Copyright © 2002 IFAC
15th Triennial World Congress, Barcelona, Spain

COVARIANCE CALCULATION FOR FLOATING POINT STATE SPACE REALIZATIONS

Sangho Ko [*,1] **Robert R. Bitmead** [*,1]

[*] *Department of Mechanical and Aerospace Engineering,*
University of California, San Diego, 9500 Gilman Drive,
La Jolla, California, 92093-0411,USA

Abstract: This paper provides a new method for analyzing floating point roundoff error for digital filters by using FSN (Finite Signal Noise) models whose noise sources have variances proportional to the the variance of the corrupted signals. With this model, the output error covariance of floating point arithmetic is derived and it is shown that optimal state space realization can be different from the optimal structure of the fixed point case. *Copyright ©2002 IFAC*

Keywords: Covariance, digital filter structures, roundoff noises, multiplicative noise, white noise

1. INTRODUCTION

From the 1970's, many researchers have conducted studies to minimize the errors in digital signal processing computations caused by finite word length effects. The finite wordlength effect may be divided into two categories (Mullis and Roberts, 1976) of *Coefficient error* and *Roundoff error.* Here, only the effect of roundoff errors will be considered.

The roundoff errors due to fixed point arithmetic are modelled by additive white noise sequences independent of the signal and with fixed variance. Mullis and Roberts (1976) and Hwang (1977) independently developed results on the properties of the output errors of the digital filters and determined the optimal fixed point state space realization.

Since the roundoff errors in floating point arithmetic are correlated with the signal that is quantized into finite precision number, the errors cannot be modelled by the standard white noise and the expression and the analysis of the errors is more complex compared to that of fixed point arithmetic. With this inherent complexity, the optimal state space realization in floating point arithmetic is known only for special cases. In case of double precision accumulation, it may be shown (Rao, 1992) that the optimal state space realization is similar in nature to that of fixed point arithmetic, and for the case of extended precision accumulation (a few additional mantissa bits, but not double length) Bomar *et al.* (1997) found that the floating point roundoff noise gain is identical in form to the fixed point gain. The previous work to optimize fixed point realization is directly applicable to the floating point realizations. In both papers, the state error covariance equation is expressed by a function of the state covariance in infinite precision, so their equations are not recursive and thereby the stability issues of the covariance equation could not be addressed.

Skelton (1994) introduced a new noise model for linear systems, so called "Finite-Signal-to-Noise Model", or simply the FSN model. Since this model assumes that the variance of the noise corrupting a signal is proportional to the variance

[1] Research supported by DARPA Grant N00014-00-1-0799 and USA National Science Foundation Grant ECS-0070146

of the signal, it is well suited to analyzing floating point roundoff error. Recently de Oliveira and Skelton (2001) provided necessary and sufficient conditions for mean square state feedback stabilization of linear systems with FSN model.

The FSN model reduces to the same mathematical problems as for the multiplicative noises. Existence conditions of state feedback stabilizability for linear systems with state and control dependent noise in continuous time were derived (Willems and Willems, 1976). More recently, a parametrization method was suggested for calculating exact stability bounds for systems with multiplicative noise (Sasagawa and Willems, 1996).

The problem of floating point arithmetic is analyzed in a different way using FSN model in the case of extended precision accumulation. The expression of the output error covariance is derived which is different from the previous floating point analysis. It is concluded that the optimal state space realization of floating point arithmetic can be significantly different from the fixed point case.

The outline of this paper is as follows. In Section 2, the definition and some stability results of linear systems with FSN models are discussed. Section 3 describes the effect of floating point roundoff errors on digital filters and formulates the filter output error covariance. Section 4 summarizes this paper. In this paper, matrices will be denoted by upper case boldface (e.g., $\mathbf{A}$), column matrices (vectors) will be denoted by lower case boldface (e.g.,$\mathbf{x}$), and scalar will be denoted by lower case (e.g., y) or upper case (e.g., Y). For a matrix $\mathbf{A}$, $\mathbf{A}^T$ denotes its transpose. For a symmetric matrices $\mathbf{P} > \mathbf{0}$ denotes the fact that $\mathbf{P}$ is positive definite.

2. FINITE SIGNAL TO NOISE MODEL

The formal definition of FSN model and stability results are described in (Skelton, 1994) and (Shi and Skelton, 1995). In this part only the result of the previous works will be presented.

