On the Transfer Function Error of State-Space Filters in Fixed-Point Context

Thibault Hilaire

On the Transfer Function Error of State-Space Filters in Fixed-Point Context

Thibault Hilaire

***Abstract*—This paper presents a new measure used for the implementation of filters/controllers in state-space form. It investigates the transfer function deviation generated by the coefficient quantization. The classical L_2-sensitivity measure is extended with precise consideration on their fixed-point representation, in order to make a more valid measure. By solving the related optimal realization problem, fixed-point accurate realizations in state-space form can be found.**

***Index Terms*—Digital filter implementation, coefficient sensitivity, fixed-point implementation.**

I. Introduction

The majority of control or signal processing systems is implemented in digital general purpose processors, DSPs[1], FPGAs[2], etc. Since these devices cannot compute with infinite precision and approximate real-number parameters with a finite binary representation, the numerical implementation of controllers (filters) leads to deterioration in characteristics and performance. This has two separate origins, corresponding to the quantization of the embedded coefficients and the roundoff errors occurring during the computations. They can be formalized as parametric errors and numerical noises, respectively. The focus of this paper are parametric errors, but one can refer to [1]–[4] for roundoff noises, where measures with fixed-point consideration already exist.

It is also well known that these Finite Word Length (FWL) effects depend on the structure of the realization. In state-space form, the realization depends on the choice of the basis of the state-vector. This motivates us to investigate the coefficient sensitivity minimization problem. It has been well studied with the L_2-measure [1], [5]. However this measure only considers how sensitive to the coefficients the transfer function is, and does not investigate the coefficients' quantization, which depends on the fixed-point representation used. In [5], the transfer function error is exhibited for the first time, however, only for quantized coefficients with the same binary-point position.

This paper investigates the transfer function deviation generated by the coefficient quantization with precise consideration on their fixed-point representation. The classical L_2-sensitivity analysis is shown in section II, whereas the new approach, based on fixed-point consideration, is presented in section III. A comparison with the L_2-sensitivity and some scaling considerations are provided. Finally, the optimal realization problem is solved in section IV and a numerical example is exhibited before conclusion.

This work has been partially funded by the NFN SISE project (National Research Network "Signal and Information Processing in Science and Engineering") and the Institute of Communications and Radio-Frequency Engineering of Vienna University of Technology.France. T. Hilaire is now with the Laboratory of Computer Science (LIP6) of the University Pierre & Marie Curie of Paris, France.

[1]Digital Signal Processors

[2]Field Programmable Gate-Array

II. L_2-sensitivity analysis

Let $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$ be a stable, controllable and observable linear discrete time Single Input Single Output (SISO) state-space system, *i.e.*

$$\left\{ \begin{array}{rcl} \boldsymbol{x}(k+1) & = & \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{b}u(k) \\ y(k) & = & \boldsymbol{c}\boldsymbol{x}(k) + du(k) \end{array} \right. \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{b} \in \mathbb{R}^{n\times 1}$, $\boldsymbol{c} \in \mathbb{R}^{1\times n}$ and $d \in \mathbb{R}$. $u(k)$ is the scalar input, $y(k)$ is the scalar output and $\boldsymbol{x}(k) \in \mathbb{R}^{n\times 1}$ is the state vector.

Its input-output relationship is given by the scalar transfer function $h : \mathbb{C} \to \mathbb{C}$ defined by:

$$h : z \mapsto \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + d. \tag{2}$$

The quantization of the coefficients introduces some uncertainties to $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d leading to $\boldsymbol{A} + \boldsymbol{\Delta A}$, $\boldsymbol{b} + \boldsymbol{\Delta b}$, $\boldsymbol{c} + \boldsymbol{\Delta c}$ and $d + \Delta d$ respectively. It is common to consider the sensitivity of the transfer function with respect to the coefficients, based on the following definitions.

Definition 1 (Transfer function sensitivity) *Consider $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ and $f : \mathbb{R}^{m\times n} \to \mathbb{C}$ differentiable with respect to all the entries of $\boldsymbol{X}$.*
The sensitivity of f with respect to $\boldsymbol{X}$ is defined by the matrix $\boldsymbol{S_X} \in \mathbb{R}^{m\times n}$:

$$\frac{\partial f}{\partial \boldsymbol{X}} \triangleq \boldsymbol{S_X} \quad \textit{with} \quad (\boldsymbol{S_X})_{i,j} \triangleq \frac{\partial f}{\partial \boldsymbol{X}_{i,j}}. \tag{3}$$

Applied to a scalar transfer function h where $h(z)$ depends on a given matrix X, $\frac{\partial h}{\partial X}$ is a transfer function of a Multiple Inputs Multiple Outputs (MIMO) system.

