

Covariance Calculation for Floating-Point State-Space Realizations

Sangho Ko and Robert R. Bitmead, *Fellow, IEEE*

Abstract—This paper provides a new method for analyzing floating-point roundoff error for digital filters by using “finite signal-to-noise” models whose noise sources have variances proportional to the variance or power of the corrupted signals. With this model, a new expression for output error covariance of floating-point arithmetic is derived in the case of double or extended precision accumulation. The output error covariance shows that the optimal state space realization for floating point is the same as that of the fixed-point case, except for two cases: when the filter has poles extremely close to the unit circle or when final quantization to precisions shorter than single precision is employed. An explicit formula is found for determining the minimum number of mantissa bits for stable realization.

Index Terms—Covariance, multiplicative white noises, optimal state space realizations, roundoff noises.

I. INTRODUCTION

SINCE the 1970s, many researchers have conducted studies to minimize the errors in digital signal processing computations caused by finite wordlength effects. The finite wordlength effect may be divided into two categories of coefficient error and roundoff error [1]. Here, only the effect of roundoff errors will be considered.

Roundoff errors due to fixed point arithmetic are modeled by additive white noise sequences independent of the signal and with fixed variance. Mullis and Roberts [1] and Hwang [2] independently developed results on the properties of 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 a finite precision number, the errors cannot be modeled by standard white noise and the expression and the analysis of the errors are 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 the case of double-precision accumulation, it may be shown [3] 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.* [4] 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 in finite precision is expressed as 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. We apply tools of multiplicative noise processes to extend and refine these earlier results.

Skelton [5] introduced a new noise model for linear systems (the so-called “finite signal-to-noise model” or, simply, the FSN model). Since this model assumes that the variance of the additive noise corrupting a signal is proportional to the variance of the signal, it is well suited to analyzing floating-point roundoff error. One important property of systems with FSN models is that they can be destabilized in mean square due to noises [5], whereas the traditional additive white noise model cannot be destabilized by noise alone. This new model describes many physical systems more realistically than the traditional additive white noise model. Recently de Oliveira and Skelton [6] provided two linear matrix inequalities (LMIs), which are necessary and sufficient conditions for mean square state feedback stabilization of linear systems with FSN models. The FSN model reduces to the same mathematical problem as for multiplicative noises. Existence conditions for state feedback stabilizability for linear systems with state and control dependent noise in continuous time were derived [7], and a parametrization method was suggested for calculating exact stability bounds for systems with multiplicative noise [8].

The problem of floating-point arithmetic is analyzed in a different way using FSN models in the case of extended or double-precision accumulation. An expression for the output error covariance is derived by solving a recursive state covariance equation. Therefore, stability problems are considered and an expression for the upper bound of final roundoff noise for stable filter realization is derived, which can be used for determining the minimum number of final mantissa bits for stability. The new expression for output error covariance compensates for the approximation which was made in the previous studies and it is reconfirmed that the optimal state space realization of floating-point arithmetic is the same as that of fixed point, except in some special cases.

In this paper, matrices will be denoted by upper case boldface (e.g., \mathbf{A}) or calligraphic uppercase (e.g., \mathcal{A}), column matrices (vectors) will be denoted by lower case boldface (e.g., \mathbf{x}), and scalars 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, and $\text{Vec}(\mathbf{A})$ denotes the vector constructed by stacking all of the columns of \mathbf{A} , the second below the first, and so on. For a symmetric matrices $\mathbf{P} > \mathbf{0}$ or $\mathbf{P} \geq \mathbf{0}$ denotes the fact that \mathbf{P} is positive definite or positive semi-definite, respectively.

Manuscript received December 30, 2002; revised October 23, 2003. This work was supported by the United States National Science Foundation under Grant ECS-0200449 and DARPA under Grant N00014-00-1-0799. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Trac D. Tran.

The authors are with the Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093-0411 USA (e-mail: sko@ucsd.edu; rbitmead@ucsd.edu).

Digital Object Identifier 10.1109/TSP.2004.837439

II. FLOATING-POINT ARITHMETIC

A. Floating-Point Arithmetic Noise

The floating-point binary representation of a number is given by

$$x = x_m 2^{x_e} \quad (1)$$

where x_m is a fractional part called the mantissa, and an integer x_e is called the exponent [9]. Typically, the mantissa is normalized such that $0.5 \leq |x_m| < 1$. 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 modeled by traditional additive white noise.

Floating-point multiplication and addition roundoff errors can be described by [10]

$$\begin{aligned} FL(x_1 x_2) &= x_1 x_2 (1 + \epsilon) \\ FL(x_1 + x_2) &= (x_1 + x_2) (1 + \delta) \end{aligned} \quad (2)$$

where $FL(\cdot)$ denotes ‘‘floating-point quantization,’’ and ϵ and δ are white noises with zero mean value and the variances of approximately [11]

$$\sigma_\epsilon^2 \simeq \sigma_\delta^2 \simeq (0.18)2^{-2B} \quad (3)$$

where B is the number of mantissa bits.

B. Final Roundoff Error [4]

Many floating-point digital signal processors in use today are classified as single-precision devices, but internally perform register-to-register calculations with additional mantissa bits. 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, whereas 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. Similarly to (2), final quantization error η can be modeled by

$$FQ(x) = x(1 + \eta) \quad (4)$$

where $FQ(\cdot)$ represents ‘‘final quantization,’’ and η is zero mean white noise with variance given by

$$\sigma_\eta^2 \simeq (0.167)2^{-2B'}. \quad (5)$$

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

C. Effects of Floating-Point Errors on Digital Filters

Digital filters are represented by the state equations

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}u_k \\ y_k &= \mathbf{C}\mathbf{x}_k + Du_k \end{aligned} \quad (6)$$

where u_k is the scalar input, y_k is the scalar output, and \mathbf{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 and final quantization, the actual filter is implemented as

$$\begin{aligned} \hat{\mathbf{x}}_{k+1} &= FQ[FL(FL(\mathbf{A}\hat{\mathbf{x}}_k) + FL(\mathbf{B}\hat{u}_k))] \\ &\triangleq \tilde{\mathbf{A}}_k \hat{\mathbf{x}}_k + \tilde{\mathbf{B}}_k \hat{u}_k \\ \hat{y}_k &= FQ[FL(FL(\mathbf{C}\hat{\mathbf{x}}_k) + FL(D\hat{u}_k))] \\ &\triangleq \tilde{\mathbf{C}}_k \hat{\mathbf{x}}_k + \tilde{D}_k \hat{u}_k \end{aligned} \quad (7)$$

where $\hat{\mathbf{x}}_k$, \hat{y}_k , and \hat{u}_k is the actual state, the actual output, and the actual input, respectively.