The linear systems with "Finite Signal to Noise" Model are described by the state space equations

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k\,(1+n_k) + \mathbf{w}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k \\ \mathbf{u}_k &= \mathbf{K}\mathbf{x}_k \end{aligned} \qquad (1)$$

where, n_k and $\mathbf{w}_k$ are zero mean independent white noise sources, and all matrices and vectors are assumed to have appropriate dimensions. The closed loop state equation will be

$$\mathbf{x}_{k+1} = (\mathbf{A}+\mathbf{B}\mathbf{K})\,\mathbf{x}_k + \mathbf{B}\mathbf{K}\mathbf{x}_k n_k + \mathbf{w}_k. \qquad (2)$$

The state covariance equation is

$$\begin{aligned} \mathbf{X}_{k+1} = \;& (\mathbf{A}+\mathbf{B}\mathbf{K})\,\mathbf{X}_k\,(\mathbf{A}+\mathbf{B}\mathbf{K})^T \\ & + \sigma^2\mathbf{B}\mathbf{K}\mathbf{X}_k\mathbf{K}^T\mathbf{B}^T + \mathbf{W} \end{aligned} \qquad (3)$$

where, $\mathbf{X}_k \triangleq \mathcal{E}\left\{\mathbf{x}_k\mathbf{x}_k^T\right\}$, $\mathbf{W} \triangleq \mathcal{E}\left\{\mathbf{w}_k\mathbf{w}_k^T\right\}$, $\sigma^2 \triangleq \mathcal{E}\left\{n_k^2\right\}$ and $\mathcal{E}(\cdot)$ represents the expectation operator.

Since the effect of noise on the states increases as the input signal u_k increases, the systems with FSN model and stable $\mathbf{A}+\mathbf{B}\mathbf{K}$ can be destabilized in mean square due to noises, whereas the traditional white noise model cannot be destabilized by noise alone. This new model describes many physical systems more realistically than the traditional white noise model.

Following is the formal definition and condition for mean square stability of systems with FSN model.

Definition 1 (Mean Square Stability) (Shi and Skelton, 1995) *The FSN closed loop system* (1-3) *is mean square stable if its steady state covariance* $\mathbf{X}$ *exists and is positive definite.*

Theorem 1 (de Oliveira and Skelton, 2001) *There exists a controller gain* $\mathbf{K}$ *such that the closed loop FSN system* (1-3) *with noise power* σ^2 *is mean square stable only if the pair* $(\mathbf{A},\mathbf{B})$ *is stabilizable and* $\left(\sigma/\sqrt{1+\sigma^2}\right)\mathbf{A}$ *is stable.*

Remark 1 *Mean square stability of the closed loop FSN system is equivalent to the existence of a matrix* $\mathbf{P} > 0$ *satisfying the LMI* (*Linear Matrix Inequality*)

$$\begin{aligned} &(\mathbf{A}+\mathbf{B}\mathbf{K})\,\mathbf{P}(\mathbf{A}+\mathbf{B}\mathbf{K})^T - \mathbf{P} \\ &\qquad + \sigma^2\mathbf{B}\mathbf{K}\mathbf{P}\mathbf{K}^T\mathbf{B}^T < \mathbf{0}. \end{aligned} \qquad (4)$$

The two conditions of Theorem 1 can be expressed also by following two LMI conditions, respectively.

$$\begin{aligned} \mathbf{B}^{\perp}\left(\mathbf{A}\mathbf{P}\mathbf{A}^T - \mathbf{P}\right)\mathbf{B}^{\perp^T} < \mathbf{0} \\ \frac{\sigma^2}{1+\sigma^2}\mathbf{A}\mathbf{P}\mathbf{A}^T - \mathbf{P} < \mathbf{0} \end{aligned} \qquad (5)$$

A sufficient conditon for mean square stabilizability of the FSN system is that the two LMI conditions (5) *should be satisfied by the same positive definite matrix* $\mathbf{P}$*. The importance of the above LMI formulation of the stabilizability conditions lies in the availability associated computational tools to test for existence of and to compute the solution matrix* $\mathbf{P}$*.*

Example
Consider the following unstable discrete system.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 0.5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Since the matrix pair $(\mathbf{A}, \mathbf{B})$ is controllable, this system can be stabilized for any values of noise variance in a traditional additive white noise model. However, following Theorem 1, this system cannot be stabilized if the noise variance $\sigma^2 > 1/\sqrt{3}$ in the FSN model, since the matrix $\left(\sigma/\sqrt{1+\sigma^2}\right)\mathbf{A}$ becomes unstable.

