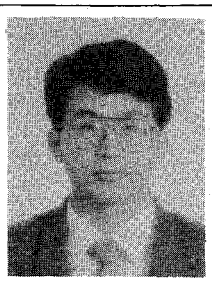
Stability Analysis of Digital Kalman Filters with Floating-Point Computation

Chih-Tsung Kuo
Tatung Institute of Technology, Taipei, Taiwan, Republic of China
Bor-Sen Chen
National Tsing Hua University, Hsinchu, Taiwan, Republic of China
and
Zeal-Sain Kuo
Chung Cheng Institute of Technology, Tahsi, Taoyuan, Taiwan, Republic of China

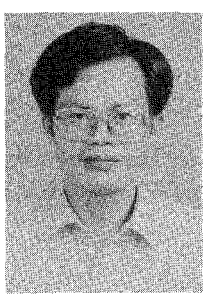
This paper is concerned with the stability analysis of an estimator with a Kalman filter having both an analog-to-digital converter (ADC) with finite wordlength and floating-point hardware with finite mantissa length. A new stability condition is introduced for the Kalman filter with roundoff noise and scaling in digital implementation. It is found that the effects of both ADC scaling and finite floating-point computation are dependent on the ADC and mantissa lengths. The derivation of the upper bound on the actual estimation error caused by roundoff noise and ADC scaling is based mainly on the Bellman-Gronwall lemma in discrete form. An example of state estimation in an inertial navigation system demonstrates the stability criterion to ensure the stability of the state estimator.

I. Introduction

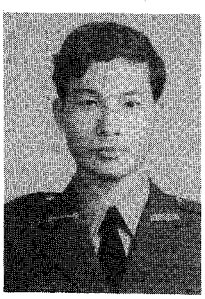
DIGITAL computers are being used increasingly to implement flight control, navigation control, or engine control in aircraft. Filtering algorithms such as Kalman filters in these systems are realized either with special-purpose digital hardware or in programs for a general-purpose digital computer. In these situations, state values and coefficients are stored in registers with a finite number of bits. The size of the available register for each value is finite, and, therefore, the values must be quantized. After every addition or multiplication (floating- or fixed-point arithmetic), the results must also be quantized.



Chih-Tsung Kuo was born in Hsin Chu, Taiwan, Republic of China, on February 28, 1962. He received the B.S. and M.S. degrees in electrical engineering from Tatung Institute of Technology, Taiwan, Republic of China, in 1984 and 1986, respectively. He is now a lecturer of electrical engineering at Tatung Institute of Technology. His current research interests are in the areas of optimal control, digital signal processing, and local area network. His mailing address is at the Department of Electrical Engineering, No. 40 Chungshan North Road, 3rd Section.



Bor-Sen Chen received the B.S. degree from Tatung Institute of Technology, the M.S. degree from National Central University, Taiwan, Republic of China, and the Ph.D. degree from the University of Southern California. He is now a professor of electrical engineering at National Tsing Hua University, Hsinchu, Taiwan, Republic of China. He was a lecturer, associate professor, and professor at Tatung Institute of Technology from 1973 to 1987. His current research interests are in robust control, adaptive control, and signal processing. His mailing address is at the Department of Electrical Engineering, Kuang Fu Road.



Zeal-Sain Kuo was born in Hsin Chu, Taiwan, Republic of China, on July 14, 1960. He received the B.S. degree in aeronautical engineering from Chung Cheng Institute of Technology in 1982 and the M.S. degree in mechanical engineering from National Taiwan University in 1986. He is now a lecturer of mechanical engineering at Chung Cheng Institute of Technology and assistant researcher in Chung Shan Institute of Science and Technology. His current research interests are in the areas of optimal control, aerodynamics, and navigation. His mailing address is at the Department of Aeronautical Engineering.

Received Sept. 6, 1989; revision received and accepted for publication July 6, 1990.

These quantizations cause roundoff errors. Hence, when implementing a filtering algorithm on a floating-point computer, one should consider the problems that arise in dealing with floating-point computation and finite wordlength.

In digital signal processing, a great deal of work has been devoted to the analysis of roundoff error.[1-6] These digital signal processing ideas serve as the basis of techniques dealing with roundoff noise for optimal state estimation in this paper. Some research has also been done on finite-wordlength analysis of digital controllers using fixed- and floating-point arithmetic.[1,8,9,11,12] For example, roundoff noise and scaling in the digital implementation of linear quadratic Gaussian compensators are considered by Moroney et al.[9] Rink and Chong[11] have determined an upper bound on the mean-square error in state regulator systems. The influence of linear optimal discrete-time systems due to finite wordlength is introduced by Van Wingerden and De Koning.[12] In Ref. 1, a new algorithm for finding the digital realization of discrete-time transfer functions by fixed-point arithmetic has been introduced. The finite-wordlength effects on the design of Kalman filters has already been given in Refs. 7 and 18. Stripad[7] analyzed expected degradation in the performance of a fixed-point implementation of a digital Kalman filter with precomputed gains. Williamson[18] has presented a more specific formulation of a similar problem in Ref. 7. In particular, he has taken the state roundoff quantization into account and applied the error spectrum shaping technique[10] in the digital Kalman filter design. However, less research has been published on the subject of digital Kalman filters employing floating-point arithmetic.