Definition 2 (L_2-Norm) *Let $\boldsymbol{H} : \mathbb{C} \to \mathbb{C}^{k\times l}$ be a function of the scalar complex variable z. $\|\boldsymbol{H}\|_2$ is the L_2-norm of $\boldsymbol{H}$, defined by:*

$$\|\boldsymbol{H}\|_2 \triangleq \sqrt{\frac{1}{2\pi}\int_0^{2\pi} \|\boldsymbol{H}\left(e^{j\omega}\right)\|_F^2 \, d\omega} \tag{4}$$

where $\|.\|_F$ is the Froebenius norm.

Proposition 1 *If $\boldsymbol{H}$ is the MIMO state-space system $(\boldsymbol{K}, \boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$, then its L_2-norm can be computed by*

$$\|\boldsymbol{H}\|_2^2 = tr(\boldsymbol{N}\boldsymbol{N}^\top + \boldsymbol{M}\boldsymbol{W}_c\boldsymbol{M}^\top) \quad (5)$$

$$= tr(\boldsymbol{N}^\top\boldsymbol{N} + \boldsymbol{L}^\top\boldsymbol{W}_o\boldsymbol{L}) \quad (6)$$

where $\boldsymbol{W}_c$ and $\boldsymbol{W}_o$ are the controllability and observability Gramians, respectively. They are solutions of the Lyapunov equations

$$\boldsymbol{W}_c = \boldsymbol{K}\boldsymbol{W}_c\boldsymbol{K}^\top + \boldsymbol{L}\boldsymbol{L}^\top, \quad \boldsymbol{W}_o = \boldsymbol{K}^\top\boldsymbol{W}_o\boldsymbol{K} + \boldsymbol{M}^\top\boldsymbol{M}. \quad (7)$$

Proof: See [1]. ∎

Gevers and Li [1] have proposed the L_2-sensitivity measure to *evaluate* the coefficient roundoff errors. It is defined by

$$M_{L_2} \triangleq \left\|\frac{\partial h}{\partial \boldsymbol{A}}\right\|_2^2 + \left\|\frac{\partial h}{\partial \boldsymbol{b}}\right\|_2^2 + \left\|\frac{\partial h}{\partial \boldsymbol{c}}\right\|_2^2 + \left\|\frac{\partial h}{\partial d}\right\|_2^2 \quad (8)$$

and can be computed by $\frac{\partial h}{\partial \boldsymbol{A}}(z) = \boldsymbol{G}^\top(z)\boldsymbol{F}^\top(z)$, $\frac{\partial h}{\partial \boldsymbol{b}}(z) = \boldsymbol{G}^\top(z)$, $\frac{\partial h}{\partial \boldsymbol{c}}(z) = \boldsymbol{F}(z)$ and $\frac{\partial h}{\partial d}(z) = 1$, with

$$\boldsymbol{F}(z) \triangleq (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}, \qquad \boldsymbol{G}(z) \triangleq \boldsymbol{c}(zI_n - \boldsymbol{A})^{-1}. \quad (9)$$

This measure is an extension of the more tractable but less natural L_1/L_2 sensitivity measure proposed by V. Tavşanoğlu and L. Thiele [6] ($\left\|\frac{\partial h}{\partial \boldsymbol{A}}\right\|_1^2$ instead of $\left\|\frac{\partial h}{\partial \boldsymbol{A}}\right\|_2^2$ in (8)).

Remark 1 To simplify the expressions, it is also possible to regroup all the coefficients in one unique matrix $\boldsymbol{Z}$:

$$\boldsymbol{Z} \triangleq \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{c} & d \end{pmatrix}. \quad (10)$$

Then, with L_2-norm property, $M_{L_2} = \left\|\frac{\partial h}{\partial \boldsymbol{Z}}\right\|_2^2$. From (9) and the associated state-spaces, the sensitivity transfer function $\frac{\partial h}{\partial \boldsymbol{Z}}$ can be described by the MIMO state-space system $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}, \tilde{\boldsymbol{D}})$ with

$$\begin{aligned} \tilde{\boldsymbol{A}} &\triangleq \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{pmatrix}, \tilde{\boldsymbol{B}} \triangleq \begin{pmatrix} \boldsymbol{0} & \boldsymbol{b} \\ \boldsymbol{I}_n & \boldsymbol{0} \end{pmatrix}, \\ \tilde{\boldsymbol{C}} &\triangleq \begin{pmatrix} \boldsymbol{I}_n & 0 \\ \boldsymbol{0} & \boldsymbol{c} \end{pmatrix}, \tilde{\boldsymbol{D}} \triangleq \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{pmatrix}. \end{aligned} \quad (11)$$

The proposition 1 is used to compute M_{L_2}. See [1] and [7] for more details.

