(R_b) = 9.9349 \times 10^3$. It is found $\eta(P_{310}) = 1.6614$ and $M_{L_2}(R_{310}) = 2.9517 \times 10^3$. From $k = 310$, $\mu_k = 10^{-4}\eta^{-1}(P_k)$. It is noted that $\eta(P_{540}) = 5.5047 \times 10^{-3}$ and $M_{L_2}(R_{540}) = 2.9515 \times 10^3$. We consider $P_{opt} = P_{540}$ and then $T_{opt} = P_{opt}^{1/2}V$, where $V$ is an arbitrary orthogonal matrix. The corresponding optimal realization $R_{opt}$ can be obtained with $T_{opt}$ and $R_b$. The following optimal realization corresponds to the choice $V = I$:

$$A_{opt} = \begin{pmatrix} 0.9658 & 0.1423 & -0.1576 & 0.0560 \\ -0.1446 & 0.9232 & -0.0385 & 0.0137 \\ 0.1445 & 0.0617 & 0.9449 & -0.1848 \\ -0.0567 & 0.0445 & 0.1089 & 0.9441 \end{pmatrix}$$

$$B_{opt} = \begin{pmatrix} 0.1562 \\ 0.0382 \\ -0.6400 \\ -1.4340 \end{pmatrix}, C_{opt}^T = \begin{pmatrix} 0.7517 \\ -1.4187 \\ -0.0484 \\ 0.0172 \end{pmatrix}.$$

We have computed $\bar{M}_{L_1/L_2}$ and $M_{L_2}$ for $R_0$, $R_b$ and $R_{opt}$, respectively. The results are given in the following table:

| Realiz. | $R_0$ | $R_b$ | $R_{opt}$ |
|---|---|---|---|
| $M_{L_2}$ | $1.9497 \times 10^7$ | $9.9349 \times 10^3$ | $2.9515 \times 10^3$ |
| $M_{L_1/L_2}$ | $2.3097 \times 10^5$ | $5.0162 \times 10^2$ | $1.7229 \times 10^3$ |

One can see that $R_0$ has a much higher $L_2$-sensitivity measure than the two others. The balanced realization $R_b$ has a relative smaller sensitivity measure, which is about 3.5 times of that for $R_{opt}$. The difference between $R^{opt}(M_{L_2})$ and $R^{opt}(\bar{M}_{L_1/L_2})$ depends on the system.

## REFERENCES


[1] L. Thiele, "On the Sensitivity of Linear State-Space Systems," *IEEE Trans. on Circuits and Systems*, vol. CAS-33, No. 5, May, pp. 502-510, 1986.

[2] G. Li, B. D. O. Anderson, M. Gevers and J. Perkins, " Optimal FWL Design of State-Space Digital Systems with Weighted Sensitivity Minimization and Sparsness Consideration," *IEEE Trans. on Circuits and Systems*, vol.39, No. 5, pp. 365-377, May 1992.

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

[4] A. Zilouchian and R. L. Carroll, "A Coefficient Sensitivity Bound in 2-D State Space Digital Filtering," *IEEE Trans. on Circuits and Systems*, vol. CAS-33, pp. 665-667, No. 7, July, 1986.

[5] T. Lin, M. Kawamata and T. Higuchi, "Minimization of Sensitivity of 2-D Systems and Its Relation to 2-D Balanced Realizations," *The Trans. of The Institute of Electronics, Information and Communication Engineers*, E 70, pp. 938-944, 1987.

[6] T. Hinamoto and T. Takao, "Synthesis of 2-D State Space Filter Structures with Low Frequency-Weighted Sensitivity," *IEEE Trans. on Circuits and Systems*, vol. CAS-39, pp. 646-651, No. 9, September, 1992.

[7] W. Y. Yan and J. B. Moore, "On $L^2$-Sensitivity Minimization of Linear State-Space Systems," *IEEE Trans. on Circuits and Systems*, vol. 39, pp. 641-648, Aug. 1992.

[8] R. P. Roesser, "A Discrete State-Space Model for Linear Image Processing," *IEEE Trans. on Automatic Control*, vol. AC-20, pp. 1-10, Feb. 1975.

[9] R. E. Crochiere, "A New Statistical Approach to The Coefficient Word Length Problem for Digital Filters," *IEEE Trans. on Circuits and Systems*, vol. CAS-22, pp. 190-196, March 1975.

[10] K. Premaratne, E.I. Jury and M. Mansour, "An Algorithm for Model Reduction of 2-D Discrete Time systems," *IEEE Trans. on Circuits and Systems*, vol. 37, No. 9, pp. 1116-1131.

[11] W. S. Lu, H.P. Wang and A. Antoniou, "An Efficient Method for the Evaluation of the Controllability and Observability Gramians of 2-D Digital Filters and Systems," *IEEE Trans. on Circuits and Systems - II*, vol. 39, No. 10, pp. 695-704, Oct. 1992.