3. FLOATING POINT ARITHMETIC

3.1 *Floating Point Arithmetic Noise*

The floating point binary representation of a number is given by

$$x = x_m 2^{x_e} \tag{6}$$

where x_m is a fractional part called the *mantissa*, and x_e is called the *exponent* (Gevers and Li, 1993). Typically, the mantissa is normalized such that $0.5 < |x_m| < 1$. Since the number is represented by a multiplication of the mantissa and the exponent, the resolution between two successive floating point numbers depends on the magnitude of these numbers, with the quantization error being proportional to this magnitude of both of these numbers. It is suggested that the noise due to floating point computation is largely due to the mantissa truncation and is therefore proportional to the signal amplitude. Hence, the quantization error in floating point arithmetic cannot be modelled by traditional additive white noise.

Floating point multiplication and addition roundoff errors can be described by (Bomar *et al.*, 1997)

$$\begin{array}{r}FL(x_1 x_2) = x_1 x_2 (1+\epsilon) \\ FL(x_1 + x_2) = (x_1 + x_2)(1+\delta)'\end{array} \tag{7}$$

where $FL(\cdot)$ denotes "floating point quantization" and ϵ and δ are white noises with zero mean value and the variances of, approximately

$$\sigma_\epsilon^2 \simeq \sigma_\delta^2 \simeq (0.18)2^{-2B} \tag{8}$$

where B is the number of mantissa bits.

The floating point digital signal processors in use today are all classified as single-precision devices, but internally perform register-to-register calculations with additional mantissa bits (Bomar *et al.*, 1997). For example, the Texas Instruments TMS320C30/C40 family of processors, the Analog Devices ADSP21020 family, and the AT&T DSP32 family all use a 32-bit mantissa for register-to-register operations, while the Motorola DSP96002 uses a 31-bit mantissa. Only the final result of a sum of products calculation is quantized back to the 24-bit-mantissa single-precision format. The mean and the variance of this final quantization are respectively zero and

$$\sigma_\eta^2 \simeq (0.167)2^{-2B'} \tag{9}$$

where, B' represents final mantissa bit(=24). Therefore we can consider only the final roundoff error, since $B > B'$ and $\sigma_\eta^2 \gg \sigma_\epsilon^2, \sigma_\delta^2$.

3.2 *Effects of Floating Point Errors on Digital Filters*

Digital filters are represented by the state equations

$$\begin{array}{rcl}\mathbf{x}_{k+1} &=& \mathbf{A}\mathbf{x}_k + \mathbf{B}u_k \\ y_k &=& \mathbf{C}\mathbf{x}_k + Du_k\end{array} \tag{10}$$

where, u_k is the scalar input, y_k is the scalar output, and x_k is the n-length state vector. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, D are $n \times n$, $n \times 1$, $1 \times n$, and 1×1 real constant matrices, respectively. Due to the floating point quantization, the actual filter is implemented as

$$\begin{array}{rcl}\hat{\mathbf{x}}_{k+1} &=& \tilde{\mathbf{A}}_k \hat{\mathbf{x}}_k + \tilde{\mathbf{B}}_k \hat{u}_k \\ \hat{y}_k &=& \tilde{\mathbf{C}}_k \hat{\mathbf{x}}_k + \tilde{D}_k \hat{u}_k\end{array} \tag{11}$$

where, $\hat{\mathbf{x}}_k$, $\hat{y}_k$, and $\hat{u}_k$ is the actual state, the actual output, and the actual input, respectively. And, the other matrices are defined as follows (Williamson, 1991):

$$\begin{array}{l}\tilde{\mathbf{A}}_k = [\alpha_{ij}(k)]\,,\ \tilde{\mathbf{B}}_k = [\beta_i(k)]\,,\ \tilde{\mathbf{C}}_k = [\bar{c}_j(k)]\,, \\ \tilde{D}_k = \tilde{D}_k.\end{array} \tag{12}$$