If the order of additions used in the inner product operation in (7) is assumed to be from left to right as in [3] and [4], $\tilde{\mathbf{A}}_k$, $\tilde{\mathbf{B}}_k$, $\tilde{\mathbf{C}}_k$, and \tilde{D}_k are given by

$$\tilde{\mathbf{A}}_k = [\alpha_{ij}(k)], \quad \tilde{\mathbf{B}}_k = [\beta_i(k)], \quad \tilde{\mathbf{C}}_k = [\bar{c}_j(k)], \quad \tilde{D}_k = \tilde{D}_k \quad (8)$$

where

$$\begin{aligned} \alpha_{ij}(k) &= \begin{cases} a_{ij}(1 + \epsilon_{ij}) \\ \quad \times \prod_{p=1}^n (1 + \delta_{ip})(1 + \eta_i), & j=1, 2 \\ a_{ij}(1 + \epsilon_{ij}) \\ \quad \times \prod_{p=j-1}^n (1 + \delta_{ip})(1 + \eta_i), & j=3, \dots, n \end{cases} \\ \beta_i(k) &= b_i(1 + \epsilon_{i,n+1})(1 + \delta_{i,n})(1 + \eta_i) \\ \bar{c}_j(k) &= \begin{cases} c_j(1 + \epsilon_{n+1,j}) \prod_{p=1}^n (1 + \delta_{n+1,p}) \\ \quad \times (1 + \eta_{m+1}), & j=1, 2 \\ c_j(1 + \epsilon_{n+1,j}) \prod_{p=j-1}^n (1 + \delta_{n+1,p}) \\ \quad \times (1 + \eta_{m+1}), & j=3, \dots, n \end{cases} \\ \tilde{D}(k) &= D(1 + \epsilon_{n+1,n+1})(1 + \delta_{n+1,n})(1 + \eta_{m+1}). \end{aligned} \quad (9)$$

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

As used in [4], ignoring the extremely small products of error terms in (9) yields

$$\begin{aligned} \hat{\mathbf{x}}_{k+1} &= (\mathbf{A} + \Delta\mathbf{A}_k)\hat{\mathbf{x}}_k + (\mathbf{B} + \Delta\mathbf{B}_k)\hat{u}_k \\ \hat{y}_k &= (\mathbf{C} + \Delta\mathbf{C}_k)\hat{\mathbf{x}}_k + (D + \Delta D_k)\hat{u}_k \end{aligned} \quad (10)$$

where

$$\begin{aligned} \Delta\mathbf{A}_k &= \begin{bmatrix} a_{11}m_{11}(k) & \dots & a_{1n}m_{1n}(k) \\ \vdots & \ddots & \vdots \\ a_{n1}m_{n1}(k) & \dots & a_{nn}m_{nn}(k) \end{bmatrix} \\ \Delta\mathbf{B}_k &= \begin{bmatrix} b_1m_{1,n+1}(k) \\ \vdots \\ b_nm_{n,n+1}(k) \end{bmatrix} \\ \Delta\mathbf{C}_k &= [c_1m_{n+1,1}(k) \quad \dots \quad c_nm_{n+1,n}(k)] \\ \Delta D_k &= Dm_{n+1,n+1}(k) \end{aligned} \quad (11)$$

and

$$\begin{aligned}
m_{ij}(k) &= \begin{cases} \epsilon_{ij} + \sum_{p=1}^n \delta_{ip} + \eta_i, & j = 1, 2 \\ \epsilon_{ij} + \sum_{p=j-1}^n \delta_{ip} + \eta_i, & j = 3, \dots, n \end{cases} \\
m_{i,n+1}(k) &= \epsilon_{i,n+1} + \delta_{i,n} + \eta_i \\
m_{n+1,j}(k) &= \begin{cases} \epsilon_{n+1,j} + \sum_{p=1}^n \delta_{n+1,p} \\ \quad + \eta_{n+1}, & j = 1, 2 \\ \epsilon_{n+1,j} + \sum_{p=j-1}^n \delta_{n+1,p} \\ \quad + \eta_{n+1}, & j = 3, \dots, n \end{cases} \\
m_{n+1,n+1}(k) &= \epsilon_{n+1,n+1} + \delta_{n+1,n} + \eta_{n+1}. \quad (12)
\end{aligned}$$

As discussed in Section II-B, 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 $\eta_i(k)$ [4]. Then, the matrices in (11) can be reduced to

$$\begin{aligned}
\Delta \mathbf{A}_k &= \begin{bmatrix} a_{11}\eta_1 & \cdots & a_{1n}\eta_1 \\ \vdots & \ddots & \vdots \\ a_{n1}\eta_n & \cdots & a_{nn}\eta_n \end{bmatrix} \\
&= \left\{ \sum_{i=1}^n \eta_i \mathbf{E}_i \right\} \mathbf{A} \\
\Delta \mathbf{B}_k &= \begin{bmatrix} \eta_1 b_1 \\ \vdots \\ \eta_n b_n \end{bmatrix} \\
&= \left\{ \sum_{i=1}^n \eta_i \mathbf{E}_i \right\} \mathbf{B} \\
\Delta \mathbf{C}_k &= [c_1 \eta_{n+1} \quad \cdots \quad c_n \eta_{n+1}] \\
&= \eta_{n+1} \mathbf{C} \\
\Delta D_k &= \eta_{n+1} D \quad (13)
\end{aligned}$$

where \mathbf{E}_i is the $n \times n$ square elementary matrix that has “1” in the $i - i$ element and zero elsewhere. [By ignoring the roundoff errors caused by intermediate register-to-register arithmetic, the assumed order of additions in (9) does not affect the result.]

Therefore, we can express the actual state and output equation by

$$\begin{aligned}
\hat{\mathbf{x}}_{k+1} &= \left(\mathbf{I} + \sum_{i=1}^n \eta_i \mathbf{E}_i \right) \mathbf{A} \hat{\mathbf{x}}_k + \left(\mathbf{I} + \sum_{i=1}^n \eta_i \mathbf{E}_i \right) \mathbf{B} \hat{u}_k \\
\hat{y}_k &= (1 + \eta_{n+1}) \mathbf{C} \hat{\mathbf{x}}_k + (1 + \eta_{n+1}) D \hat{u}_k. \quad (14)
\end{aligned}$$

D. 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 (16) and (18).

$$\mathbf{e}_{k+1} \triangleq \mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1} \quad (15)$$

$$= \mathbf{A} \mathbf{e}_k - \sum_{i=1}^n \eta_i \mathbf{E}_i \mathbf{A} \hat{\mathbf{x}}_k - \sum_{i=1}^n \eta_i \mathbf{E}_i \mathbf{B} u_k \quad (16)$$

$$\Delta y_k \triangleq y_k - \hat{y}_k \quad (17)$$

$$= \mathbf{C} \mathbf{e}_k - \eta_{n+1} \mathbf{C} \hat{\mathbf{x}}_k - \eta_{n+1} D u_k. \quad (18)$$

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 [9].