Applying a coordinate transformation, defined by $\bar{\boldsymbol{x}}(k) \triangleq \boldsymbol{\mathcal{U}}^{-1}\boldsymbol{x}(k)$ to the state-space system $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$, leads to a new equivalent realization $(\boldsymbol{\mathcal{U}}^{-1}\boldsymbol{A}\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{U}}^{-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{\mathcal{U}}, d)$.

Since these two realizations are equivalent in infinite precision but are no more equivalent in finite precision (fixed point arithmetic, floating-point arithmetic, etc.), the L_2-sensitivity then depends on $\boldsymbol{\mathcal{U}}$, and is denoted $M_{L_2}(\boldsymbol{\mathcal{U}})$.
In this case, it is natural to define the following problem:

Problem 1 (optimal L_2-sensitivity problem) *Considering a state-space realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$, the optimal L_2-sensitivity problem consists of finding the coordinate transformation $\boldsymbol{\mathcal{U}}_{opt}$ that minimizes M_{L_2}:*

$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ invertible}}{\arg\min} M_{L_2}(\boldsymbol{\mathcal{U}}). \quad (12)$$

In [1], it is shown that the problem has one unique solution. Hence, for example, a gradient method can be used to solve it.

III. Transfer function error

A. Fixed-point implementation

In this paper, the notation (β, γ) is used for the fixed-point representation of a variable or coefficient (2's complement scheme), according to Figure 1. β is the total wordlength of the representation in bits, whereas γ is the wordlength of the fractional part (it determines the position of the binary-point). They are fixed for each variable (input, states, output) and each coefficient, and implicit (unlike the floating-point representation). β and γ will be suffixed by the variable/coefficient they refer to. These parameters could be scalars, vectors or matrices, according to the variables they refer to.

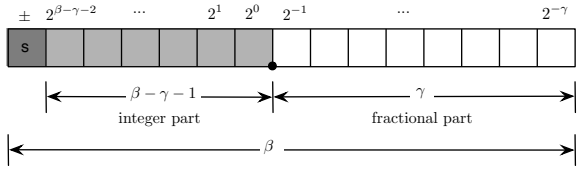

Fig. 1. Fixed-point representation

Let us suppose that the wordlength of the coefficients $\boldsymbol{\beta}_{\boldsymbol{Z}}$ is given[3]. Then, the coefficients $\boldsymbol{Z}_{ij}$ are represented in fixed-point by $(\beta_{\boldsymbol{Z}_{ij}}, \gamma_{\boldsymbol{Z}_{ij}})$ with:

$$\gamma_{\boldsymbol{Z}_{ij}} = \beta_{\boldsymbol{Z}_{ij}} - 2 - \left\lfloor \log_2 |\boldsymbol{Z}_{ij}| \right\rfloor. \quad (13)$$

where the $\lfloor a \rfloor$ operation rounds a to the nearest integer less or equal to a (for positive numbers $\lfloor a \rfloor$ is the integer part).

Remark 2 The binary point position is not defined for null coefficients, however this is no problem because these coefficients will not be represented in the final algorithm (the null multiplications are removed).
So, in order to consider coefficients that will be quantized without error, we introduced a *weighting* matrix $\boldsymbol{\delta}_{\boldsymbol{Z}}$ such that

$$(\delta_{\boldsymbol{Z}})_{ij} \triangleq \begin{cases} 0 & \text{if } \boldsymbol{Z}_{ij} \text{ is exactly implemented} \\ 1 & \text{otherwise.} \end{cases} \quad (14)$$

The exactly implemented coefficients are 0, ± 1 and also positive and negative coefficients of power of 2.

Remark 3 In some specific computational cases the fixed-point representation chosen for the coefficients is not always the best one as defined in (13). For example, in the *Roundoff Before Multiplication* scheme, some extra quantizations are added to the coefficients, in order to avoid shift operations after multiplications [2]. Only the classical case (corresponding to the *Roundoff After Multiplication*) is considered here, as defined by equation (13).

Remark 4 It is also possible to choose the same fixed-point representation for all the coefficients (determined by the

[3]In FPGA or ASIC, it is of interest to consider the wordlength as optimization variables, in order to find hardware realizations that minimize hardware criteria like power consumption or surface, under certain numerical accuracy constraints, like L_2-sensitivity ones [8]. This is not considered in this paper.

coefficient with the highest magnitude). But in that case, the lowest coefficients (in magnitude) do not have an appropriate representation. They are coded with less meaningful bits and have a higher relative error. When the ratio between the greatest and lowest magnitude is too high, then underflows occurs for the lowest coefficients that cannot be represented. For example, this is common for the Direct Form realizations with high (or low) L_2-gain.