Fixed-point arithmetic can be implemented quickly and economically in filtering algorithms; however, to avoid arithmetic overflow, the arithmetic must be handled by scaling inputs, outputs, coefficients, and states. In considering floating-point arithmetic, such dynamic range considerations generally can be neglected because of the large range of representable numbers, but quantization is introduced for both multiplication and addition. These quantizations lead to a deterioration in the ideal (i.e., infinite wordlength) performances of filtering algorithms. In extreme cases, these quantizations can even lead to an ideally stable filtering algorithm becoming unstable. This instability issue deeply influences the design of Kalman filters requiring the realization of a cluster of poles near the unit circle for state estimation.[18] This paper is concerned with an estimator system influenced by finite wordlength, where floating-point arithmetic is used in the optimal estimation computation.

In this paper, the effect of computation roundoff error, which is considered as perturbations of the Kalman filter's system due to multiplication and addition, is discussed. Based on the Bellman-Gronwall lemma, a robust stability criterion is derived to tolerate the perturbation due to the finite-wordlength effect, without leading to instability. According to this robust stability criterion, an adequate finite wordlength of the estimator using the Kalman filter under the consideration of roundoff error is proposed. At the same time, the actual estimation error bound and the state signal bound in this system are also discussed.

Assume that the continuous-time system is given in the following state-space form:

$$\frac{dx(t)}{dt} = \phi x(t) + \bar{v}(t); \qquad x(t) \in R^N \tag{1a}$$

$$y(t) = Cx(t) + \bar{e}(t); \qquad y(t) \in R^M \tag{1b}$$

where $\mathrm{col}[\bar{v}(t)\ \bar{e}(t)]$ is a zero-mean white noise process with intensity

$$\begin{bmatrix} V_1 & 0 \\ 0 & S_2 \end{bmatrix}; \qquad V_1 \geq 0,\ S_2 \geq 0 \tag{2}$$

When the state at the sampling time t_k is given, and the sampling period $h = t_{k+1} - t_k$,

$$t_k = k \cdot h$$

the discrete model of Eqs. (1) can be expressed as

$$x(kh + h) = Ax(kh) + v(kh); \qquad x(kh) \in R^N \tag{3a}$$

$$y(kh) = Cx(kh) + \bar{e}(kh); \qquad y(kh) \in R^M \tag{3b}$$

where

$$A = e^{\phi h}$$

$$v(kh) = \int_0^h e^{\phi(h-s')} \bar{v}(s' + kh)\, ds'$$

$v(kh)$ and $\bar{e}(kh)$ form sequences of white noise processes with zero-mean values and the covariances

$$E[v(kh)v^T(kh)] = S_1 \tag{4a}$$

$$E[v(kh)\bar{e}^T(kh)] = 0 \tag{4b}$$

$$E[\bar{e}(kh)\bar{e}^T(kh)] = S_2 \tag{4c}$$

Note that, in the case of a high sampling rate, the sampling period h will approach zero such that the eigenvalues of A near the unit circle and the covariance S_1 become small. We shall see that, under these situations, the effects of roundoff errors become more important, particularly for the cases of large S_2.[15, 18]

II. Problem Formulation

We consider the optimal state-estimation problem for a discrete-time system that is described by the state equation[16]

$$x(k + 1) = Ax(k) + v(k); \qquad x(k) \in R^N \tag{5a}$$

$$y(k) = Cx(k) + \bar{e}(k); \qquad y(k) \in R^M \tag{5b}$$

where $v(k)$ and $\bar{e}(k)$ are sequences of white noise processes with zero-mean values and the covariances

$$E[v(k)v^T(k)] = S_1 \tag{6a}$$

$$E[v(k)\bar{e}^T(k)] = 0 \tag{6b}$$

$$E[\bar{e}(k)\bar{e}^T(k)] = S_2 \tag{6c}$$

Here $E[\cdot]$ denotes the expectation of $[\cdot]$. For simplicity, the sampling time is used as the time unit, $h = 1$.