Here

$$\begin{array}{l}\alpha_{ij}(k) = \begin{cases} a_{ij}(1+\epsilon_{ij}) \displaystyle\prod_{p=1}^{n} (1+\delta_{ip})(1+\eta_i), \\ \qquad (for\ j = 1, 2) \\ a_{ij}(1+\epsilon_{ij}) \displaystyle\prod_{p=j-1}^{n} (1+\delta_{ip})(1+\eta_i), \\ \qquad (for\ j = 3, \ldots, n) \end{cases} \\ \beta_i(k) = b_i(1+\epsilon_{i,n+1})(1+\delta_{i,n})(1+\eta_i) \\ \bar{c}_j(k) = \begin{cases} c_j(1+\epsilon_{n+1,j}) \displaystyle\prod_{p=1}^{n} (1+\delta_{n+1,p})(1+\eta_{n+1}), \\ \qquad (for\ j = 1, 2) \\ c_j(1+\epsilon_{n+1,j}) \displaystyle\prod_{p=j-1}^{n} (1+\delta_{n+1,p})(1+\eta_{n+1}), \\ \qquad (for\ j = 3, \ldots, n) \end{cases} \\ \tilde{D}(k) = D(1+\epsilon_{n+1,n+1})(1+\delta_{n+1,n})(1+\eta_{n+1})\end{array} \tag{13}$$

where, $\epsilon_{ij} = \epsilon_{ij}(k)$, $\delta_{ij} = \delta_{ij}(k)$ are multiplication and addition errors and $\eta_i = \eta_i(k)$ are final quantization errors, all of which are zero mean independent white noises.

If the products of very small error terms are neglected in (13), then (11) is reduced to (14).

$$\begin{aligned}\hat{\mathbf{x}}_{k+1} &= (\mathbf{A}+\Delta\mathbf{A}_k)\hat{\mathbf{x}}_k + (\mathbf{B}+\Delta\mathbf{B}_k)\hat{u}_k \\ \hat{y}_k &= (\mathbf{C}+\Delta\mathbf{C}_k)\hat{\mathbf{x}}_k + (D+\Delta D_k)\hat{u}_k\end{aligned} \tag{14}$$

where,

$$\begin{aligned}\Delta\mathbf{A}_k &= \begin{bmatrix} a_{11}m_{11}(k) & \dots & a_{1n}m_{1n}(k) \\ \vdots & \ddots & \vdots \\ a_{n1}m_{n1}(k) & \dots & a_{nn}m_{nn}(k)\end{bmatrix} \\ \Delta\mathbf{B}_k &= \begin{bmatrix} b_1 m_{1,n+1}(k) \\ \vdots \\ b_n m_{n,n+1}(k)\end{bmatrix} \\ \Delta\mathbf{C}_k &= \begin{bmatrix} c_1 m_{n+1,1}(k) & \dots & c_n m_{n+1,n}(k)\end{bmatrix} \\ \Delta D_k &= D m_{n+1,n+1}(k).\end{aligned} \tag{15}$$

And,

$$\begin{aligned} m_{ij}(k) &= \begin{cases} \epsilon_{ij} + \sum_{p=1}^{n}\delta_{ip} + \eta_i, \\ \qquad (for\ j = 1,2) \\ \epsilon_{ij} + \sum_{p=j-1}^{n}\delta_{ip} + \eta_i, \\ \qquad (for\ j = 3,\dots,n)\end{cases} \\ m_{i,n+1}(k) &= \epsilon_{i,n+1} + \delta_{i,n} + \eta_i \\ m_{n+1,j}(k) &= \begin{cases} \epsilon_{n+1,j} + \sum_{p=1}^{n}\delta_{n+1,p} + \eta_{n+1}, \\ \qquad (for\ j = 1,2) \\ \epsilon_{n+1,j} + \sum_{p=j-1}^{n}\delta_{n+1,p} + \eta_{n+1}, \\ \qquad (for\ j = 3,\dots,n)\end{cases} \\ m_{n+1,n+1}(k) &= \epsilon_{n+1,n+1} + \delta_{n+1,n} + \eta_{n+1}.\end{aligned} \tag{16}$$

As discussed in Section 3.1, since the error of final quantization into 24-bit mantissa is much bigger than the other errors caused by intermediate register-to-register arithmetic, we can consider only the final roundoff errors $\eta_i(k)$. Then, the matrices in (15) can be reduced to the matrices in (17).