During the quantization process, the coefficients are changed from $\boldsymbol{Z}$ into $\boldsymbol{Z}^\dagger \triangleq \boldsymbol{Z} + \boldsymbol{\Delta Z}$. For a best-roundoff quantization, the $\{\boldsymbol{\Delta Z}_{i,j}\}$ are independent centered random variables uniformly distributed [9] within the ranges $-2^{-\gamma_{\boldsymbol{Z}_{ij}}-1} \leqslant \boldsymbol{Z}_{i,j} < 2^{-\gamma_{\boldsymbol{Z}_{ij}}-1}$, so their second-order moments are given by

$$\sigma^2_{\boldsymbol{Z}_{ij}} \triangleq E\left\{(\boldsymbol{\Delta Z}_{ij})^2\right\} \tag{15}$$

$$= \frac{2^{-2\gamma_{\boldsymbol{Z}_{ij}}}}{12}\delta_{\boldsymbol{Z}_{ij}}, \tag{16}$$

where $E\{.\}$ is the mean operator.

B. Transfer function error

Due to the quantization of the coefficients, the transfer function is changed from h to $h^\dagger \triangleq h + \Delta h$. This degradation can be evaluated in a statistical way with the following definition.

Definition 3 (Transfer function error) *A measure of the transfer function error can be statistically defined by [5]*

$$\sigma^2_{\Delta h} \triangleq \frac{1}{2\pi}\int_0^{2\pi} E\left\{\left|\Delta h\left(e^{j\omega}\right)\right|^2\right\} d\omega. \tag{17}$$

The transfer function error is a tractable measure that can be evaluated with the following propostion.

Proposition 2 *The transfer function error is given by:*

$$\sigma^2_{\Delta h} = \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \boldsymbol{\Xi}_{\boldsymbol{Z}} \right\|_2^2 \tag{18}$$

where $\times$ is the Schur product, $\boldsymbol{\Xi}_{\boldsymbol{Z}} \in \mathbb{R}^{(n+1)\times(n+1)}$ defined by:

$$(\boldsymbol{\Xi}_{\boldsymbol{Z}})_{ij} \triangleq \begin{cases} \frac{2^{-\boldsymbol{\beta}_{\boldsymbol{Z}_{ij}}+1}}{\sqrt{3}} \lfloor \boldsymbol{Z}_{ij} \rfloor_2 (\boldsymbol{\delta}_{\boldsymbol{Z}})_{ij} & \text{if } \boldsymbol{Z}_{ij} \neq 0 \\ 0 & \text{if } \boldsymbol{Z}_{ij} = 0 \end{cases} \tag{19}$$

and $\lfloor x \rfloor_2$ is nearest power of 2 lower than $|x|$:

$$\lfloor x \rfloor_2 \triangleq 2^{\lfloor \log_2 |x| \rfloor}. \tag{20}$$

Proof: A first order approximation gives

$$\Delta h(z) = \sum_{i,j} \frac{\partial h}{\partial \boldsymbol{Z}_{ij}}(z)\boldsymbol{\Delta Z}_{ij}, \qquad \forall z \in \mathbb{C}. \tag{21}$$

Hence

$$E\left\{\left|\Delta h\left(e^{j\omega}\right)\right|^2\right\} = \sum_{i,j}\left|\frac{\partial h(e^{j\omega})}{\partial \boldsymbol{Z}_{ij}}\right|^2 \sigma^2_{\boldsymbol{Z}_{ij}} \tag{22}$$

because the random variables $\Delta \boldsymbol{Z}_{ij}$ are independent. Considering (13) and (16) for non-null coefficients, we get

$$\sigma^2_{\boldsymbol{Z}_{ij}} = \frac{4}{3} 2^{-2\beta_{\boldsymbol{Z}_{ij}}} \lfloor \boldsymbol{Z}_{ij} \rfloor_2^2 (\delta_{\boldsymbol{Z}})_{ij} \tag{23}$$

and

$$\sigma^2_{\Delta h} = \sum_{i,j}\left\| (\boldsymbol{\Xi})_{ij} \frac{\partial h}{\partial \boldsymbol{Z}_{ij}} \right\|_2^2 \tag{24}$$

Then, with $\left(\frac{\partial h}{\partial \boldsymbol{Z}}\right)_{ij} = \frac{\partial h}{\partial \boldsymbol{Z}_{ij}}$ and (4), eq. (18) holds. ■

Remark 5 In the classical case where the wordlength of the coefficients are all the same (equal to β), we can define a normalized transfer error $\bar{\sigma}^2_{\Delta h}$ defined by:

$$\bar{\sigma}^2_{\Delta h} \triangleq \frac{3\sigma^2_{\Delta h}}{2^{-2\beta+2}}. \tag{25}$$

This measure is now independent of the wordlength, and can be used for some comparisons.

