OPTIMAL SYNTHETIC FWL DESIGN OF STATE-SPACE DIGITAL FILTERS

Gang Li Michel Gevers

Laboratoire d'Automatique, Dynamique et Analyse des Systèmes
Université Catholique de Louvain
Bâtiment Maxwell, place du Levant 3, B-1348 Louvain-la-Neuve, Belgium

Abstract: The optimal Finite Word Length (FWL) design problem of state-space filters is investigated. Instead of the usual L_1/L_2-mixed sensitivity measure, it is argued that a sensitivity measure based on the L_2 norm only is natural and reasonable. The minimization problem of this newly defined sensitivity measure is studied. The set of optimal realizations minimizing this measure is characterized. It is shown that the FWL effects can be synthesized by what is called FWL Noise Gain (FNG) which is a linear combination of the classical roundoff noise gain and the L_2 sensitivity measure. By minimizing the FNG with dynamical constraint, the optimal synthetical FWL state-space design problem is formulated. The existence of optimal realizations is shown. The necessary and sufficient condition equation that should be satisfied by the optimal realizations is given.

1 Introduction

The optimal Finite Word Length FWL state-space design has been considered as one of the most effective and elegant methods [1]-[5] in digital filter design. It is well known that any linear system can be represented by state-space equations and that this state-space model is not unique. In the infinite precision case, all these realizations are equivalent since they yield one and the same transfer function. But different realizations have different numerical properties such as sensitivity and error propagation. This means that they are no longer equivalent in the finite precision case. The optimal FWL state-space design is to identify those realizations which minimize the degradation of the system performance due to the FWL effects. The often used measures are the transfer function sensitivity [3-4] and the roundoff noise gain [1-2].

A global sensitivity measure of the transfer function with respect to the parameters of the state space model was first proposed by Tavsanoglu and Thiele [3]. In this definition of sensitivity measure two different norms, L_1 and L_2, have been mixed. The rather illogical combination of an L_1 norm with respect to some parameters and an L_2 norm with respect to others is opportunistic rather than intuitive and reasonable.

It has also been noted that in classical analysis, the FWL effects due to coefficient truncation and to arithmetic roundoff are investigated separately. Since parameter perturbation and signal roundoff exist simultaneously in actual implementations, and since both of them degrade the performance of a filter, it would be better to have a measure that simultaneously takes into account these two FWL effects. The main objective of this paper is to find a proper measure with which the FWL effects, parameter truncation and signal roundoff errors, can be unified.

2 L_2-sensitivity minimization

In this paper we consider the implementation of a discrete linear time-invariant single input, single output system which can be implemented by a minimal state-space realization:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + du(t) \end{aligned} \qquad (1)$$

with $A \in \mathbf{R}^{n\times n}, B \in \mathbf{R}^n, C^T \in \mathbf{R}^n$ and $d \in \mathbf{R}$. The transfer function can be expressed in terms of the state matrices as $H(z) = C(zI - A)^{-1}B + d$. We now define a realization set S_H of this system as follows: $S_H = \{(A, B, C, d)\}$, where (A, B, C, d) gives the same transfer function $H(z)$. Clearly, if (A, B, C, d) belongs to S_H, so does $(T^{-1}AT, T^{-1}B, CT, d)$ for any similarity transformation T. This means that S_H is an infinite set.

In practice it is impossible to realize the coefficients in (A, B, C, d) exactly due to Finite Word Length (FWL) constraints. As a result, the actual system has a transfer function $H^*(z)$ computed with the

0-7803-0532-9/92 $3.00

FWL version (A^*, B^*, C^*, d^*) of (A, B, C, d) will differ from $H(z)$. A often used measure of this effect is the sensitivity measure.

Definition 1 : Let $M \in \mathbb{R}^{n\times m}$ be a matrix and let $f(M) \in \mathbb{C}$ be a scalar complex function of M, differentiable w.r.t. all the elements of M. We then define the sensitivity function of f w.r.t. M as

$$S_M \triangleq \frac{\partial f}{\partial M} \text{ with } (S_M)_{ij} \triangleq \frac{\partial f}{\partial m_{ij}} \tag{2}$$

where m_{ij} is the $(i,j)^{th}$ element of the matrix M. ■

With these notations it is easy to show [3] that

$$\begin{aligned} \frac{\partial H(z)}{\partial A} &= G(z)F^T(z), \frac{\partial H(z)}{\partial B} = G(z) \\ \frac{\partial H(z)}{\partial C^T} &= F(z), \frac{\partial H(z)}{\partial d} = 1 \end{aligned} \tag{3}$$

where

$$\begin{aligned} F(z) &\triangleq (zI - A)^{-1}B = [f_1(z) \dots f_n(z)]^T \\ G^T(z) &\triangleq C(zI - A)^{-1} = [g_1(z) \dots g_n(z)] \end{aligned} \tag{4}$$

Since the sensitivity function w.r.t. d is constant, it will be ignored in the subsequent analysis.