$$\begin{aligned}\Delta\mathbf{A}_k &= \begin{bmatrix} a_{11}\eta_1 & \dots & a_{1n}\eta_1 \\ \vdots & \ddots & \vdots \\ a_{n1}\eta_n & \dots & a_{nn}\eta_n\end{bmatrix} \\ &= \mathbf{diag}\{\ \eta_1,\ \eta_2,\ \dots,\ \eta_n\ \}\,\mathbf{A} \\ &= \left\{\sum_{i=1}^{n}\eta_i\mathbf{E}_i\right\}\mathbf{A} \\ \Delta\mathbf{B}_k &= \begin{bmatrix} b_1\eta_1 \\ \vdots \\ b_n\eta_n\end{bmatrix} = \mathbf{diag}\{\ \eta_1,\ \dots,\ \eta_n\ \}\,\mathbf{B} \\ &= \left\{\sum_{i=1}^{n}\eta_i\mathbf{E}_i\right\}\mathbf{B} \\ \Delta\mathbf{C}_k &= \begin{bmatrix} c_1\eta_{n+1} & \dots & c_n\eta_{n+1}\end{bmatrix} = \eta_{n+1}\mathbf{C} \\ \Delta D_k &= \eta_{n+1}D\end{aligned} \tag{17}$$

where, $\mathbf{E}_i$ is the *elementary matrix* that has “1” in the $i-i$ element and zero elsewhere. Therefore, we can express the actual state and output equation by (18).

$$\begin{aligned}\hat{\mathbf{x}}_{k+1} &= \left(\mathbf{I} + \sum_{i=1}^{n}\eta_i\mathbf{E}_i\right)\mathbf{A}\hat{\mathbf{x}}_k + \left(\mathbf{I} + \sum_{i=1}^{n}\eta_i\mathbf{E}_i\right)\mathbf{B}\hat{u}_k \\ \hat{y}_k &= (1+\eta_{n+1})\,\mathbf{C}\hat{\mathbf{x}}_k + (1+\eta_{n+1})\,D\hat{u}_k\end{aligned} \tag{18}$$

When the input is a white noise with variance of σ_u^2, the state covariance equation of (18) will be

$$\begin{aligned}\hat{\mathbf{X}}_{k+1} &= \mathbf{A}\hat{\mathbf{X}}_k\mathbf{A}^T + \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\hat{\mathbf{X}}_k\mathbf{A}^T\mathbf{E}_i \\ &\quad + \sigma_u^2\mathbf{B}\mathbf{B}^T + \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\end{aligned} \tag{19}$$

where, $\hat{\mathbf{X}}_k \triangleq \mathcal{E}\{\hat{\mathbf{x}}_k\hat{\mathbf{x}}_k^T\}$.

The second and fourth terms of the right hand side of (19) are caused by the floating point roundoff error. Similarly to systems with FSN models, the above recursion can be destabilized by the second term when the variance of floating point error is relatively greater than the stability margin of the matrix $\mathbf{A}$. Hence, the floating point errors can cause stability problems for systems like very narrow band filters where poles are very near to the unit circle.

3.3 *Calculation of Floating Point Output Error Covariance*

To find the expression of output error covariance we define the state error $\mathbf{e}_k$ and the output error Δy_k as in (20). Here it is assumed that $\hat{u}_k = u_k$. This is often the case in practice when the input itself has been generated by a finite wordlength device, and hence is known exactly.

$$\begin{aligned}\mathbf{e}_{k+1} &\triangleq \mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1} \\ &= \mathbf{A}\mathbf{e}_k - \sum_{i=1}^{n}\eta_i\mathbf{E}_i\mathbf{A}\hat{\mathbf{x}}_k - \sum_{i=1}^{n}\eta_i\mathbf{E}_i\mathbf{B}u_k \\ \\ \Delta y_k &\triangleq y_k - \hat{y}_k \\ &= \mathbf{C}\mathbf{e}_k - \eta_{n+1}\mathbf{C}\hat{\mathbf{x}}_k - \eta_{n+1}Du_k\end{aligned} \tag{20}$$

$$\begin{aligned}\mathbf{z}_{k+1} &= \bar{\mathbf{A}}\mathbf{z}_k + \bar{\mathbf{I}}\sum_{i=1}^{n}\eta_i\mathbf{E}_i\mathbf{A}\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{z}_k \\ &\quad + \bar{\mathbf{I}}\sum_{i=1}^{n}\eta_i\mathbf{E}_i\mathbf{B}u_k\end{aligned} \tag{21}$$

$$\Delta y_k = \begin{bmatrix}\mathbf{C} & -\eta_{n+1}\mathbf{C}\end{bmatrix}\mathbf{z}_k - \eta_{n+1}Du_k$$

where,

$$\mathbf{z}_k \triangleq \begin{bmatrix}\mathbf{e}_k\\ \hat{\mathbf{x}}_k\end{bmatrix}, \quad \bar{\mathbf{A}} \triangleq \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\end{bmatrix}, \quad \bar{\mathbf{I}} \triangleq \begin{bmatrix}-\mathbf{I}\\ \mathbf{I}\end{bmatrix}. \qquad (22)$$