When the input is a zero mean white noise sequence with variance of σ_u^2 in (14) and (16), the state and state error covariance equation will be given by

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

$$\begin{aligned}
\bar{\mathbf{E}}_{k+1} &= \mathbf{A} \bar{\mathbf{E}}_k \mathbf{A}^T + \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \bar{\mathbf{X}}_k \mathbf{A}^T \mathbf{E}_i \\
&\quad + \sigma_u^2 \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{B} \mathbf{B}^T \mathbf{E}_i \quad (20)
\end{aligned}$$

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

From (19) and (20), finiteness of both covariances depends only on $\bar{\mathbf{X}}_k$. That is, if $\bar{\mathbf{X}} \triangleq \lim_{k \rightarrow \infty} \bar{\mathbf{X}}_k$ exists, then $\bar{\mathbf{E}} \triangleq \lim_{k \rightarrow \infty} \bar{\mathbf{E}}_k$ also exists.

The second and third terms of the right-hand side of (19) are caused by the floating-point roundoff error. The state covariance recursion can be destabilized by the second term when the variance of the 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. Since $\sigma_\eta^2 \ll 1$, the third term on the right-hand side of $\bar{\mathbf{X}}_k$ in (19) can be neglected compared to the fourth term. [We note that because (19) is linear, the effect of this term is to increase the covariance $\bar{\mathbf{X}}_k$ proportionally with the magnitude of this neglected term relative to the final term. It cannot affect stability of the recursion.] Hence, (19) is reduced to

$$\bar{\mathbf{X}}_{k+1} = \mathbf{A} \bar{\mathbf{X}}_k \mathbf{A}^T + \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \bar{\mathbf{X}}_k \mathbf{A}^T \mathbf{E}_i + \sigma_u^2 \mathbf{B} \mathbf{B}^T. \quad (21)$$

Equation (21) can be reexpressed by using Kronecker products [12] as

$$\begin{aligned}
\bar{\mathbf{x}}_{k+1} &= \left[(\mathbf{A} \otimes \mathbf{A}) + \sigma_\eta^2 \sum_{i=1}^n (\mathbf{E}_i \otimes \mathbf{E}_i) (\mathbf{A} \otimes \mathbf{A}) \right] \bar{\mathbf{x}}_k + \mathbf{q} \\
&= \mathcal{A} \bar{\mathbf{x}}_k + \mathbf{q} \quad (22)
\end{aligned}$$

where $\mathcal{A} \triangleq [(\mathbf{A} \otimes \mathbf{A}) + \sigma_\eta^2 \sum_{i=1}^n (\mathbf{E}_i \otimes \mathbf{E}_i) (\mathbf{A} \otimes \mathbf{A})]$, $\bar{\mathbf{x}}_k \triangleq \text{Vec}(\bar{\mathbf{X}}_k)$, and $\mathbf{q} \triangleq \text{Vec}(\sigma_u^2 \mathbf{B} \mathbf{B}^T)$. Then, mean square stability of the state covariance $\bar{\mathbf{X}}_k$ is equivalent to $\rho(\mathcal{A}) < 1$, where $\rho(\cdot)$ represents the spectral radius of a matrix. The following lemma gives equivalent conditions for mean square stability of (21).

Lemma 1: Consider the system represented by (22), and suppose (\mathbf{A}, \mathbf{B}) is a controllable pair with \mathbf{A} having all eigenvalues inside the unit circle. Denote $\mathcal{A}_1 \triangleq \mathbf{I} - (\mathbf{A} \otimes \mathbf{A})$ and $\mathcal{A}_2 \triangleq \sigma_\eta^2 \sum_{i=1}^n (\mathbf{E}_i \otimes \mathbf{E}_i) (\mathbf{A} \otimes \mathbf{A})$. Then, $\rho(\mathcal{A}) < 1$ if and only if $\rho(\mathcal{A}_2 \mathcal{A}_1^{-1}) < 1$.

Proof: Suppose $\rho(\mathcal{A}) < 1$. Then, $\bar{\mathbf{X}} = \lim_{k \rightarrow \infty} \bar{\mathbf{X}}_k$ is finite, non-negative definite and satisfies the following algebraic equation:

$$\bar{\mathbf{X}} = \mathbf{A}\bar{\mathbf{X}}\mathbf{A}^T + \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \bar{\mathbf{X}} \mathbf{A}^T \mathbf{E}_i + \sigma_u^2 \mathbf{B} \mathbf{B}^T. \quad (23)$$

Further, if $\lim_{k \rightarrow \infty} \bar{\mathbf{X}}_k$ is finite, then $\rho(\mathcal{A}) < 1$. We next establish positivity and monotonicity properties of $\bar{\mathbf{X}}$ as a function of σ_η^2 .

Consider two values $\sigma_\eta^2 = \sigma_1^2$, $\sigma_\eta^2 = \sigma_2^2$ with $\sigma_1^2 \geq \sigma_2^2$ for which $\bar{\mathbf{X}}_k$ is finite and converges to limits denoted by $\bar{\mathbf{X}}_{\sigma_1}$ and $\bar{\mathbf{X}}_{\sigma_2}$, respectively. Then, the difference $\tilde{\mathbf{X}}_\sigma \triangleq \bar{\mathbf{X}}_{\sigma_1} - \bar{\mathbf{X}}_{\sigma_2}$ satisfies

$$\begin{aligned} \tilde{\mathbf{X}}_\sigma &= \mathbf{A}\tilde{\mathbf{X}}_\sigma\mathbf{A}^T + \sigma_1^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \tilde{\mathbf{X}}_\sigma \mathbf{A}^T \mathbf{E}_i \\ &\quad + (\sigma_1^2 - \sigma_2^2) \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \bar{\mathbf{X}}_{\sigma_2} \mathbf{A}^T \mathbf{E}_i. \end{aligned} \quad (24)$$

The non-negativity of the driving term of (24) and $\rho(\mathcal{A}) < 1$ ensures that $\tilde{\mathbf{X}}_\sigma \geq \mathbf{0}$. Therefore

$$\sigma_1^2 \geq \sigma_2^2 \Rightarrow \bar{\mathbf{X}}_{\sigma_1} \geq \bar{\mathbf{X}}_{\sigma_2}. \quad (25)$$

This is the desired monotonicity result.

When $\sigma_\eta = 0$, $\tilde{\mathbf{X}}_\sigma$ satisfies $\tilde{\mathbf{X}}_\sigma = \mathbf{A}\tilde{\mathbf{X}}_\sigma\mathbf{A}^T + \sigma_u^2 \mathbf{B} \mathbf{B}^T$ with \mathbf{A} stable and (\mathbf{A}, \mathbf{B}) controllable. Thus, $\tilde{\mathbf{X}}_\sigma > \mathbf{0}$ [13]. By monotonicity, we have

$$\bar{\mathbf{X}}_{\sigma_1} \geq \bar{\mathbf{X}}_{\sigma_2} > \mathbf{0}. \quad (26)$$

The monotonicity of $\bar{\mathbf{X}}$ with respect to σ_η^2 implies that either $\lim_{k \rightarrow \infty} \bar{\mathbf{X}}_k$ is finite for all σ_η^2 (which is demonstrably untrue [6]), or there exists a σ_{sup}^2 such that $\sigma_\eta^2 < \sigma_{\text{sup}}^2$ implies that $\lim_{k \rightarrow \infty} \bar{\mathbf{X}}_k$ is finite and positive definite, and $\sigma_\eta^2 \geq \sigma_{\text{sup}}^2$ implies that $\{\bar{\mathbf{X}}_k\}$ is unbounded. Thus, $\sigma_\eta^2 < \sigma_{\text{sup}}^2$ implies $\rho(\mathcal{A}) < 1$, and $\sigma_\eta^2 \leq \sigma_{\text{sup}}^2$ implies $\rho(\mathcal{A}) \geq 1$. By continuity of the eigenvalues of \mathcal{A} with σ_η^2 , we have at $\sigma_\eta^2 = \sigma_{\text{sup}}^2$, $\rho(\mathcal{A}) = 1$.

