

# Low Parametric Sensitivity Realizations with relaxed $L_2$ -dynamic-range-scaling constraints

Thibault Hilaire, *Member IEEE*

**Abstract**—This paper presents a new dynamic-range scaling for the implementation of filters/controllers in state-space form. Relaxing the classical  $L_2$ -scaling constraints by specific fixed-point considerations allows for a higher degree of freedom for the optimal  $L_2$ -parametric sensitivity problem. However, overflows in the implementation are still prevented. The underlying constrained problem is converted into an unconstrained problem for which a solution can be provided. This leads to realizations which are still scaled but less sensitive.

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

## I. INTRODUCTION

The majority of control (or signal processing) systems is implemented in digital general purpose processors, DSPs<sup>1</sup>, FPGAs<sup>2</sup>, 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.

It is also well known that these Finite Word Length (FWL) effects depend on the structure of the realization. This motivates to investigate the coefficient sensitivity minimization problem. It has been widely studied since Thiele published [5], [6] and the definition of a tractable input-output sensitivity norm (the  $L_1/L_2$ -sensitivity). This work has been extended with a more natural and reasonable measure, the  $L_2$ -sensitivity ([1], [7]). The dynamic-range-scaling constraints have been introduced in [8] and [9] to prevent overflow and underflow during the evaluation of the state-vector, and as well as the state and criteria normalization. These constraints have to be considered in the  $L_2$ -sensitivity minimization problem, for which [10] proposes an efficient quasi-Newton algorithm to solve it.

This paper investigates the  $L_2$ -dynamic-range-scaling problem by considering concrete fixed-point implementation of state-space realizations. It reveals that the classical  $L_2$ -scaling is only a sufficient condition to prevent overflows and thus

it can be slightly relaxed in order to extend the degrees of freedom for the optimization process. New relaxed- $L_2$ -dynamic-range-scalings are then presented with respect to the described computational scheme. Finally, the  $L_2$ -sensitivity minimization problem with relaxed  $L_2$ -scaling constraints is solved. A numerical example illustrates that the proposed constraints can offer reduced  $L_2$ -sensitivity with overflow protection.

## II. $L_2$ -SENSITIVITY ANALYSIS

Let  $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$  be a stable, controllable and observable linear discrete time SISO<sup>3</sup> state-space system, i.e.

$$\begin{cases} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{b}u(k) \\ y(k) &= \mathbf{c}\mathbf{x}(k) + du(k) \end{cases} \quad (1)$$

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

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

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

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

**Definition 1 (Transfer function sensitivity)** Consider  $\mathbf{X} \in \mathbb{R}^{m \times n}$  a matrix and  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{C}$  a scalar complex function, differentiable with respect to all the entries of  $\mathbf{X}$ . The sensitivity of  $f$  with respect to  $\mathbf{X}$  is defined by the matrix  $\mathbf{S}_{\mathbf{X}} \in \mathbb{R}^{m \times n}$ :

$$\frac{\partial f}{\partial \mathbf{X}} \triangleq \mathbf{S}_{\mathbf{X}} \quad \text{with} \quad (\mathbf{S}_{\mathbf{X}})_{i,j} \triangleq \frac{\partial f}{\partial \mathbf{X}_{i,j}} \quad (3)$$

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

$$\|\mathbf{H}\|_p \triangleq \left( \frac{1}{2\pi} \int_0^{2\pi} \|\mathbf{H}(e^{j\omega})\|_F^p d\omega \right)^{\frac{1}{p}} \quad (4)$$

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

<sup>3</sup>Single Input Single Output

Manuscript received February 24, 2009.

This work began with the Cairn Project and INRIA Lab (France). The author is now with the Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology, Austria.

This work has been funded in parts by the NFN SISE project (National Research Network "Signal and Information Processing in Science and Engineering").