Then, the state covariance equation of $\mathbf{z}_k$ and the output error covariance equation will be given by (23), when the input signal u_k is chosen as a white noise test signal with zero mean and variance of σ_u^2.

$$\begin{aligned}\mathbf{M}_{k+1} &= \bar{\mathbf{A}}\mathbf{M}_k\bar{\mathbf{A}}^T + \sigma_u^2\sigma_\eta^2\bar{\mathbf{I}}\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\bar{\mathbf{I}}^T\\ &\quad + \sigma_\eta^2\bar{\mathbf{I}}\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{M}_k\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\mathbf{A}^T\mathbf{E}_i\bar{\mathbf{I}}^T\end{aligned}$$

$$\begin{aligned}\Delta Y_k &\triangleq \mathcal{E}\left\{\Delta y_k^2\right\}\\ &= \mathbf{C}\bar{\mathbf{E}}_k\mathbf{C}^T + \sigma_\eta^2\mathbf{C}\bar{\mathbf{X}}_k\mathbf{C}^T + \sigma_u^2\sigma_\eta^2 D^2\end{aligned} \qquad (23)$$

where,

$$\begin{aligned}&\mathbf{M}_k \triangleq \mathcal{E}\left\{\mathbf{z}_k\mathbf{z}_k^T\right\} \triangleq \begin{bmatrix}\bar{\mathbf{E}}_k & \mathbf{Z}_k\\ \mathbf{Z}_k^T & \bar{\mathbf{X}}_k\end{bmatrix},\\ &\sigma_\eta^2 \triangleq \mathcal{E}\{\eta_i^2\}.\end{aligned} \qquad (24)$$

The above state covariance equation (23) can be divided into the following three coupled matrix equations.

$$\begin{aligned}\bar{\mathbf{E}}_{k+1} &= \mathbf{A}\bar{\mathbf{E}}_k\mathbf{A}^T + \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\bar{\mathbf{X}}_k\mathbf{A}^T\mathbf{E}_i\\ &\quad + \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\\ \bar{\mathbf{X}}_{k+1} &= \mathbf{A}\bar{\mathbf{X}}_k\mathbf{A}^T + \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\bar{\mathbf{X}}_k\mathbf{A}^T\mathbf{E}_i\\ &\quad + \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\\ \mathbf{Z}_{k+1} &= \mathbf{A}\mathbf{Z}_k\mathbf{A}^T - \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\bar{\mathbf{X}}_k\mathbf{A}^T\mathbf{E}_i\\ &\quad - \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\end{aligned} \qquad (25)$$

In (25), the stability depends only on $\bar{\mathbf{X}}_k$. That is, if $\bar{\mathbf{X}} \triangleq \lim_{k\to\infty}\bar{\mathbf{X}}_k$ exists, then $\mathbf{M}_\infty \triangleq \lim_{k\to\infty}\mathbf{M}_k$ exists. It can be shown that this existence condition is equivalent to the existence of a matrix $\mathbf{P} > \mathbf{0}$ satisfying the LMI

$$\begin{aligned}&\mathbf{A}\mathbf{P}\mathbf{A}^T - \mathbf{P} + \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\mathbf{P}\mathbf{A}^T\mathbf{E}_i\\ &\qquad + \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i < \mathbf{0}.\end{aligned} \qquad (26)$$

This LMI condition (26) can be used to determine the number of final mantissa bits that is required for stable realization of filters which have poles very near to the unit circle. If (26) is satisfied, then

$$\bar{\mathbf{X}} = \mathbf{A}\bar{\mathbf{X}}\mathbf{A}^T + \sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{A}\bar{\mathbf{X}}\mathbf{A}^T\mathbf{E}_i + \mathbf{Q}_1 \qquad (27)$$

where, $\mathbf{Q}_1 \triangleq \sigma_u^2\sigma_\eta^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i$.