C. Comparison with the classical M_{L_2} measure

It is of interest to remark the relationship with the classical M_{L_2} measure. In [5] where the transfer function error appears for the first time, the coefficients are supposed to have the same fixed-point representation, so their second-order moment $(\sigma^2_{\boldsymbol{Z}_{ij}})$ are all equal and denoted σ^2_0. The M_{L_2} satisfies then

$$M_{L_2} = \frac{\sigma^2_{\Delta h}}{\sigma^2_0}. \tag{26}$$

Here, the transfer function error $\sigma^2_{\Delta h}$ can be seen as an extension of the M_{L_2} measure with fixed-point considerations. The sensitivity is weighted according to the variance of the quantization noise of each coefficient.

The M_{L_2} measure considers each coefficient with the same weight, even if the quantization of one particular coefficient induces a small or big modification of this coefficient. To avoid this problem, a *normalization* has been created. Since a coordinate transformation $\boldsymbol{\mathcal{U}} = \alpha \boldsymbol{I}_n$ does not change $\boldsymbol{A}$, but only multiplies the coefficients of $\boldsymbol{b}$ by $\frac{1}{\alpha}$, and those of $\boldsymbol{c}$ by α, it was of interest to set a condition on $\boldsymbol{b}$ or $\boldsymbol{c}$ for normalization.

The L_2-dynamic-range-scaling constraints have been introduced by Jackson in [10] and Hwang in [11]. It consists in scaling the state-variable vector so as to prevent overflows or underflows during its evaluation and it imposes a condition on $\boldsymbol{b}$ (it avoids big values for $\boldsymbol{c}$ and small for $\boldsymbol{b}$, and *vice versa*).

Definition 4 (L_2-scaling) *A state-space realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$ is said to be L_2-scaled if the transfer functions from the input to each state have unitary L_2-norms,* i.e.*:*

$$\left\| \boldsymbol{e}_i^\top (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}) \right\|_2 = 1, \quad \forall 1 \leqslant i \leqslant n \tag{27}$$

where $\boldsymbol{e}_i$ is the column vector of appropriate dimension and with all elements being 0 except for the i^{th} element which is 1.
With proposition 1 applied to the system $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{e}_i^\top, \boldsymbol{0})$, the L_2-scaling constraints (27) *can be expressed as:*

$$(\boldsymbol{W}_c)_{i,i} = 1, \quad \forall 1 \leqslant i \leqslant n \tag{28}$$

where $\boldsymbol{W}_c$ is the controllability Gramian of the state-space system $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$.

In addition, it has been shown in [12] that the L_2-scaling is not necessary to implement a realization without overflows. Two equivalent choices are possible for the implementation:

- define a binary-point position for each state (for example, the same as the input), and apply a scaling to them in order to adapt the peak values of each state to the chosen binary-point position. This scaling can also be based on a *relaxed* L_2-scaling, as in [12].
- or set the binary-point position for each state, according to (29), to make sure that the fixed-point representation of the states avoids state-overflows:

$$\gamma_{\boldsymbol{x}_i} = \beta_{\boldsymbol{x}_i} - 2 - \left\lfloor \log_2 \overset{\text{up}}{\boldsymbol{x}}_i \right\rfloor. \tag{29}$$

where $\overset{\text{up}}{\boldsymbol{x}}_i$ is an upper-bound of the i^{th} state, given by a L_1-norm:

$$\overset{\text{up}}{\boldsymbol{x}}_i = \left\| e_i^\top (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} \right\|_1 \overset{\max}{u}, \tag{30}$$

or estimated by a L_2-norm [12], [13]:

$$\overset{\text{up}}{\boldsymbol{x}}_i \simeq \delta \left\| e_i^\top (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} \right\|_2 \overset{\max}{u}. \tag{31}$$

The L_2-scaling constraints have to be applied in the first case, whereas no other constraint has to be applied in the second case..

Contrary to the L_2-sensitivity measure M_{L_2}, the transfer function error $\sigma^2_{\Delta h}$ can be applied in a general cases and be meaningful. In the very particular case where all the coefficients have the same fixed-point representation, these two measure are equivalent. But for other cases, only the $\sigma^2_{\Delta h}$ measure is a meaningful measure of the degradation of the transfer function due to the quantization of the coefficients.