We have that if a solution to (23) exists, then it is given by $\bar{\mathbf{x}} \triangleq \text{Vec}(\bar{\mathbf{X}}) = (\mathbf{I} - \mathcal{A})^{-1} \mathbf{q}$. We know further, from the monotonicity argument, that this $\bar{\mathbf{x}}$ exists for any $\sigma_\eta^2 < \sigma_{\text{sup}}^2$ but is discontinuous at $\sigma_\eta^2 = \sigma_{\text{sup}}^2$. [We actually know that $\bar{\mathbf{X}}$, which is the solution to (23), has a different number of positive eigenvalues either side of σ_{sup}^2 , and from (26), we also know that for $\sigma_\eta^2 < \sigma_{\text{sup}}^2$, $\bar{\mathbf{X}}$ is bounded below.] This implies that the matrix $(\mathbf{I} - \mathcal{A})$ is singular at this point and so $\lambda(\mathcal{A}) = 1$. Then, from the relation $(\mathbf{I} - \mathcal{A}) = (\mathbf{I} - \mathcal{A}_2 \mathcal{A}_1^{-1}) \mathcal{A}_1$, the matrix $(\mathbf{I} - \mathcal{A}_2 \mathcal{A}_1^{-1})$ is singular with invertible \mathcal{A}_1 and thus $\lambda(\mathcal{A}_2 \mathcal{A}_1^{-1}) = 1$, and so, $\rho(\mathcal{A}_2 \mathcal{A}_1^{-1}) = 1$, since the matrix $\mathcal{A}_2 \mathcal{A}_1^{-1}$ is proportional to σ_η^2 . Hence, $\rho(\mathcal{A}) < 1$ if and only if $\rho(\mathcal{A}_2 \mathcal{A}_1^{-1}) < 1$. ■

Theorem 1: The following statements are equivalent.

- i) The floating-point state covariance represented by (21) is mean square stable.
- ii) There exists $\mathbf{P} > \mathbf{0}$ satisfying the LMI

$$\mathbf{A} \mathbf{P} \mathbf{A}^T - \mathbf{P} + \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \mathbf{P} \mathbf{A}^T \mathbf{E}_i < \mathbf{0}. \quad (27)$$

iii)

$$\sigma_\eta^2 < \left[\rho \left\{ \sum_{i=1}^n (\mathbf{E}_i \otimes \mathbf{E}_i) (\mathbf{A} \otimes \mathbf{A}) (\mathbf{I} - \mathbf{A} \otimes \mathbf{A})^{-1} \right\} \right]^{-1}. \quad (28)$$

Proof: It is shown [14] that i) is equivalent to ii). By Lemma 1, $\rho(\mathcal{A}) < 1$ is equivalent to $\rho(\mathcal{A}_2 \mathcal{A}_1^{-1}) < 1$, from which iii) can be obtained, and hence, i) is equivalent to iii). ■

Together with (5), the upper bound for final roundoff error variance given by (28) can be used to determine the minimum number of final mantissa bits that is required for stable realization of filters which have poles very close to the unit circle.

From (22) and if $\rho(\mathcal{A}_2 \mathcal{A}_1^{-1}) < 1$

$$\bar{\mathbf{x}} = \mathcal{A}_1^{-1} \sum_{k=0}^{\infty} (\mathcal{A}_2 \mathcal{A}_1^{-1})^k \mathbf{q}. \quad (29)$$

Multiplying on both sides by $\mathcal{A}_1 = \mathbf{I} - (\mathbf{A} \otimes \mathbf{A})$ yields

$$\mathcal{A}_1 \bar{\mathbf{x}} = \sum_{k=0}^{\infty} (\mathcal{A}_2 \mathcal{A}_1^{-1})^k \mathbf{q}. \quad (30)$$

Therefore

$$\bar{\mathbf{x}} = (\mathbf{A} \otimes \mathbf{A}) \bar{\mathbf{x}} + \mathbf{q} + \sum_{k=1}^{\infty} \mathbf{r}_k \quad (31)$$

where

$$\mathbf{r}_k \triangleq \sigma_\eta^{2k} \left[\sum_{i=1}^n \sum_{j=1}^{\infty} (\mathbf{E}_i \mathbf{A}^j) \otimes (\mathbf{E}_i \mathbf{A}^j) \right]^k \mathbf{q}. \quad (32)$$

Then, unvectorizing $\bar{\mathbf{x}}$ yields

$$\begin{aligned} \bar{\mathbf{X}} &= \mathbf{A} \bar{\mathbf{X}} \mathbf{A}^T + \mathbf{Q} + \sum_{k=1}^{\infty} \mathbf{R}_k \\ &= \sum_{i=0}^{\infty} \mathbf{A}^i \left\{ \mathbf{Q} + \sum_{k=1}^{\infty} \mathbf{R}_k \right\} (\mathbf{A}^T)^i \\ &= \sigma_u^2 \mathbf{K} + \sum_{i=0}^{\infty} \mathbf{A}^i \left(\sum_{k=1}^{\infty} \mathbf{R}_k \right) (\mathbf{A}^T)^i \end{aligned} \quad (33)$$

where $\mathbf{q} \triangleq \text{Vec}(\mathbf{Q})$, $\mathbf{r}_k \triangleq \text{Vec}(\mathbf{R}_k)$, and \mathbf{R}_k can be expressed recursively from (32).

$$\begin{aligned} \mathbf{r}_k &= \sigma_\eta^{2k} \left[\sum_{i=1}^n \sum_{j=1}^{\infty} (\mathbf{E}_i \mathbf{A}^j) \otimes (\mathbf{E}_i \mathbf{A}^j) \right]^k \mathbf{q} \\ &= \sigma_\eta^2 \left[\sum_{i=1}^n \sum_{j=1}^{\infty} (\mathbf{E}_i \mathbf{A}^j) \otimes (\mathbf{E}_i \mathbf{A}^j) \right] \text{Vec}(\mathbf{R}_{k-1}) \\ &= \sigma_\eta^2 \sum_{i=1}^n \sum_{j=1}^{\infty} \text{Vec} \left[\mathbf{E}_i \mathbf{A}^j \mathbf{R}_{k-1} (\mathbf{A}^T)^j \mathbf{E}_i \right]. \end{aligned} \quad (34)$$

Therefore

$$\mathbf{R}_k = \sigma_\eta^2 \sum_{i=1}^n \sum_{j=1}^{\infty} \mathbf{E}_i \mathbf{A}^j \mathbf{R}_{k-1} (\mathbf{A}^T)^j \mathbf{E}_i \quad (35)$$

with

$$\begin{aligned} \mathbf{R}_1 &= \sigma_u^2 \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \left\{ \sum_{j=1}^{\infty} \mathbf{A}^j \mathbf{B} \mathbf{B}^T (\mathbf{A}^T)^j \right\} \mathbf{E}_i \\ &= \sigma_u^2 \sigma_\eta^2 \text{diag}(\mathbf{A} \mathbf{K} \mathbf{A}^T) \end{aligned} \quad (36)$$

where $\text{diag}(\mathbf{A} \mathbf{K} \mathbf{A}^T)$ represents a diagonal matrix whose diagonal elements are the same as those of $\mathbf{A} \mathbf{K} \mathbf{A}^T$.