<sup>1</sup>Digital Signal Processors

<sup>2</sup>Field Programmable Gate-Array

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

$$M_{L_2} \triangleq \left\| \frac{\partial h}{\partial \mathbf{A}} \right\|_2^2 + \left\| \frac{\partial h}{\partial \mathbf{b}} \right\|_2^2 + \left\| \frac{\partial h}{\partial \mathbf{c}} \right\|_2^2 + \left\| \frac{\partial h}{\partial d} \right\|_2^2 \quad (5)$$

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

$$\mathbf{F}(z) \triangleq (z\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{b}, \quad \mathbf{G}(z) \triangleq \mathbf{c}(z\mathbf{I}_n - \mathbf{A})^{-1}. \quad (6)$$

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 [5] ( $\left\| \frac{\partial h}{\partial \mathbf{A}} \right\|_1^2$  instead of  $\left\| \frac{\partial h}{\partial \mathbf{A}} \right\|_2^2$  in (5)).

**Remark 1** It is also possible to regroup all the coefficients in one unique matrix

$$\mathbf{Z} \triangleq \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{c} & d \end{pmatrix}. \quad (7)$$

Then, with  $L_2$ -norm property,  $M_{L_2} = \left\| \frac{\partial h}{\partial \mathbf{Z}} \right\|_2^2$ . From (6) and the associated state-spaces, the sensitivity transfer function  $\frac{\partial h}{\partial \mathbf{Z}}$  can be described by the MIMO<sup>4</sup> state-space system  $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{D}})$  with

$$\begin{aligned} \tilde{\mathbf{A}} &\triangleq \begin{pmatrix} \mathbf{A} & \mathbf{bc} \\ \mathbf{0} & \mathbf{A} \end{pmatrix}, \tilde{\mathbf{B}} \triangleq \begin{pmatrix} \mathbf{0} & \mathbf{b} \\ \mathbf{I}_n & \mathbf{0} \end{pmatrix}, \\ \tilde{\mathbf{C}} &\triangleq \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{c} \end{pmatrix}, \tilde{\mathbf{D}} \triangleq \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}. \end{aligned} \quad (8)$$

See [1] and [11] for more details.

The following proposition allows to compute  $M_{L_2}$ .

**Proposition 1** *Let us consider  $\mathbf{H}$  the MIMO state-space system  $(\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N})$ . Its  $L_2$ -norm can be computed by*

$$\begin{aligned} \|\mathbf{H}\|_2^2 &= \text{tr}(\mathbf{N}\mathbf{N}^\top + \mathbf{M}\mathbf{W}_c\mathbf{M}^\top) \\ &= \text{tr}(\mathbf{N}^\top\mathbf{N} + \mathbf{L}^\top\mathbf{W}_o\mathbf{L}) \end{aligned} \quad (9)$$

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

$$\mathbf{W}_c = \mathbf{K}\mathbf{W}_c\mathbf{K}^\top + \mathbf{L}\mathbf{L}^\top, \quad \mathbf{W}_o = \mathbf{K}^\top\mathbf{W}_o\mathbf{K} + \mathbf{M}^\top\mathbf{M}. \quad (11)$$

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

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  $\mathbf{T}$ , and is denoted  $M_{L_2}(\mathbf{T})$ .

In this case, it is natural to define the following problem:

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

$$\mathbf{T}_{opt} = \arg \min_{\mathbf{T} \text{ invertible}} M_{L_2}(\mathbf{T}). \quad (12)$$

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

### III. $L_p$ -DYNAMIC-RANGE SCALING

The  $L_p$ -dynamic-range-scaling constraints have been introduced by Jackson in [8] and Hwang in [9]. It consists in scaling the state-variable vector such that overflows or underflows during its evaluation are prevented.

**Definition 3 ( $L_p$ -scaling)** *A state-space realization  $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$  is said to be  $L_p$ -scaled if the  $L_p$ -norms of the transfer functions from the input to each state are set to 1, i.e.:*

$$\left\| \mathbf{e}_i^\top (z\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{b} \right\|_p = 1, \quad \forall 1 \leq i \leq n \quad (13)$$

where  $\mathbf{e}_i$  is the column vector of appropriate dimension and with all elements being 0 except from the  $i^{\text{th}}$  element which is 1.