D. Scaling

Let us consider a scaling of the states: $\boldsymbol{x}(k)$ is changed in $\boldsymbol{\mathcal{U}}^{-1}\boldsymbol{x}(k)$ with $\boldsymbol{\mathcal{U}}$ an invertible diagonal matrix. The realization $\boldsymbol{Z}_0$ is changed into $\boldsymbol{Z}_1$:

$$\boldsymbol{Z}_1 = \boldsymbol{\mathcal{T}}^{-1}\boldsymbol{Z}_0\boldsymbol{\mathcal{T}}, \text{ with } \boldsymbol{\mathcal{T}} = \begin{pmatrix} \boldsymbol{\mathcal{U}}^{-1} & \\ & \boldsymbol{I}_n \end{pmatrix} \tag{32}$$

Proposition 3 (Invariance to scaling) *A scaling with powers of 2 ($\boldsymbol{\mathcal{U}}$ diagonal with $\boldsymbol{\mathcal{U}}_{ii} = 2^{p_i}, p_i \in \mathbb{Z}, 1 \leqslant i \leqslant n$) does not change the transfer function error $\sigma^2_{\Delta h}$.*

Proof: Let $\mathscr{F}_2(x)$ denotes the fractional value of $\log_2|x|$:

$$\mathscr{F}_2(x) \triangleq \log_2|x| - \lfloor \log_2|x| \rfloor \tag{33}$$

Then the operator $\lfloor . \rfloor_2$ satisfies

$$\lfloor ab \rfloor_2 = \lfloor a \rfloor_2 \lfloor b \rfloor_2 2^{\lfloor \mathscr{F}_2(a) + \mathscr{F}_2(b) \rfloor} \tag{34}$$

and hence

$$\left\lfloor (\boldsymbol{Z}_1)_{ij} \right\rfloor_2 = \left\lfloor \boldsymbol{\mathcal{T}}_{ii}^{-1} \right\rfloor_2 \left\lfloor (\boldsymbol{Z}_0)_{ij} \right\rfloor_2 \left\lfloor \boldsymbol{\mathcal{T}}_{jj} \right\rfloor_2 \Phi_{ij} \tag{35}$$

with $\Phi_{ij} \triangleq 2^{\left\lfloor \mathscr{F}_2(\boldsymbol{\mathcal{T}}_{ii}^{-1}) + \mathscr{F}_2(\boldsymbol{\mathcal{T}}_{jj}) + \mathscr{F}_2((\boldsymbol{Z}_0)_{ij}) \right\rfloor}$. So, $\boldsymbol{\Xi}_{\boldsymbol{Z}}\big|_{\boldsymbol{Z}_1}$ is deduced from $\boldsymbol{\Xi}_{\boldsymbol{Z}}\big|_{\boldsymbol{Z}_0}$ by

$$\left(\boldsymbol{\Xi}\big|_{\boldsymbol{Z}_1}\right)_{ij} = \left(\boldsymbol{\Xi}\big|_{\boldsymbol{Z}_0}\right)_{ij} \left\lfloor \boldsymbol{\mathcal{T}}_{ii}^{-1} \right\rfloor_2 \left\lfloor \boldsymbol{\mathcal{T}}_{jj} \right\rfloor_2 \Phi_{ij}. \tag{36}$$

The similarity on $\boldsymbol{Z}_0$ changes the sensitivity such that

$$\left.\frac{\partial h}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_1} = \boldsymbol{\mathcal{T}}^\top \left.\frac{\partial h}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_0} \boldsymbol{\mathcal{T}}^{-1}, \tag{37}$$

then

$$\left.\left(\frac{\partial h}{\partial \boldsymbol{Z}_{ij}}\Xi_{ij}\right)\right|_{Z_1} = \left.\left(\frac{\partial h}{\partial \boldsymbol{Z}_{ij}}\Xi_{ij}\right)\right|_{Z_0} \frac{\left\lfloor \boldsymbol{\mathcal{T}}_{ii}^{-1} \right\rfloor_2}{\boldsymbol{\mathcal{T}}_{ii}^{-1}} \frac{\lfloor \boldsymbol{\mathcal{T}}_{jj} \rfloor_2}{\boldsymbol{\mathcal{T}}_{jj}} \Phi_{ij}. \tag{38}$$

Now we can remark that $\Phi_{ij} \in \{1, 2, 4\}$ and $\Phi_{ij} = 1$ if the power of 2 are used for the scaling. Also $\frac{\lfloor a \rfloor_2}{a} = 1$ if a is a power of 2. ■

IV. Optimal realizations

It is now possible to consider the optimal transfer function error problem. It consists of finding the optimal coordinate transformation $\boldsymbol{\mathcal{U}}_{opt}$:

$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ invertible}}{\arg\ \min}\ \sigma^2_{\Delta h}(\boldsymbol{\mathcal{U}}). \tag{39}$$

Since $\sigma^2_{\Delta h}$ is invariant to power-of-2 scaling, this optimization problem has an infinite number of solutions. So it could be of interest to *normalize* all the coordinate transforms with regard to an extra consideration. For example, this could be a L_2-scaling constraints, even if it is not necessary here.