The steady-state covariance of the output error is given by

$$\Delta Y \triangleq \lim_{k \rightarrow \infty} \mathcal{E} \{ \Delta y^2(k) \} = \mathbf{C} \bar{\mathbf{E}} \mathbf{C}^T + \sigma_\eta^2 \mathbf{C} \bar{\mathbf{X}} \mathbf{C}^T + \sigma_u^2 \sigma_\eta^2 D^2 \quad (37)$$

where $\bar{\mathbf{E}}$ satisfies

$$\begin{aligned} \bar{\mathbf{E}} &\triangleq \lim_{k \rightarrow \infty} \bar{\mathbf{E}}_k \\ &= \mathbf{A} \bar{\mathbf{E}} \mathbf{A}^T + \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{A} \bar{\mathbf{X}} \mathbf{A}^T \mathbf{E}_i \\ &\quad + \sigma_u^2 \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{B} \mathbf{B}^T \mathbf{E}_i. \end{aligned} \quad (38)$$

In (33), $\bar{\mathbf{X}}$ consists of two terms and lets us see separately the effect of each on $\bar{\mathbf{E}}$ and ΔY . Let $\bar{\mathbf{X}}^{(1)}$, $\bar{\mathbf{E}}^{(1)}$, and $\Delta Y^{(1)}$ represent the effect of the first term and $\bar{\mathbf{X}}^{(2)}$, $\bar{\mathbf{E}}^{(2)}$, and $\Delta Y^{(2)}$ represent the effect of the second term.

First, when only the first term of $\bar{\mathbf{X}}$ in (33) is considered, that is $\bar{\mathbf{X}}^{(1)} = \sigma_u^2 \mathbf{K}$, we will have from (38)

$$\begin{aligned} \bar{\mathbf{E}}^{(1)} &= \mathbf{A} \bar{\mathbf{E}}^{(1)} \mathbf{A}^T + \sigma_u^2 \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i (\mathbf{A} \mathbf{K} \mathbf{A}^T + \mathbf{B} \mathbf{B}^T) \mathbf{E}_i \\ &= \mathbf{A} \bar{\mathbf{E}}^{(1)} \mathbf{A}^T + \sigma_u^2 \sigma_\eta^2 \sum_{i=1}^n \mathbf{E}_i \mathbf{K} \mathbf{E}_i \\ &= \sigma_u^2 \sigma_\eta^2 \sum_{k=0}^{\infty} \mathbf{A}^k \left[\sum_{i=1}^n \mathbf{E}_i \mathbf{K} \mathbf{E}_i \right] (\mathbf{A}^T)^k. \end{aligned} \quad (39)$$

Then, from (37)

$$\begin{aligned} \frac{\Delta Y^{(1)}}{\sigma_u^2 \sigma_\eta^2} &= \mathbf{C} \left\{ \sum_{k=0}^{\infty} \mathbf{A}^k \sum_{i=1}^n \mathbf{E}_i \mathbf{K} \mathbf{E}_i (\mathbf{A}^T)^k \right\} \mathbf{C}^T + \mathbf{C} \mathbf{K} \mathbf{C}^T + D^2 \\ &= \text{tr} \left\{ \sum_{k=0}^{\infty} (\mathbf{A}^T)^k \mathbf{C}^T \mathbf{C} \mathbf{A}^k \sum_{i=1}^n \mathbf{E}_i \mathbf{K} \mathbf{E}_i \right\} + \mathbf{C} \mathbf{K} \mathbf{C}^T + D^2 \\ &= \text{tr} \left\{ \mathbf{W} \sum_{i=1}^n \mathbf{E}_i \mathbf{K} \mathbf{E}_i \right\} + \mathbf{C} \mathbf{K} \mathbf{C}^T + D^2. \end{aligned} \quad (40)$$

Therefore

$$\Delta Y^{(1)} = \sigma_u^2 \sigma_\eta^2 \left\{ \mathbf{C} \mathbf{K} \mathbf{C}^T + D^2 + \sum_{i=1}^n \mathbf{K}_{ii} \mathbf{W}_{ii} \right\} \quad (41)$$

which is the same as the expression for output error covariance in [3] and [4]. Therefore, it can be said that if the second term of state covariance recursion in (21) is neglected, then the output error covariance will be the same as the previous studies.

Now, to see the effect of the second term of $\bar{\mathbf{X}}$, denote the effect of particular \mathbf{R}_k on $\Delta Y^{(2)}$, $\bar{\mathbf{X}}^{(2)}$ and $\bar{\mathbf{E}}^{(2)}$ by $\Delta Y_k^{(2)}$, $\bar{\mathbf{X}}(\mathbf{R}_k)^{(2)}$ and $\bar{\mathbf{E}}(\mathbf{R}_k)^{(2)}$, respectively.

Then, we have

$$\begin{aligned} \Delta Y^{(2)} &= \sum_{k=1}^{\infty} \Delta Y_k^{(2)} \\ &= \sum_{k=1}^{\infty} \mathbf{C} \bar{\mathbf{E}}(\mathbf{R}_k)^{(2)} \mathbf{C}^T + \sigma_\eta^2 \mathbf{C} \bar{\mathbf{X}}(\mathbf{R}_k)^{(2)} \mathbf{C}^T. \end{aligned} \quad (42)$$