Definition 2: Let $f(z) \in \mathbb{C}^{n\times m}$ be any complex matrix valued function of the complex variable z. We then define the L_p-norm of $f(z)$ as

$$\| f \|_p \triangleq \left(\frac{1}{2\pi} \int_0^{2\pi} \| f(e^{j\omega}) \|_F^p \, d\omega \right)^{1/p} \tag{5}$$

where $\| f(e^{j\omega}) \|_F$ is the Frobenius norm of the matrix $f(e^{j\omega})$:

$$\begin{aligned} \| f(e^{j\omega}) \|_F &\triangleq (\sum_{i,k} |f_{ik}(e^{j\omega})|^2)^{1/2} \\ &= \{tr[f^T(e^{-j\omega})f(e^{j\omega})]\}^{1/2}. \end{aligned} \tag{6}$$

■

Tavsanoglu and Thiele [3] have proposed the following overall sensitivity measure of the transfer function $H(z)$:

$$M_a \triangleq \| \frac{\partial H}{\partial A} \|_1^2 + \| \frac{\partial H}{\partial B} \|_2^2 + \| \frac{\partial H}{\partial C^T} \|_2^2 . \tag{7}$$

The mixing of different measures in the overall sensitivity measure above is motivated by the analytic properties of the first term on the right of (7), which allow one to derive an analytic minimization procedure for M_a: see [3] and [4]. The optimization of a more logical L_2 measure is much harder and has only recently been solved by the authors [5] and, independently, by Helmke and Moore [6].

The L_2-sensitivity measure is defined as follows [5-6]:

$$M_{L2} \triangleq \| \frac{\partial H(z)}{\partial A} \|_2^2 + \| \frac{\partial H(z)}{\partial B} \|_2^2 + \| \frac{\partial H(z)}{\partial C^T} \|_2^2 . \tag{8}$$

In this section we examine the L_2-sensitivity measure minimization problem, and show how to compute the optimal FWL realizations that minimize this measure.

It can be shown with (3), (5) and (6) that

$$\begin{aligned} \| \frac{\partial H(z)}{\partial A} \|_2^2 &= tr(W_A), \| \frac{\partial H(z)}{\partial B} \|_2^2 = tr(W_o), \\ \| \frac{\partial H(z)}{\partial C^T} \|_2^2 &= tr(W_c) \end{aligned} \tag{9}$$

where

$$W_A \triangleq \sum_{p=0}^{\infty} h(p)h^T(p), h(p) \triangleq \sum_{i+k=p} [(CA^i)^T][A^kB]^T \tag{10}$$

and (W_c, W_o) is the controllability and observability Gramian pair of (A, B, C).

If the realization (A, B, C), having $(W_c, W_o, h(p))$, is obtained from an initial realization (A_0, B_0, C_0), corresponding to $(W_c^0, W_o^0, h_0(p))$, by a similarity transformation T, we have the following equations

$$\begin{aligned} W_c &= T^{-1}W_c^0T^{-T}, W_o = T^TW_o^0T, \\ h(p) &= T^Th_0(p)T^{-T}. \end{aligned} \tag{11}$$

We can now express our new sensitivity measure M_{L2} explicitly as a function of the transformation matrix T or $P \triangleq TT^T$:

$$\begin{aligned} M_{L2}(T) &= tr\left(\sum_{p=0}^{\infty} Ph_0(p)P^{-1}h_0^T(p)\right) + tr(PW_o^0) \\ &+ tr(P^{-1}W_c^0) \triangleq R(P) \end{aligned} \tag{12}$$

The optimal FWL design problem is to identify those realizations that minimize M_{L2}:

$$\boxed{\min_{(A,B,C)\in S_H} M_{L2} \Longrightarrow (A, B, C)_{opt}.} \tag{13}$$

This is equivalent to

$$\min_{P>0} R(P) \Longrightarrow P_{opt}. \tag{14}$$

We now show how to construct such solution by the following theorem.