For $\sigma_\eta^2 \ll 1$, (27) can be approximately written as

$$\bar{\mathbf{X}} = \mathbf{A}_p\bar{\mathbf{X}}\mathbf{A}_p^T + \mathbf{Q}_1 = \sum_{i=0}^{\infty}\mathbf{A}_p^i\mathbf{Q}_1\left(\mathbf{A}_p^T\right)^i, \qquad (28)$$

where,

$$\mathbf{A}_p \triangleq \sqrt{1+\sigma_\eta^2}\;\mathbf{A}. \qquad (29)$$

Substituting (28) into the steady state version of (25) yields

$$\bar{\mathbf{E}} = \sum_{i=0}^{\infty}\mathbf{A}_p^i\mathbf{Q}\left(\mathbf{A}_p^T\right)^i \qquad (30)$$

where,

$$\begin{aligned}\mathbf{Q} &\triangleq \sigma_\eta^2\sum_{j=1}^{n}\mathbf{E}_j\mathbf{A}\left\{\sum_{l=0}^{\infty}\mathbf{A}_p^l\mathbf{Q}_1\left(\mathbf{A}_p^T\right)^l\right\}\mathbf{A}^T\mathbf{E}_j + \mathbf{Q}_1\\ &= \sigma_\eta^4\sigma_u^2\sum_{j=1}^{n}\mathbf{E}_j\mathbf{A}\left\{\sum_{l=0}^{\infty}\mathbf{A}_p^l\mathbf{V}\left(\mathbf{A}_p^T\right)^l\right\}\mathbf{A}^T\mathbf{E}_j\\ &\quad + \sigma_\eta^2\sigma_u^2\sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i\end{aligned} \qquad (31)$$

where,

$$\mathbf{V} = \sum_{i=1}^{n}\mathbf{E}_i\mathbf{B}\mathbf{B}^T\mathbf{E}_i. \qquad (32)$$

Therefore, the steady state covariance of output error is expressed by (33).

$$\begin{aligned}\Delta Y &\triangleq \lim_{k\to\infty}\mathcal{E}\left\{\Delta y^2(k)\right\}\\ &= \mathbf{C}\bar{\mathbf{E}}\mathbf{C}^T + \sigma_\eta^2\mathbf{C}\bar{\mathbf{X}}\mathbf{C}^T + \sigma_u^2\sigma_\eta^2 D^2\\ &= \sigma_\eta^4\sigma_u^2\mathbf{C}\sum_{s=0}^{\infty}\mathbf{A}^s\sum_{j=1}^{n}\mathbf{E}_j\mathbf{A}\left\{\sum_{l=0}^{\infty}\mathbf{A}_p^l\mathbf{V}\right.\\ &\qquad\left.\left(\mathbf{A}_p^T\right)^l\right\}\mathbf{A}^T\mathbf{E}_j\left(\mathbf{A}^T\right)^s\mathbf{C}^T\\ &\quad + \sigma_\eta^2\sigma_u^2\mathbf{C}\sum_{s=0}^{\infty}\mathbf{A}^s\mathbf{V}\left(\mathbf{A}^T\right)^s\mathbf{C}^T\\ &\quad + \sigma_\eta^4\sigma_u^2\mathbf{C}\sum_{s=0}^{\infty}\mathbf{A}^s\mathbf{V}\left(\mathbf{A}^T\right)^s\mathbf{C}^T\\ &\quad + \sigma_u^2\sigma_\eta^2 D^2\end{aligned} \qquad (33)$$

Since usually $\sigma_\eta^4 \ll \sigma_\eta^2$, σ_η^4-terms in (33) can be neglected. Then, the output error covariance will be (34).

$$\Delta Y = \sigma_\eta^2 \sigma_u^2 \left[\mathbf{C} \sum_{s=0}^{\infty} \mathbf{A}^s \mathbf{V} \left(\mathbf{A}^T\right)^s \mathbf{C}^T + D^2 \right] \tag{34}$$

Or,

$$\begin{aligned} \Delta Y &= \sigma_\eta^2 \sigma_u^2 \left[\mathbf{tr}\left\{\mathbf{VW}\right\} + D^2 \right] \\ &= \sigma_\eta^2 \sigma_u^2 \left[\sum_{i=1}^{n} W_{ii} b_i^2 + D^2 \right] \end{aligned} \tag{35}$$

where,

$$\mathbf{W} = \sum_{i=0}^{\infty} \left(\mathbf{A}^T\right)^i \mathbf{C}^T \mathbf{C} \mathbf{A}^i, \tag{36}$$

and W_{ii} is the $i-i$th element of $\mathbf{W}$ and b_i is the i-th element of $\mathbf{B}$. This result (35) can be formally stated as following Theorem 2.