Let  $u^{\max}$  denote the maximum value of the input  $u$ :

$$u^{\max} \triangleq \max_{k \in \mathbb{N}} |u(k)| \quad (14)$$

The  $L_1$ -scaling guarantees that the dynamic of each state  $x_i$  is lower than  $u^{\max}$ , whereas the  $L_2$ -scaling guarantees that the variance of each state is unitary for a unit-variance centered white noise input.  $L_2$ -scaling doesn't completely prevent overflow as does  $L_1$ , but it is less conservative and more realistic, so it is widely used [12].

With proposition 1 applied to the system  $(\mathbf{A}, \mathbf{b}, \mathbf{e}_i^\top, 0)$ , the  $L_2$ -scaling constraints (13) can be expressed as:

$$(\mathbf{W}_c)_{i,i} = 1, \quad \forall 1 \leq i \leq n \quad (15)$$

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

**Problem 2 (sensitivity problem with  $L_2$ -scaling constraints)**

*The optimal  $L_2$ -sensitivity problem with  $L_2$ -scaling constraints can be formulated as the optimization problem 1, subject to the constraints in (15).*

Moreover, it is possible to  $L_2$ -scale a realization with the following proposition.

**Proposition 2 (a posteriori  $L_2$ -scaling)** *Considering a state-space realization  $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$ , it is also possible to a posteriori  $L_2$ -scale it with a diagonal coordinate transformation  $\mathbf{T}$  such that:*

$$\mathbf{T}_{i,i} = \sqrt{(\mathbf{W}_c)_{i,i}}, \quad \forall 1 \leq i \leq n \quad (16)$$

*Then, there exist infinite transformation matrices  $\mathbf{T}$  (not necessary diagonal) that produces  $L_2$ -scaled realizations: Let us consider the invertible matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$ , then also the transformation matrix  $\mathbf{T} = \mathbf{U}\mathbf{V}$  produces  $L_2$ -scaling with  $\mathbf{V}$  diagonal such that:*

$$\mathbf{V}_{i,i} = \sqrt{(\mathbf{U}^{-1}\mathbf{W}_c\mathbf{U}^{-\top})_{i,i}}, \quad \forall 1 \leq i \leq n \quad (17)$$

*Proof:* A transformation matrix  $\mathbf{T}$  that transforms  $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$  into  $(\mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \mathbf{T}^{-1}\mathbf{b}, \mathbf{c}\mathbf{T}, d)$  changes the controllability Gramian  $\mathbf{W}_c$  into  $\mathbf{T}^{-1}\mathbf{W}_c\mathbf{T}^{-\top}$ .

Since  $\mathbf{T}$  is diagonal, the constraints  $(\mathbf{W}_c)_{i,i} = 1$  imply  $\mathbf{T}_{i,i} = \sqrt{(\mathbf{W}_c)_{i,i}}$ .

<sup>4</sup>Multiple Inputs Multiple Outputs

Moreover, it is also possible to successively apply two transformation matrices  $U$  and  $V$  on  $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$ . If  $V$  is composed according to (17), then the transformation  $T = UV$  performs the  $L_2$ -scaling. ■

This proposition can be used to transform the constrained problem 2 into an unconstrained problem. Then an optimization algorithm, like quasi-Newton, can be used to solve it. Other analytical algorithms for this problem can be found in [10] and [3].

#### IV. FIXED-POINT IMPLEMENTATION

##### A. Fixed-point representation

In this paper, the notation  $(\beta, \gamma)$  is used for the fixed-point representation of a variable or coefficient (2's complement scheme), according to Figure 1.  $\beta$  is the total wordlength of the representation in bits, whereas  $\gamma$  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).  $\beta$  and  $\gamma$  will be suffixed by the variable/coefficient they refer to.