Theorem 1: For an asymptotically stable minimal system $H(z)$, $R(P)$ has a unique global minimum; the solution of (14) is the unique solution of

$$P\left[W_o^0 + \sum_{k=0}^{\infty} h_0(k)P^{-1}h_0^T(k)\right]P = W_c^0 + \sum_{k=0}^{\infty} h_0^T(k)Ph_0(k). \qquad (15)$$

■

Proof: For the detailed proof, we refer to [5]. ■

It seems impossible to derive an explicit expression of the optimal P from (15). The difficulty in finding the optimal realizations that minimize this measure can be overcome by using an analog computation via Gradient Flows (see [6] and [7]) and an iterative algorithm proposed in [5].

Comments :

1. T_{opt} is not unique even so is $P_{opt} = T_{opt}T_{opt}^T$. This degree of freedom can be used to obtain some additional performance improvements by the use of special forms, such as Schur or system Hessenberg forms [5].

2. We can show that the two optimal realization sets minimizing M_a and M_{L2}, respectively, are not identical. The latter is always bigger than the former. (see [5])

3 Optimal FWL Synthetic Design

For a fixed-point implementation with roundoff before multiplication [5], the actual computational model of the system is:

$$\begin{aligned} x'(t+1) &= A^*Q[x'(t)] + B^*u(t) \\ y'(t) &= C^*Q[x'(t)] + d^*u(t) \end{aligned} \qquad (16)$$

where $y'(t)$ is the actual output of the realization and $Q[.]$ denotes the quantization operation. The quantizer $Q[.]$ makes $Q[x]$ have a B_s bit expression where x has $(B_s + B_c)$ or more than $(B_s + B_c)$ bits. The input $u(t)$ is assumed to have already been rounded off to B_s bits.

We define the roundoff noise

$$e_x(t) \triangleq x'(t) - Q[x'(t)]. \qquad (17)$$

In numerical analysis, such a roundoff noise proces is modelled as independent white noise having zero-mean and a variance $\sigma_n^2 = (1/12)2^{-2Bs}$.

The degradation of the output of the realization due to a FWL implementation of the parameters in (A, B, C, d) and to a quantization of the states can be measured by the difference between the desired output of the plant, $y(t)$, and the actual output of the realization, $y'(t)$:

$$\begin{aligned} \Delta y(t) &= y(t) - y'(t) = [y(t) - y^*(t)] + [y^*(t) - y'(t)] \\ &\triangleq \Delta y^*(t) + \Delta y'(t). \end{aligned} \qquad (18)$$

In (18) we have conceptually separated the overall output error $\Delta y(t)$ into two errors $\Delta y^*(t) \triangleq y(t) - y^*(t)$ and $\Delta y'(t) \triangleq y^*(t) - y'(t)$, which account for the FWL effects on coefficients and on arithmetic operations, respectively. By a detailed analysis, we can show that $\Delta y'(t)$ can be interpreted as the output of a $MISO$ system excited by the roundoff noise $e_x(t)$ of the state $x'(t)$ and that $\Delta y_i^*(t)$, as the output of a system whose impulse response is the difference between the impulse responses of the ideal and the FWL transfer function excited by $u(t)$. Since $e_x(t)$ and $u(t)$ are independent, so are $\Delta y'(t)$ and $\Delta y_i^*(t)$. The steady-state variance of the plant output error $\Delta y(t)$ is therefore the sum of the variances of these two independent error signals :

$$\sigma^2 \triangleq \lim_{t\to\infty} E[\Delta y(t)^2] = \sigma_1^2 + \sigma_2^2 \qquad (19)$$

where $\sigma_1^2 \triangleq \lim_{t\to\infty} E[\Delta y^*(t)^2]$ and $\sigma_2^2 \triangleq \lim_{t\to\infty} E[\Delta y'(t)^2]$.

Without giving any detail, we mention that by adopting a statistical approach [5]

$$\sigma_1^2 = (M_{L2} + 1)\sigma_c^2 \qquad (20)$$

with $\sigma_c^2 = (1/12)2^{-2B_c}$, and

$$\sigma_2^2 = tr(W_o)\sigma_n^2. \qquad (21)$$

Denote by $\Sigma^2(T)$ the coordinate dependent part of the variance σ^2 of (19). Then one has, using (20) and (21):

$$\Sigma^2(T) = tr(W_o^*)\sigma_n^2 + M_{L2}\sigma_c^2, \qquad (22)$$

where σ_n^2 and σ_c^2 are the variances of the roundoff noise and of the statistical coefficient random noise both due to the FWL effects and dependent on the wordlengths of signals and coefficients, respectively.