One possible *normalization* is to apply on a realization the power-of-2 scaling $\boldsymbol{\mathcal{V}}$:

$$\boldsymbol{\mathcal{V}}_{ii} = \left\lfloor \sqrt{(\boldsymbol{W}_c)_{ii}} \right\rfloor_2, \qquad 1 \leqslant i \leqslant n \tag{40}$$

where $\boldsymbol{W}_c$ is the controlability Gramian of this realization.

Then, the *normalized* realization has the following property:

Proposition 4 *A* normalized *realization satisfies the* relaxed-L_2*-scaling constraints,* i.e.*:*

$$1 \leqslant (\boldsymbol{W}_c)_{ii} < 4, \qquad 1 \leqslant i \leqslant n. \tag{41}$$

This relaxed-*constraints were proposed in [12] as an extension of the strict-L_2-scaling constraints* (28) *that still prevents the implementation from overflows.*

Proof: After the *normalization*, the new realization will have a controlability Gramian changed into $\boldsymbol{W}'_c = \boldsymbol{\mathcal{V}}^{-1}\boldsymbol{W}_c\boldsymbol{\mathcal{V}}^{-\top}$, hence

$$(\boldsymbol{W}'_c)_{ii} = \left(\frac{\sqrt{(\boldsymbol{W}_c)_{ii}}}{\left\lfloor \sqrt{(\boldsymbol{W}_c)_{ii}} \right\rfloor_2} \right)^2. \tag{42}$$

We can remark that $\forall x, 1 \leqslant \frac{x}{\lfloor x \rfloor_2} < 2$, so the *normalized* realization satisfies (41). ■

Of course, any other normalization is possible, but this one allows us to use the existing L_2-scaling or relaxed L_2-scaling methods [12], [14].

Then, the optimal problem can be defined as

Problem 2 (normalized $\sigma^2_{\Delta h}$-optimal realization)
Considering a state-space realization $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$, the $\sigma^2_{\Delta h}$-optimal realization can be found by solving the following optimization problem:

$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ invertible}}{\arg\min}\ \sigma^2_{\Delta h}(\boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{V}}), \tag{43}$$

where $\boldsymbol{\mathcal{V}}$ is a diagonal scaling matrix such that

$$\boldsymbol{\mathcal{V}}_{ii} = \left\lfloor \sqrt{(\boldsymbol{\mathcal{U}}^{-1}\boldsymbol{W}_c\boldsymbol{\mathcal{U}}^{-\top})_{ii}} \right\rfloor_2 \tag{44}$$

Proof: $\boldsymbol{\mathcal{V}}$ is built in such way that the coordinate transformation $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{V}}$ applied on the realization performes the *normalization*. ∎

Since the $\sigma^2_{\Delta h}$ measure is non smooth, this optimization problem can be solved with a global optimization method such as the Adaptive Simulated Algorithm (ASA) [15], [16]. A gradient-base method such as the quasi-Newton algorithm leads to local optima, which in practice are however not too far from the global optimum.
The FWR Toolbox[4] was used for the numerical examples, and few minutes of computation were here required on a desktop computer.

V. Numerical example

Let us consider the filter with coefficients given by the Matlab command `butter(4,0.05)`, and some equivalent (in infinite precision) realizations:

- $\boldsymbol{Z}_1$: the Direct Form II,
- $\boldsymbol{Z}_2$: the balanced realization,
- $\boldsymbol{Z}_3$: the normalized $\bar{\sigma}^2_{\Delta h}$-optimal realization (obtained with ASA and proposition 2).

The following table gives the $\bar{\sigma}^2_{\Delta h}$ measure of these different realizations. This could be compared to the *a posteriori* difference between the transfer function h and the transfer function with quantized coefficients (denoted $h^\dagger$). The wordlengths used are 16, 14 and 10 bits (but 10 bits are not enough to preserve the stability of the Direct Form II, that cannot be used with so few bits):

realization	$\bar{\sigma}^2_{\Delta h}$	$\left\lVert h - h^\dagger \right\rVert_2$		
		16 bits	14 bits	10 bits
$\boldsymbol{Z}_1$	$5.690e+6$	$2.055e-2$	0.1578	$N.A.$
$\boldsymbol{Z}_2$	3.693	$3.678e-5$	$1.6994e-4$	$3.0375e-3$
$\boldsymbol{Z}_3$	1.439	$2.189e-5$	$2.3148e-5$	$1.8358e-4$