From (33), (35), and (38), it can be shown that

$$\begin{aligned} &\sigma_\eta^2 \mathbf{C} \bar{\mathbf{X}}(\mathbf{R}_k)^{(2)} \mathbf{C}^T \\ &= \sigma_\eta^4 \text{tr} \left\{ \sum_{l=1}^{\infty} \mathbf{C} \mathbf{A}^l \sum_{s=1}^n \sum_{t=1}^{\infty} \mathbf{E}_s \mathbf{A}^t \mathbf{R}_{k-1} (\mathbf{A}^T)^t \mathbf{E}_s (\mathbf{A}^T)^l \mathbf{C}^T \right\} \\ &= \sigma_\eta^4 \text{tr} \left\{ \sum_{t=1}^{\infty} (\mathbf{A}^T)^t \text{diag}(\mathbf{W}) \mathbf{A}^t \mathbf{R}_{k-1} \right\} \\ &= \sigma_\eta^4 \text{tr} \{ \mathbf{W}_1 \mathbf{R}_{k-1} \} \end{aligned} \quad (43)$$

and

$$\begin{aligned} &\mathbf{C} \bar{\mathbf{E}}(\mathbf{R}_k)^{(2)} \mathbf{C}^T \\ &= \sigma_\eta^2 \text{tr} \left\{ \mathbf{C} \sum_{l=0}^{\infty} \mathbf{A}^l \sum_{s=1}^n \sum_{t=1}^{\infty} \mathbf{E}_s \mathbf{A}^t \mathbf{R}_k (\mathbf{A}^T)^t \mathbf{E}_s (\mathbf{A}^T)^l \mathbf{C}^T \right\} \\ &= \sigma_\eta^2 \text{tr} \left\{ \text{diag}(\mathbf{W}) \sum_{t=1}^{\infty} \mathbf{A}^t \mathbf{R}_k (\mathbf{A}^T)^t \right\} \\ &= \sigma_\eta^4 \text{tr} \left\{ \sum_{t=1}^{\infty} (\mathbf{A}^T)^t \text{diag}(\mathbf{W}) \mathbf{A}^t \right. \\ &\quad \left. \times \sum_{s=1}^n \sum_{l=1}^{\infty} \mathbf{E}_s \mathbf{A}^l \mathbf{R}_{k-1} (\mathbf{A}^T)^l \mathbf{E}_s \right\} \\ &= \sigma_\eta^4 \text{tr} \{ \mathbf{W}_2 \mathbf{R}_{k-1} \} \end{aligned} \quad (44)$$

where

$$\mathbf{W}_k \triangleq \sum_{l=1}^{\infty} (\mathbf{A}^T)^l \text{diag}(\mathbf{W}_{k-1}) (\mathbf{A})^l \text{ with } \mathbf{W}_0 = \mathbf{W}. \quad (45)$$

Hence

$$\begin{aligned} \Delta Y_k^{(2)} &= \sigma_\eta^4 \text{tr} \{ (\mathbf{W}_1 + \mathbf{W}_2) \mathbf{R}_{k-1} \} \\ &= \sigma_\eta^6 \text{tr} \left\{ \sum_{t=1}^{\infty} (\mathbf{A}^T)^t \text{diag}(\mathbf{W}_1 + \mathbf{W}_2) \mathbf{A}^t \mathbf{R}_{k-2} \right\} \\ &= \sigma_\eta^6 \text{tr} \{ (\mathbf{W}_2 + \mathbf{W}_3) \mathbf{R}_{k-2} \} \\ &\quad \vdots \\ &= \sigma_\eta^2 \text{tr} \{ (\mathbf{W}_{k-1} + \mathbf{W}_k) \mathbf{R}_1 \} \\ &= \sigma_u^2 \sigma_\eta^{2(k+1)} \\ &\quad \times \text{tr} \left\{ (\mathbf{W}_{k-1} + \mathbf{W}_k) \text{diag}(\mathbf{A} \mathbf{K} \mathbf{A}^T) \right\}. \end{aligned} \quad (46)$$

TABLE I
 COMPARISON OF OUTPUT ERROR COVARIANCE

Realization	L=10			L=42		
	$\Delta Y^{(1)}$	$\Delta Y^{(2)}$	Bits	$\Delta Y^{(1)}$	$\Delta Y^{(2)}$	Bits
Phase-variable	1.91×10^8	2.12×10^{-2}	7.46	3.76×10^{27}	3.65×10^{27}	23.6
Fixed point optimal	5.25×10^5	1.59×10^{-7}	3.21	1.03×10^{25}	1.39×10^{22}	19.2

Therefore, we have

$$\begin{aligned}
 \frac{\Delta Y}{\sigma_u^2 \sigma_\eta^2} &= \frac{\Delta Y^{(1)} + \sum_{k=1}^{\infty} \Delta Y_k^{(2)}}{\sigma_u^2 \sigma_\eta^2} \\
 &= \mathbf{CKC}^T + D^2 + \mathbf{tr}\{\mathbf{W}\mathbf{diag}(\mathbf{K})\} \\
 &\quad + \sum_{k=1}^{\infty} \sigma_\eta^{2k} \mathbf{tr}\left\{(\mathbf{W}_{k-1} + \mathbf{W}_k)\mathbf{diag}(\mathbf{AKA}^T)\right\} \\
 &= \mathbf{CKC}^T + D^2 + \mathbf{tr}\{\mathbf{W}\mathbf{diag}(\mathbf{K})\} \\
 &\quad + \sigma_\eta^2 \mathbf{tr}\left\{\mathbf{W}\mathbf{diag}(\mathbf{AKA}^T)\right\} \\
 &\quad + (1 + \sigma_\eta^2) \sum_{k=1}^{\infty} \sigma_\eta^{2k} \mathbf{tr}\left\{\mathbf{W}_k \mathbf{diag}(\mathbf{AKA}^T)\right\} \\
 &\simeq \mathbf{CKC}^T + D^2 + \mathbf{tr}\{\mathbf{W}\mathbf{diag}(\mathbf{K})\} \\
 &\quad + \sum_{k=1}^{\infty} \sigma_\eta^{2k} \mathbf{tr}\left\{\mathbf{W}_k \mathbf{diag}(\mathbf{AKA}^T)\right\} \quad (47)
 \end{aligned}$$

where $\sigma_\eta^2 \ll 1$ and $\mathbf{K} \geq \mathbf{AKA}^T$ are used.

Lemma 2:

$$\sum_{k=1}^{\infty} \sigma_\eta^{2k} \mathbf{tr}\left\{\mathbf{W}_k \mathbf{diag}(\mathbf{AKA}^T)\right\} \quad (48)$$

is a geometric series, and the ratio of each two consecutive terms r is given by

$$r = \frac{\mathbf{tr}\left\{\mathbf{diag}\left(\sum_{l=1}^{\infty} \mathbf{A}^l \mathbf{diag}(\mathbf{AKA}^T)(\mathbf{A}^T)^l\right) \mathbf{W}_1\right\}}{\mathbf{tr}\left\{\mathbf{diag}(\mathbf{AKA}^T) \mathbf{W}_1\right\}} \quad (49)$$

where

$$\mathbf{W}_1 = \sum_{l=1}^{\infty} (\mathbf{A}^T)^l \mathbf{diag}(\mathbf{W}) \mathbf{A}^l. \quad (50)$$

Proof: See the Appendix. \blacksquare

Direct application of Lemma 2 yields Theorem 2.