One can define a measure $G_F(T)$ called *FWL Noise Gain (FNG)* as follows:

$$G_F(T) \triangleq M_{L2} + \frac{\sigma_n^2}{\sigma_c^2} tr(W_o) \triangleq M_{L2} + \rho^2 G. \quad (23)$$

A l_2-norm scaling is usually used in order to maintain the amplitudes of the states within an acceptable range which is determined by hardware, hence to reduce the probability of overflow. This leads to the following constraint on the realizations (or coordinate choice of the system) [1-2]:

$$W_c(i,i) = 1 \quad \forall i, \quad (24)$$

where W_c is the controllability Gramian of the realization as defined before.

So, the optimal FWL system design can be formulated as the following constrained minimization problem:

$$\min_{(A,B,C)\in S_H} G_F(T) = M_{L2} + \rho^2 G \quad (25)$$

with the constraint (24).

With $M_{10} = W_c$, $M_{20} = (1+\rho^2)W_o$ and $M_{30} = W_c$, the FWL noise gain can be rewritten as

$$G_F(T) = tr\left[\sum_{i=0}^{\infty} P h_o(i) P^{-1} h_0^T(i)\right] + tr(P^{-1} M_{10}) + tr(P M_{20}) \triangleq R(P) \quad (26)$$

with $P = TT^T$. It then follows from the above that

$$\min_{T:detT\neq 0} G_F(T) \Longleftrightarrow \min_{P:P>0} R(P) \quad (27)$$

both with the constraint

$$(T^{-1} M_{30} T^{-T})_{ii} = 1 \quad \forall i. \quad (28)$$

Theorem 2: (27)-(28) has a unique solution and

$$\sum_{i=0}^{\infty} \{h_0(i) P^{-1} h_0^T(i) - P^{-1} h_0^T(i) P h_0(i) P^{-1}\} + M_{20} - P^{-1} M_{10} P^{-1} - \lambda P^{-1} M_{30} P^{-1} = 0, \quad (29)$$

with

$$tr[M_{30} P^{-1}] - n = 0 \quad (30)$$

is the sufficient and necessary condition that is satisfied by this solution. ■

Proof: For the detailed proof, we refer to [5]. ■

An algorithm has been given in [5] for solving (29)-(30). And once we get the optimal P_{opt}, using a SVD of P_{opt} we can obtain the optimal transformation matrices T_{opt}. In [2], Hwang gave an algorithm to find such a T_{opt}.

4 Conclusions

We have investigated the optimal synthetic FWL design problem of state-space filters in this paper. Our contribution is twofold. The first one is to study a new sensitivity measure which is more natural and reasonable. The second one is to derive a measure that unifies the FWL effects. By minimizing these two measures, the optimal realization sets have been found, respectively. The necessary and sufficient conditions the optimal realizations have to satisfy are given for each case.

REFERENCES

[1] C.T. Mullis and R.A. Roberts (1976), "Filter structures which minimize roundoff noise in fixed-point digital filters", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 505-508.

[2] S.Y. Hwang (1977), "Minimum Uncorrelated Unit Noise in State-Space Digital Filtering", IEEE Trans. on Acoust. Speech and Signal Processing, Vol. ASSP-25, No 4 - Aug., pp.273-281.

[3] V. Tavsanoglu and L. Thiele (1984), "Optimal Design of State-Space Digital Filters by Simultaneous Minimization of Sensitivity and Roundoff Noise", IEEE Trans. on Circuits and Systems, Vol. CAS-31, No 10 - Oct., pp. 884-888.

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

[5] M. Gevers and G. Li (1992), "Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects", to be published by Springer-Verlag, Communication and Control Engineering Series, New York.

[6] U. Helmke and J.B. Moore (1991), "L^2 sensitivity minimization of linear system representations via gradient flows", submitted to J. of Mathematical Systems, Estimation and Control.

[7] J.E. Perkins, U. Helmke and J.B. Moore (1990), "Balanced Realizations via Gradient Flow Techniques," Systems and Control Letters 14, pp. 369-380.