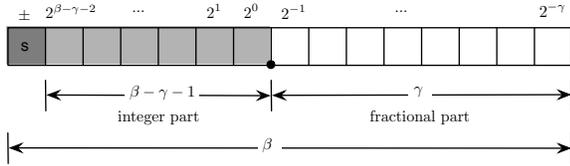


Fig. 1. Fixed-point representation

To represent a value  $x$  without overflow, a fixed-point representation  $(\beta_x, \gamma_x)$  may satisfy:

$$\beta_x - \gamma_x - 1 \geq \lceil \log_2 |x| \rceil + 1 \quad (18)$$

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

An important fixed-point issue is to find a valid fixed-point representation, such that (18) is satisfied for all values which  $x$  can assume during the execution of the algorithm.

##### B. State-overflow

**Definition 4 (State-overflow)** *The overflow of the state variables  $(\mathbf{x}_i)_{1 \leq i \leq n}$  can be strictly avoided iff  $(1 \leq i \leq n)$*

$$\forall k, \quad -2^{\beta_{\mathbf{x}_i} - \gamma_{\mathbf{x}_i} - 1} \leq \mathbf{x}_i(k) < 2^{\beta_{\mathbf{x}_i} - \gamma_{\mathbf{x}_i} - 1}. \quad (19)$$

*The overflows are avoided if the binary-point position of each state is carefully chosen, such that*

$$\gamma_{\mathbf{x}_i} = \beta_{\mathbf{x}_i} - 2 - \left\lceil \log_2 \max \mathbf{x}_i \right\rceil, \quad (20)$$

where  $\max \mathbf{x}_i$  is the maximum magnitude for the  $i^{\text{th}}$  state:

$$\max \mathbf{x}_i \triangleq \max_{k \in \mathbb{N}} |\mathbf{x}_i(k)|. \quad (21)$$

However only upper bounds can be computed. A first upper bound  $\mathbf{x}_i^{\text{up}}$  can be obtained by an  $L_1$ -norm:

$$\mathbf{x}_i^{\text{up}} = \left\| \mathbf{e}_i^\top (z\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{b} \right\|_1^{\max}, \quad (22)$$

and a second one can be estimated by an  $L_2$ -norm [12]:

$$\mathbf{x}_i^{\text{up}} \simeq \delta \left\| \mathbf{e}_i^\top (z\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{b} \right\|_2^{\max} u. \quad (23)$$

Here, the parameter  $\delta$  can be interpreted as a representation of the number of standard deviations of  $\mathbf{x}_i$ , if the input is unit-variance white centered noise ( $\delta \geq 1$ ). Since the  $L_2$ -norm in (23) doesn't give a strict bound (contrary to (22)),  $\delta$  can be seen as a *safety* parameter [12].

Finally, these upper bounds are used to define the binary-point positions:

$$\gamma_{\mathbf{x}_i} = \beta_{\mathbf{x}_i} - 2 - \left\lceil \log_2 \mathbf{x}_i^{\text{up}} \right\rceil. \quad (24)$$

In general, the  $L_1$  and  $L_2$  estimations of  $\mathbf{x}_i^{\text{up}}$  approximately leads to the same binary-point position, with 1 or 2 bits deviation. However, since the  $L_2$ -norm is more tractable (with proposition 1) and the  $L_1$ -norm too conservative ( $\mathbf{x}_i^{\max} \ll \mathbf{x}_i^{\text{up}}$ ), in practice (23) is used, with  $\delta = 1$ . After implementation, a simulation-based estimation like in [13] or [14] can also be used to verify *in situ* the peak values and the binary point positions, according to the inputs.

##### C. Computational scheme

In order to implement a realization without overflows, two equivalent choices are possible:

- set the binary-point position for each state, according to (24), to make sure that the fixed-point representation of the states avoids state-overflows;
- or define a binary-point position for each state, and apply a scaling to them in order to adapt the peak values of each state to the chosen binary-point position.

Here, we here focus on the 2<sup>nd</sup> choice, referring to dynamic-range-scaling constraints.