Theorem 2: For a given infinite precision digital filter represented by (6), the steady-state covariance of output error ΔY is given by

$$\Delta Y = \Delta Y^{(1)} + \Delta Y^{(2)} \quad (51)$$

$$\Delta Y^{(1)} = \sigma_u^2 \sigma_\eta^2 \left\{ \mathbf{CKC}^T + D^2 + \mathbf{tr}\left(\mathbf{W}\mathbf{diag}(\mathbf{K})\right) \right\} \quad (52)$$

$$\Delta Y^{(2)} = \sigma_u^2 \sigma_\eta^2 \frac{\sigma_\eta^2}{1-r} \mathbf{tr}\left\{\mathbf{W}_1 \mathbf{diag}(\mathbf{AKA}^T)\right\} \quad (53)$$

for a zero mean white noise input signal of variance σ_u^2 when the filter (6) is implemented with a digital signal processor using accumulation for internal register-to-register arithmetic. Here, W_{ii} is the i - i th element of the observability gramian \mathbf{W} of matrix pair (\mathbf{A}, \mathbf{C}) , \mathbf{K} is the controllability gramian of matrix pair (\mathbf{A}, \mathbf{B}) , σ_η^2 is the variance of the final quantization error given by (5), and r is given by (49).

According to the two previous studies in [3] and [4], the output error covariance is given by

$$\Delta Y_{\text{previous}} = \sigma_u^2 \sigma_\eta^2 \left\{ \mathbf{CKC}^T + D^2 + \sum_{i=1}^n K_{ii} W_{ii} \right\} \quad (54)$$

which is the same as $\Delta Y^{(1)}$ in Theorem 2. For very stable \mathbf{A} , $\Delta Y^{(1)}$, that is $\Delta Y_{\text{previous}}$ well approximates to ΔY , since for $\sigma_\eta^2 \ll 1$, usually

$$\frac{\sigma_\eta^2}{1-r\sigma_\eta^2} \mathbf{tr}\left\{\mathbf{W}_1 \mathbf{diag}(\mathbf{AKA}^T)\right\} \ll \mathbf{tr}\left(\mathbf{W}\mathbf{diag}(\mathbf{K})\right)$$

and therefore, $\Delta Y^{(2)}$ can be neglected in ΔY expression (51). However, for matrices \mathbf{A} that have poles very near to the unit circle, it cannot be neglected. Since the term $\Delta Y^{(2)}$ comes from the second term in the right-hand side of the state covariance equation given by (21), if that term is neglected, then it yields the same expression for output covariance as the one given by the two previous studies, which is illustrated in the following example of a frequency sampling filter that has poles on a circle with radius slightly less than one.

Example 1: Consider the first resonator of linear phase type 2 frequency sampling filter ([15], [16]) for FIR filters, with real-valued impulse response coefficients, and with length $N = 60$. The transfer function is given by

$$H(z) = \frac{2|H_0| \sin\left(\frac{\pi}{N}\right)}{1 - 2az^{-1} \cos\left(\frac{\pi}{N}\right) + a^2 z^{-2}} \quad (55)$$

with $|H_0| = 1$ and $a = 1 - 2^{-L}$. This filter has two zeros at the origin and two poles near to the unit circle. The bigger L , the closer to the unit circle. To see the effect of the final quantization, consider two cases: one for $L = 10$ and the other $L = 42$, which is an extreme case. Table I compares the output error covariance of phase-variable form [17] state-space realization and fixed point optimal realization for each L . In Table I, ‘‘Bits’’-column represents the minimal required mantissa bits

for stable realization calculated from (5) and (28). As shown, when $L = 10$, $\Delta Y^{(2)}$ can be neglected compared with $\Delta Y^{(1)}$, but in case of the extreme value $L = 42$, $\Delta Y^{(2)}$ is as big as $\Delta Y^{(1)}$. Table I shows also that the fixed point optimal realization still results in much smaller output error covariance than the phase-variable form, even for the case of $L = 42$.

As shown in Example 1, when the poles are not extremely close to the unit circle, the output covariance due to the second term of state covariance (21) can be neglected and in this case the optimal realization is the same as the fixed point case, but if a filter has poles extremely close to the unit circle, the additional term $\Delta Y^{(2)}$ can be as big as $\Delta Y^{(1)}$ and cannot be neglected. In addition to this case, the additional term $\Delta Y^{(2)}$ can be also big when σ_η^2 is big, that is, when the final quantization to precision shorter than single precision (e.g., final quantization to short precision) is employed. So, in these cases, the optimal realization structure will be different from that of fixed point case, since the optimal realization structure should satisfy simultaneously the following conditions.

- Minimize $\text{tr}(\mathbf{W} \text{diag}(\mathbf{K}))$.
- Minimize $\text{tr}\{\mathbf{W}_1 \text{diag}(\mathbf{AKA}^T)\}$.
- Minimize r , which is given by (49).

Solving the above minimization problem to find the optimal structure will be very complex, but we can observe that the fixed point optimal realization which minimizes $\text{tr}(\mathbf{W})$ with the constraint $\mathbf{K}_{ii} = 1$ for all $i = 1, \dots, n$ also reduces the output covariance, although it does not minimize it. This is confirmed also in Example 1.

III. CONCLUDING REMARKS

The floating-point roundoff errors coming from implementing digital filters 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 effects on digital filters were 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 accumulation, and it was reconfirmed that the optimal state space digital filter realization of floating-point arithmetic will be the same as that of fixed-point arithmetic.

APPENDIX PROOF OF LEMMA 2

Consider the ratio of the $(k+1)$ th term to k th term of the series (48), which is denoted by r_k .

$$r_k = \sigma_\eta^2 \frac{\text{tr}\{\text{diag}(\mathbf{AKA}^T)\mathbf{W}_{k+1}\}}{\text{tr}\{\text{diag}(\mathbf{AKA}^T)\mathbf{W}_k\}} = \sigma_\eta^2 \frac{\text{tr}\{\text{diag}[\sum_{l=1}^{\infty} \mathbf{A}^l \text{diag}[\mathbf{AKA}^T]_{ii} (\mathbf{A}^T)^l] \mathbf{W}_k\}}{\text{tr}\{\text{diag}(\mathbf{AKA}^T)\mathbf{W}_k\}}. \quad (56)$$

The numerator and the denominator of r_k can be expressed by the Vec-operator as follows:

$$\begin{aligned} & \text{Numerator of } r_k \\ &= \sigma_\eta^2 \left\{ \text{Vec} \left(\sum_{s=1}^n \mathbf{E}_s \sum_{l=1}^{\infty} \mathbf{A}^l \text{diag}(\mathbf{AKA}^T) (\mathbf{A}^T)^l \mathbf{E}_s \right) \right\}^T \\ & \quad \times \text{Vec}(\mathbf{W}_k) \\ &= \sigma_\eta^2 \left\{ \sum_{s=1}^n \sum_{l=1}^{\infty} (\mathbf{E}_s \mathbf{A}^l \otimes \mathbf{E}_s \mathbf{A}^l) \text{Vec} \left(\text{diag}(\mathbf{AKA}^T) \right) \right\}^T \\ & \quad \times \text{Vec}(\mathbf{W}_k) \\ &= \sigma_\eta^2 \text{Vec}(\mathbf{K})^T (\mathbf{A} \otimes \mathbf{A})^T \sum_{s=1}^n (\mathbf{E}_s \otimes \mathbf{E}_s) \sum_{l=1}^{\infty} (\mathbf{A}^T \otimes \mathbf{A}^T)^l \\ & \quad \times \sum_{s=1}^n (\mathbf{E}_s \otimes \mathbf{E}_s) \text{Vec}(\mathbf{W}_k). \quad (57) \end{aligned}$$