Theorem 2 *For a given infinite precision digital filter represented by*

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}u_k \\ y_k &= \mathbf{C}\mathbf{x}_k + Du_k, \end{aligned} \tag{37}$$

the steady state covariance of output error,ΔY*, is given by*

$$\Delta Y = \sigma_\eta^2 \sigma_u^2 \left[\sum_{i=1}^{n} W_{ii} b_i^2 + D^2 \right] \tag{38}$$

for a white noise input signal of variance σ_u^2 *when the filter* (37) *is implemented in digital signal processor utilizing extended precision arithmetic for internal register-to-register, where* W_{ii} *is the* $i-i$*th element of the observability Gramian* $\mathbf{W}$ *of matrix pair* $(\mathbf{A}, \mathbf{C})$ *,* b_i *is the* i*-th element of* $\mathbf{B}$*, and* σ_η^2 *is the variance of the final quantization error given by* (9)*.*

The expression of output error covariance (35) is different from that of Bomar *et al.*(1997),

$$\Delta Y = \sigma_\eta^2 \sigma_u^2 \left[\mathbf{CKC}^T + D^2 + \sum_{i=1}^{n} K_{ii} W_{ii} \right] \tag{39}$$

where, $\mathbf{K}$ is the controllability Gramian of matrix pair $(\mathbf{A}, \mathbf{B})$. In (39), because only the third term in the bracket depends on the realization and is the fixed point roundoff noise gain, Bomar *et al.* (1997) concluded that the optimal state space digital filters realized on floating point digital signal processors are the same as that of fixed point case. Hence, this difference between (38) and (39) points to the fact that the optimal state space realization of floating point arithmetic can be different from the structure of fixed point arithmetic. Finding optimal realizations in floating point arithmetic requires more study.

4. CONCLUSION

The floating point roundoff errors coming from implementing digital filters and controllers appear in the form of multiplicative noises, so the traditional method of using additive white noise cannot be applied to analyze the effect of the errors. In this paper, the floating point error effect on digital filters was analyzed by using the newly introduced FSN models which have noise sources whose variances are linearly proportional to the variances of the corrupted signal. With this new model, a new expression for the output error covariance of digital filters was derived when implemented in floating point digital signal processor using extended precision and it was concluded that the optimal state space digital filter realization of floating point arithmetic can be different from that of fixed point arithmetic.

5. REFERENCES

Bomar, B. W., L. M. Smith and R. D.Joseph (1997). Roundoff noise analysis of state space digital filters implemented on floating point digital signal processors. *IEEE Trans. Circuits Syst.* **44**, 952–955.

de Oliveira, M. C. and R. E. Skelton (2001). State feedback control of linear systems in the presence of devices with finite signal-to-noise ratio. *Int. J. Control* **74**, 1501–1509.

Gevers, M. and G. Li (1993). *Parametrizations in control, estimation and filtering problems.* Springer-Verag. London.

Mullis, C. T. and R. A. Roberts (1976). Synthesis of minimum roundoff noise fixed point digital filters. *IEEE Trans. Circuits Syst.* **CAS-23**, 551–562.

Rao, B. D. (1992). Floating point arithmetic and digital filters. *IEEE Trans. Signal Processing* **40**, 85–95.

Sasagawa, T. and J. L. Willems (1996). Parametrization mathod for calculating exact stability bounds of stochastic linear systems with multiplicative noise. *Automatica* **32**, 1741–1747.

Shi, G. and R. E. Skelton (1995). State feedback covariance control for linear finite signal-to-noise ratio models. In: *Proceedings of the 34th Conference on Decision and Control.* New Orleans, LA, USA. pp. 3423–3428.

Skelton, R. E. (1994). Robust control of aerospace systems. In: *Proceedings IFAC Symposium on Robust Control Design.* Rio de Janeiro, Brazil. pp. 24–32.

Willems, J. L. and J. C. Willems (1976). Feedback stabilizability for stochastic systems with state and control dependent noise. *Automatica* **12**, 277–283.

Williamson, D. (1991). *Digital control and implementatation.* Prentice Hall. Australia.