Let us consider in detail the fixed-point implementation of the system given in (1). It leads to  $(n + 1)$  scalar products to be evaluated, of the form:

$$S = \sum_{i=1}^N \mathbf{p}_i \mathbf{q}_i \quad (25)$$

where the  $(\mathbf{p}_i)$  are given coefficients and  $(\mathbf{q}_i)$  are bounded variables.

To avoid bit-shift operations between each addition in the evaluation of eq. (25), the binary-point positions of each partial product of the sum should be equal.

Then, two computational schemes are possible: the *Roundoff After Multiplication* scheme, where shifts are added after each product to align the operands of the sum ( $\mathbf{p}_i \mathbf{q}_i$  is implemented as  $(\mathbf{p}'_i * \mathbf{q}'_i) \gg d_i$ ) and the *Roundoff Before Multiplication* scheme, where the required shifts are reported into the coefficients ( $\mathbf{p}_i \mathbf{q}_i$  is implemented as  $(\mathbf{p}'_i \gg d_i) * \mathbf{q}'_i$ ).

The main idea of the scaling is to scale each variable  $(\mathbf{q}_i)$  such that the shifts ( $d_i = 0, \forall i$ ) are prevented. In fixed-point representation, the scaling only implies that all the  $(\mathbf{q}_i)$  have a common format, and so have the  $(\mathbf{p}_i)$ . See [2], [15] for more details on implementation schemes.

Applied to the state-space realization (1), this yields that all the states must have the same binary-point position as the

input and the coefficients  $\mathbf{A}$ ,  $\mathbf{b}$ ,  $\mathbf{c}$  and  $d$ .

Besides, since they have the same fractional part, their quantization's errors  $\Delta\mathbf{A}$ ,  $\Delta\mathbf{b}$ ,  $\Delta\mathbf{c}$  and  $\Delta d$  have the same magnitude  $2^{-\gamma z^{-1}}$ , and the  $L_2$ -sensitivity measure represents a meaningful bound on the transfer function error  $\Delta h$ :

$$\begin{aligned} \|\Delta h\|_2^2 &\leq \left\| \frac{\partial h}{\partial \mathbf{A}} \times \Delta\mathbf{A} \right\|_2^2 + \left\| \frac{\partial h}{\partial \mathbf{b}} \times \Delta\mathbf{b} \right\|_2^2 \\ &\quad + \left\| \frac{\partial h}{\partial \mathbf{c}} \times \Delta\mathbf{c} \right\|_2^2 + \left\| \frac{\partial h}{\partial d} \times \Delta d \right\|_2^2 \\ &\leq 2^{-2(\gamma z+1)} M_{L_2} \end{aligned} \quad (26)$$

#### D. New $L_2$ -scaling constraints

Taken this into consideration, the overflows will be avoided by setting the same binary-point position for the states and the input, and by applying an appropriate scaling on the states such that the constraints (20) are satisfied.

Compared to strict  $L_2$ -scaling where the states must satisfy  $\mathbf{x}_i = \overset{\max}{u}$ , here, the constraints are relaxed (but still restrictive enough to guarantee the protection against overflow) and replaced by  $\gamma_{\mathbf{x}_i} = \gamma_u$ .

**Proposition 3 (relaxed- $L_2$ -scaling constraints)** *Since the input and the states may have the same binary-point position, the  $L_2$ -scaling constraints (15) are now transformed into*

$$\frac{2^{2\alpha_i}}{\delta^2} \leq (\mathbf{W}_c)_{i,i} < 4 \frac{2^{2\alpha_i}}{\delta^2}, \quad \forall 1 \leq i \leq n \quad (28)$$

where

$$\alpha_i \triangleq \beta_{\mathbf{x}_i} - \beta_u - \mathcal{F}_2 \left( \overset{\max}{u} \right) \quad (29)$$

and  $\mathcal{F}_2(x)$  is defined as the fractional value of  $\log_2(x)$ :

$$\mathcal{F}_2(x) \triangleq \log_2(x) - \lfloor \log_2(x) \rfloor. \quad (30)$$