Similarly

$$\begin{aligned} \text{denominator of } r_k &= \{\text{Vec}(\mathbf{K})\}^T (\mathbf{A} \otimes \mathbf{A})^T \\ & \quad \times \sum_{s=1}^n (\mathbf{E}_s \otimes \mathbf{E}_s) \text{Vec}(\mathbf{W}_k). \quad (58) \end{aligned}$$

Let $\mathbf{M}_e \triangleq \sum_{i=1}^n (\mathbf{E}_i \otimes \mathbf{E}_i)$, $\mathbf{M}_a \triangleq \sum_{l=1}^{\infty} (\mathbf{A} \otimes \mathbf{A})^l$, $\mathbf{A}^{[2]} \triangleq (\mathbf{A} \otimes \mathbf{A})$, $\mathbf{v}_k \triangleq \text{Vec}(\mathbf{W}_k)$, and $\mathbf{k} \triangleq \text{Vec}(\mathbf{K})$. From (45),

$$\mathbf{v}_k = \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1}. \quad (59)$$

Then

$$\begin{aligned} \frac{r_k}{r_{k-1}} &= \frac{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_k}{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{v}_k} \cdot \frac{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{v}_{k-1}}{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1}} \\ &= \frac{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1}}{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1}} \\ & \quad \cdot \frac{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{v}_{k-1}}{\mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1}} \quad (60) \end{aligned}$$

and

$$\begin{aligned} & \text{Numerator of } \frac{r_k}{r_{k-1}} \\ &= \mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \\ &= \text{tr} \left(\mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \mathbf{k}^T \mathbf{A}^{[2]T} \right. \\ & \quad \left. \times \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \right). \quad (61) \end{aligned}$$

Since $\mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e = \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e$

$$\begin{aligned} & \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \\ &= \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \\ &= \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \\ &= \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e. \quad (62) \end{aligned}$$

Therefore

$$\text{Numerator of } \frac{r_k}{r_{k-1}} = \text{tr} \left(\mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \mathbf{k}^T \right. \\ \left. \times \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \right). \quad (63)$$

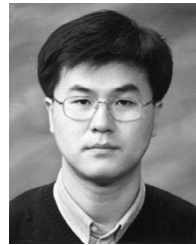
Similarly

$$\text{denominator of } \frac{r_k}{r_{k-1}} \\ = \mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \\ = \text{tr} \left(\mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \right) \\ = \text{tr} \left(\mathbf{M}_a \mathbf{M}_e \mathbf{A}^{[2]} \mathbf{k} \mathbf{k}^T \mathbf{A}^{[2]T} \mathbf{M}_e \mathbf{M}_a^T \right. \\ \left. \times \mathbf{M}_e \mathbf{M}_e \mathbf{v}_{k-1} \mathbf{v}_{k-1}^T \mathbf{M}_e \right). \quad (64)$$

From (63) and (64), $r_k/r_{k-1} = 1$. Hence, r_k is constant for all k . Taking $k = 1$ yields (49). ■

REFERENCES

- [1] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551–562, Sept. 1976.
- [2] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.
- [3] B. D. Rao, "Floating point arithmetic and digital filters," *IEEE Trans. Signal Processing*, vol. 40, pp. 85–95, Jan. 1992.
- [4] B. W. Bomar, L. M. Smith, and R. D. Joseph, "Roundoff noise analysis of state-space digital filters implemented on floating-point digital signal processors," *IEEE Trans. Circuits Syst. II*, vol. 44, pp. 952–955, Nov. 1997.
- [5] G. Shi and R. E. Skelton, "State feedback covariance control for linear finite signal-to-noise ratio models," in *Proc. CDC*, New Orleans, LA, 1995, pp. 3423–3428.
- [6] M. C. de Oliveira and R. E. Skelton, "State feedback control of linear systems in the presence of devices with finite signal-to-noise ratio," *Int. J. Contr.*, vol. 74, pp. 1501–1509, Oct. 2001.
- [7] J. L. Willems and J. C. Willems, "Feedback stabilizability for stochastic systems with state and control dependent noise," *Automatica*, vol. 12, pp. 277–283, May 1976.
- [8] T. Sasagawa and J. L. Willems, "Parametrization method for calculating exact stability bounds of stochastic linear systems with multiplicative noise," *Automatica*, vol. 32, pp. 1741–1747, Dec. 1996.
- [9] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems*. London, U.K.: Springer-Verlag, 1993.
- [10] T. Kaneko and B. Liu, "On local roundoff errors in floating-point arithmetic," *J. Assoc. Comput. Mach.*, vol. 20, pp. 391–398, July 1973.
- [11] L. M. Smith, B. W. Bomar, R. D. Joseph, and G. C. Yang, "Floating-point roundoff noise analysis of second-order state-space digital filter structures," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 90–98, Feb. 1992.
- [12] J. W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 772–781, Sept. 1978.
- [13] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*. London, U.K.: Taylor and Francis, 1998.
- [14] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: SIAM, 1994.
- [15] L. R. Rabiner and R. W. Schafer, "Recursive and nonrecursive realizations of digital filters designed by frequency sampling techniques," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 200–207, Sept. 1971.
- [16] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
- [17] R. E. Skelton, *Dynamic Systems Control*. New York: Wiley, 1988.
- [18] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1985.



Sangho Ko was born in Seoul, Korea, in 1967. He received the B.Sc. degree in aerospace and mechanical engineering from Hankuk Aviation University, Goyang, Korea, and the M.Sc. degree in aerospace engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1989 and 1992, respectively. He has been pursuing the Ph.D. degree with the Department of Mechanical and Aerospace Engineering, the University of California at San Diego, La Jolla, since 2000.

He was with Samsung Aerospace Industries, Ltd., Kyungnam, Korea, from 1992 to 1999, where he was involved in designing digital flight control system of the advanced jet trainer T-50 for the Republic of Korea Air Force. His current research focuses on state estimation and control problems with state constraints.



Robert R. Bitmead (F'91) was born in Sydney, Australia, in 1954. He received the B.Sc. degree in applied mathematics from the University of Sydney and the M.E. and Ph.D. degrees from the University of Newcastle, Callaghan, Australia, in 1976, 1977, and 1979, respectively.

He has held the Cymer Endowed Chair in the Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, since 1999 and has held faculty positions at the Australian National University, Canberra, from 1982 to 1999, and James Cook University of North Queensland, Townsville, Australia, from 1980 to 1982. He has held visiting faculty positions at Cornell University, Ithaca, NY; University of Louvain, Louvain-la-Neuve, Belgium; INRIA, Paris, France, and Kyoto University, Kyoto, Japan. He is the Editor of Adaptive and Intelligent Control for *Automatica*. His research interests are in the areas of adaptive systems, estimation, control design, modeling, and telecommunications.

Dr. Bitmead is a Fellow of the Australian Academy of Technological Sciences and Engineering.