The realization $\boldsymbol{Z}_3$ has of course the lowest transfer function error. Even with few bits, the degradation of realization Z_3 remain low, compared to Z_1 and Z_2.
Even if it is non-optimal, Z_2 performs quite well with 16-bit implementation, but is less accurate than Z_3 with less bits.
As we can remark, the coefficients of these three realizations have different magnitudes, and that legitimates the use of $\sigma^2_{\Delta h}$.
Here are the 16-bit fixed-point coefficients of $\boldsymbol{Z}_3^\dagger$ (each one has a different binary-point position):

$$\boldsymbol{z}_3^\dagger = \begin{pmatrix} +29648\cdot 2^{-15} & +27141\cdot 2^{-18} & +20820\cdot 2^{-20} & -30467\cdot 2^{-19} & -32227\cdot 2^{-19} \\ +24569\cdot 2^{-20} & +29679\cdot 2^{-15} & +22295\cdot 2^{-17} & -31725\cdot 2^{-20} & +19083\cdot 2^{-22} \\ -31503\cdot 2^{-20} & -31152\cdot 2^{-19} & +29148\cdot 2^{-15} & +30424\cdot 2^{-22} & -32633\cdot 2^{-15} \\ +22733\cdot 2^{-17} & +21076\cdot 2^{-20} & -32727\cdot 2^{-21} & +29154\cdot 2^{-15} & -26416\cdot 2^{-31} \\ +28776\cdot 2^{-24} & -32739\cdot 2^{-22} & -25371\cdot 2^{-26} & -32767\cdot 2^{-18} & +16771\cdot 2^{-29} \end{pmatrix}$$

[4] sources available at http://fwrtoolbox.gforge.inria.fr

VI. Conclusion

This paper has presented the optimal realization problem in fixed-point context and exhibited a new measure to evaluate the coefficients roundoff errors. Compared to the classical L_2-sensitivity measure, the transfer function error is a meaningful measure for every realization, even for non-L_2-scaled ones.
A normalization procedure used to solve the optimal realization problem has been presented with a numerical example, but there is still further work to be done, specially to develop an *ad hoc* optimization algorithm.
This measure could be easily adapted to the Specialized Implicit Framework [17] that allows to encompass all the existing structures (direct forms, cascade/parallel decompositions, lattices, δ or ρ-operator based realizations, etc.). Some other measures, like the pole-sensitivity measure, could be extended to consider fixed-point coefficients.

References

[1] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Probems*. Springer-Verlag, 1993.
[2] T. Hilaire, D. Ménard, and O. Sentieys, "Bit accurate roundoff noise analysis of fixed-point linear controllers," in *Proc.* IEEE Int. Symposium on Computer-Aided Control System Design *(CACSD'08)*, Sept. 2008.
[3] S. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. 25, no. 4, pp. 273–281, August 1977.
[4] C. Mullis and R. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," in *IEEE Transactions on Circuits and Systems*, vol. CAS-23, no. 9, September 1976.
[5] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W. Lu, "Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability gramians," in *IEEE Transactions on Circuits and Systems, Fundamental Theory and Applications*, vol. 49, no. 9, september 2002.
[6] V. Tavşanoğlu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensibility and roundoff noise," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. CAS-31, October 1984.
[7] T. Hilaire and P. Chevrel, "On the compact formulation of the derivation of a transfer matrix with respect to another matrix," INRIA, Tech. Rep. RR-6760, 2008.
[8] R. Rocher, D. Ménard, N. Hervé, and O. Sentieys, "Fixed-point configurable hardware components," *EURASIP Signal of Embedded Systems*, no. 1, p. 20, January 2006.
[9] A. Sripad and D. Snyder, "A necessary and sufficient condition for quantization error to be uniform and white," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 5, no. 10, 1977, pp. 442–448.
[10] L. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *Audio and Electroacoustics, IEEE Transactions on*, vol. 18, no. 2, pp. 107–122, June 1970.
[11] S. Hwang, "Dynamic range constraint in state-space digital filtering," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 23, 1975, pp. 591–593.
[12] T. Hilaire, "Low parametric sensitivity realizations with relaxed l_2-dynamic-range-scaling constraints," *IEEE Trans. on Circuits & Systems II*, vol. 56, no. 7, pp. 590–594, July 2009.
[13] K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation of Digital Controllers*. John Wiley & Sons, 1999.
[14] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Minimization of l2 sensitivity of one- and two dimensional state-space digital filters subject to l2-dynamic-range-scaling constraints," *IEEE Trans. on Circuits and Systems-II*, vol. 52, no. 10, pp. 641–645, October 2005.
[15] L. Ingber, "Adaptive simulated annealing (ASA): Lessons learned," *Control and Cybernetics*, vol. 25, no. 1, pp. 33–54, 1996.
[16] S. Chen and B. Luk, "Adaptive Simulated Annealing for optimization in signal processing applications," *Signal Processing*, vol. 79, pp. 117–128, 1999.
[17] T. Hilaire, P. Chevrel, and J. Whidborne, "A unifying framework for finite wordlength realizations," *IEEE Trans. on Circuits and Systems*, vol. 8, no. 54, pp. 1765–1774, August 2007.