For microcontroller or DSP implementations (contrary to FPGA or some ASIC implementations), the wordlength of all variables is equal, i.e.  $\beta_u = \beta_{\mathbf{x}_i}$  ( $1 \leq i \leq n$ ). Also  $\overset{\max}{u}$  could be set to a power of 2. Then, if  $\delta$  is set to unity (as for classical  $L_2$ -scaling constraints), the relaxed- $L_2$ -scaling constraints (28) become:

$$1 \leq (\mathbf{W}_c)_{i,i} < 4, \quad \forall 1 \leq i \leq n \quad (31)$$

*Proof:* The binary-point position of the input is set to  $\gamma_u = \beta_u - 2 - \lfloor \log_2 \overset{\max}{u} \rfloor$ . Hence, with (24), the constraints  $\gamma_u = \gamma_{\mathbf{x}_i}$  lead to

$$\beta_u - \lfloor \log_2 \overset{\max}{u} \rfloor = \beta_{\mathbf{x}_i} - \lfloor \log_2 \left( \delta \left\| \mathbf{e}_i^\top (z\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{b} \right\|_2 \overset{\max}{u} \right) \rfloor$$

and

$$\left\lfloor \log_2 \left( \delta \sqrt{(\mathbf{W}_c)_{i,i}} + \mathcal{F}_2 \left( \overset{\max}{u} \right) \right) \right\rfloor = \beta_{\mathbf{x}_i} - \beta_u \quad (32)$$

And finally

$$2^{\alpha_i} \leq \delta \sqrt{(\mathbf{W}_c)_{i,i}} < 2^{\alpha_i+1} \quad (33)$$

It is important to remark that these new constraints allow more freedom for the scaling and introduce a new degree of freedom for the search for optimal realizations. Even though not considered in this paper, moreover it could give more freedom for the minimization of the roundoff noise power. ■

## V. OPTIMAL $L_2$ -SENSITIVITY REALIZATION WITH RELAXED $L_2$ -NORM DYNAMIC-RANGE-SCALING CONSTRAINTS

Then, these relaxed constraints can be applied to a new sensitivity problem:

**Problem 3 (relaxed sensitivity problem)** *The optimal  $L_2$ -sensitivity problem with relaxed  $L_2$ -norm dynamic-range-scaling constraints can be expressed in the form of the constrained problem 2 subject to constraints in (28).*

This constrained problem can be solved by two different means.

First, in addition to the  $n^2$  free parameters of the transformation matrix  $U$  applied to the system,  $n$  extra parameters  $(\gamma_i)_{1 \leq i \leq n}$  can be considered. These  $(\gamma_i)$  represent the desired  $L_2$ -scaling and will be constrained by

$$\frac{2^{2\alpha_i}}{\delta^2} \leq \gamma_i < 4 \frac{2^{2\alpha_i}}{\delta^2}, \quad \forall 1 \leq i \leq n. \quad (34)$$

Then a diagonal transformation matrix  $\mathbf{V}_\gamma$  is applied, with

$$(\mathbf{V}_\gamma)_{i,i} = \sqrt{\frac{(\mathbf{W}_c)_{i,i}}{\gamma_i}}. \quad (35)$$

In this case, a constrained optimization algorithm (like quasi-Newton one, implemented in `fmincon` with Matlab) can then be used to solve the following problem:

$$(\mathbf{U}_{opt}, \gamma_{opt}) = \arg \min_{\substack{U \text{ invertible} \\ \gamma \text{ satisfying (34)}}} M_{L_2}(\mathbf{U}\mathbf{V}_\gamma) \quad (36)$$

The optimal realization satisfying the relaxed- $L_2$  constraints is then obtained by applying the transformation matrix  $\mathbf{T}_{opt} = \mathbf{U}_{opt} \mathbf{V}_{\gamma_{opt}}$ .

The other approach is to scale the system after each transformation to ensure that the constraints are met:

**Proposition 4 (a posteriori relaxed scaling)** *Considering a state-space realization, it is possible to a posteriori scale it with a diagonal transformation matrix  $\mathbf{T}$  given by*

$$\mathbf{T}_{i,i} = \delta \sqrt{(\mathbf{W}_c)_{i,i}} 2^{-\mathcal{F}_2(\delta \sqrt{(\mathbf{W}_c)_{i,i}}) - \alpha_i}, \quad (37)$$

such that the constraints (28) are satisfied. Moreover, it is possible to build all the transformation matrices that meet the constraints (28): Let us consider an invertible matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$ , then the transformation matrix  $\mathbf{T} = \mathbf{U}\mathbf{V}$  with  $\mathbf{V}$  diagonal such that:

$$\mathbf{V}_{i,i} = \delta \sqrt{(\mathbf{U}^{-1} \mathbf{W}_c \mathbf{U}^{-\top})_{i,i}} 2^{-\mathcal{F}_2(\delta \sqrt{(\mathbf{U}^{-1} \mathbf{W}_c \mathbf{U}^{-\top})_{i,i}}) - \alpha_i} \quad (38)$$

produces the relaxed- $L_2$ -scaling.

*Proof:*  $\mathcal{F}_2$  acts as a modulo operator. For  $x \in \mathbb{R}$ ,  $\bar{x} \triangleq 2^{\mathcal{F}_2(x)+a}$  is such that  $2^a \leq \bar{x} < 2^{a+1}$ . Since the constraints (28) are equal to

$$2^{\alpha_i} \leq \delta \sqrt{(\mathbf{W}_c)_{i,i}} < 2^{\alpha_i+1} \quad (39)$$

and  $T$  transforms  $(W_c)_{i,i}$  into  $T_{i,i}^{-2}(W_c)_{i,i}$ , then  $T_{i,i}$  has to be of the form:

$$\delta T_{i,i}^{-1} \sqrt{(W_c)_{i,i}} = 2^{\mathcal{F}_2(\delta \sqrt{(W_c)_{i,i}}) + \alpha_i} \quad (40)$$

Thus, the optimization problem is given by

$$U_{opt} = \arg \min_{\substack{U \text{ invertible} \\ V \text{ defined by (38)}}} M_{L_2}(UV). \quad (41)$$

These two ways of solving problem 3 are implemented in the FWR toolbox<sup>5</sup> for Matlab, with `fminsearch`, `fmincon` and `fminunc` functions, and they both give same results with similar numbers of iterations.

Of course, the use of matrices  $V_\gamma$  and  $V$ , that are merely used to eliminate the constraints and solve an unconstrained minimization problem, increases the degree of non-linearity for the objective function to minimize. However, this seems not to be a problem, since in our tests, the optimal realizations found seem to be global optima.

## VI. EXAMPLE

Let us consider the following state-space digital controller, given in modal form<sup>6</sup>:

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 0.3820 & 0 & 0 \\ 0 & 0.7964 & 0.5598 \\ 0 & -0.5598 & 0.7964 \end{pmatrix}, & \mathbf{b} &= \begin{pmatrix} 0.5391 \\ -0.8417 \\ 0.6232 \end{pmatrix}, \\ \mathbf{c} &= (0.1664 \quad 0.1639 \quad 0.2047), & d &= 0.0159 \end{aligned} \quad (42)$$

and its multiple equivalent (in infinite precision) realizations:

- $\mathcal{R}_1$  is the original realization given by (42)
- $\mathcal{R}_2$  is the optimal  $L_2$ -scaled realization (solution of problem 2). It is obtained with proposition 2 and a quasi-newton algorithm.
- $\mathcal{R}_3$  is the optimal relaxed- $L_2$ -scaled realization (problem 3), with  $u^{\max}$  a power of 2, and  $\delta = 1$ . It is obtained with proposition 4.

The following table gives the  $M_{L_2}$  sensitivities of these different realizations:

realization	$M_{L_2}$ sensitivity
$\mathcal{R}_1$	6355.5
$\mathcal{R}_2$	530.0964
$\mathcal{R}_3$	528.2532

In this example, the relaxed  $L_2$ -scaled realization  $\mathcal{R}_3$  achieves lower sensitivity than the strict  $L_2$ -scaled optimal realization  $\mathcal{R}_2$  while protecting implementation from overflows. But it is not always the case : if we consider the example in [10], the optimal relaxed- $L_2$ -scaled realization satisfies  $(W_c)_{i,i} = 1$  and is then also a strict  $L_2$ -scaled realization. This depends on the diagonal terms of the controllability Gramians of the (non scaled) optimal realization.

It is also interesting to notice that a good estimation of  $u^{\max}$  (if it is not a power of 2) can allow to achieve lower sensitivity by moving the constraints (it could also be the case for the example in [10]).

<sup>5</sup>sources available at <http://fwrtoolbox.gforge.inria.fr/>

<sup>6</sup>Due to a lack of space, only 4 digits are given, but more may be required to completely define the system.

## VII. CONCLUSION

This paper has presented the  $L_2$ -sensitivity minimization problem and the associated  $L_2$ -scaling constraints. These constraints that prevent from overflows have been considered with concrete fixed-point implementation schemes. Novel  $L_2$ -dynamic-range constraints have been exhibited.

Even if the goal of this paper is not a detailed optimization algorithm like in [10], two different means to solve the constrained optimization problem have been exhibited and applied on a numerical example.

These relaxed constraints could also be very important for some other realizations, like  $\delta$ -operator state-space or the  $\rho$ -Direct Form II transposed [16]. For these realizations where a parameter  $\Delta$  should be used to achieve the  $L_2$ -scaling, a relaxed- $L_2$ -scaling permits to fix this parameter as a power of 2, in order to decrease the amount of computations.

To apply this work to other classical structures, it will be soon extended to the Specialized Implicit Framework [17] that allows to encompass existing structures in an implicit state-space form.

## REFERENCES

- [1] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems*. Springer-Verlag, 1993.
- [2] T. Hilaire, D. Ménard, and O. Sentieys, "Bit accurate roundoff noise analysis of fixed-point linear controllers," in *Proc. IEEE Intern. Symp. 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] 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.
- [6] L. Thiele, "Design of sensitivity and round-off noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39–46, 1984.
- [7] T. Hinamoto and Y. Sugie, "L2-sensitivity analysis and minimization of 2-d separable-denominator state-space digital filters," *Signal Processing, IEEE Transactions on*, vol. 50, no. 12, pp. 3107–3114, Dec 2002.
- [8] 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.
- [9] S. Hwang, "Dynamic range constraint in state-space digital filtering," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 23, 1975.
- [10] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Minimization of  $L_2$  sensitivity of one- and two dimensional state-space digital filters subject to  $L_2$ -dynamic-range-scaling constraints," *IEEE Trans. on Circuits and Systems-II*, vol. 52, no. 10, pp. 641–645, October 2005.
- [11] 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.
- [12] K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation of Digital Controllers*. John Wiley & Sons, 1999.
- [13] P. Belanovic and M. Rupp, "Automated floating-point to fixed-point conversion with the Fixify environment," *The 16th IEEE Intern. Workshop on Rapid System Prototyping, 2005.*, pp. 172–178, June 2005.
- [14] S. Kim, K. Kum, and W. Sung, "Fixed-point optimization utility for C and C++ based digital signal processing programs," *IEEE Transactions on Circuits and Systems*, vol. 45, no. 11, pp. 1455–1464, Nov. 1998.
- [15] G. Constantinides, P. Cheung, and W. Luk, *Synthesis and Optimization of DSP Algorithms*, Springer, Ed. Kluwer Academic Publishers, 2004.
- [16] G. Li and Z. Zhao, "On the generalized DFII structure and its state-space realization in digital filter implementation," *IEEE Trans. on Circuits and Systems*, vol. 51, no. 4, pp. 769–778, April 2004.
- [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.