Suppose $y^*(k)$ is the output measurement of an L-bit analog-to-digital converter (ADC) whose input is the sampled output $y(k)$, and the ADC is using a symmetrical reference system. That is,

$$y^*(k) = y(k) + d(k) \tag{7}$$

where $d(k)$ is the ADC quantization error occurring at the kth sample time and satisfying $|d(k)| \leq 2^{-L}$ of full scale (FS). The probability distribution of $d(k)$ is uniform over the quantization error and also a white noise process.[3] Sometimes, the consideration of ADC problems is very troublesome, since an ADC may reach saturation and depend highly on the hardware specifications. However, it is necessary to convert analog signals to digital form using an ADC in most digital systems

containing analog as well as digital components. Then the discrete state equations [Eqs. (5)] can then be rewritten as

$$x(k+1)=Ax(k)+v(k); \qquad x(k)\in R^N \tag{8a}$$

$$y^*(k)=y(k)+e(k); \qquad y^*(k)\in R^M \tag{8b}$$

where

$$e(k)=\bar{e}(k)+d(k)$$

is a white noise process with mean zero and covariance $[S_2+(2^{-L}\cdot \mathrm{FS}\cdot 2)^2\cdot I_M/12]$, where I_M is an identity matrix with appropriate dimensions.

In the one-step-ahead prediction problem, let the process be described by Eqs. (5). The state estimator has the form

$$\hat{x}(k+1)=[A-K(k)C]\hat{x}(k)+K(k)y^*(k) \tag{9}$$

and the performance is defined as

$$J=E[\epsilon^T(k)\epsilon(k)] \tag{10}$$

where $\epsilon(k)=\hat{x}(k)-x(k)$, called estimation error. Then the Kalman filter is to choose $K(k)$ in Eq. (9) to minimize the performance J, and in an infinite-precision case, the solution is given as

$$K(k)=AP(k)C^T[S_2+CP(k)C^T]^{-1} \tag{11}$$

$$P(k+1)=AP(k)A^T+S_1-AP(k)C^T[S_2+CP(k)C^T]^{-1} \times CP^T(k)A^T \tag{12}$$

To implement a digital Kalman filter for state estimation, where we consider the roundoff errors due to the finite-word-length effects, let $\mathrm{fl}_i\{a+b\}$ denote the result of floating-point addition and $\mathrm{FL}_i[ab]$ denote the result of floating-point multiplication. A state estimator with a digital Kalman filter with floating-point computation is given by

$$\hat{x}^*(k+1)=\mathrm{fl}_2\Big[\mathrm{FL}_1[A\hat{x}^*(k)]+\mathrm{FL}_3\Big(K\cdot \mathrm{fl}_1\{y^*(k)-\mathrm{FL}_2[C\hat{x}^*(k)]\}\Big)\Big] \tag{13}$$

where $\hat{x}^*(k+1)$ is assumed to be a (signed) floating-point representation with a W-bit mantissa part.

It is assumed that the values of A and C here are retrieved from memory when they are needed, and the Kalman gain K is precomputed by solving the algebraic Riccati equation (12) in forward time and is stored in the computer. Thus these values have already been rounded to exact values with finite wordlengths. Now we can precisely state the main problem considered in this paper.

Problem. Consider the state estimator [Eq. (13)] in which the coefficient matrices A, C, and K exist as finite words. Under what conditions will the state estimator that has been deteriorated by the effects of the finite wordlengths still be stable?

III. Stability Analysis of Kalman Filters

The objective of this section will be to obtain a sufficient condition of stability for the state estimator [Eq. (13)] in the nonideal situation, where roundoff errors may occur with every elementary floating-point addition and multiplication. An upper bound on the estimation error degraded for finite-wordlength effects is derived. Before further analysis, some mathematic tools and definitions needed for solving our problem are introduced. Let the norm of real stochastic vector $x\in R^N$, denoted by $\|x\|$, be defined by[14]

$$\|x\|=\sqrt{E[x^Tx]} \tag{14}$$

Then

$$\|Ax\|^2=E[x^TA^TAx]=\mathrm{tr}(E[x^TA^TAx])=\mathrm{tr}(E[A^TAxx^T]) \tag{15}$$

where tr denotes the trace operator, and

$$\|Ax\|^2\le \|A\|^2\|x\|^2$$

where $\|A\|$ denotes the induced norm defined as follows:

$$\|A\|\triangleq \begin{cases}\sqrt{\lambda_{\max}(A^TA)} \text{ for } A & \text{is deterministic}\\ \sqrt{\lambda_{\max}(E[A^TA])} \text{ for } A & \text{is stochastic}\end{cases} \tag{16}$$

It is shown in Appendix A that the rounded floating-point sum of two M-vectors can be expressed by

$$\mathrm{fl}\{a+b\}=(I+\Delta R)(a+b) \tag{17}$$

where $\Delta=2^{-W}$, W being the mantissa length, and

$$R=\mathrm{diag}[r_1\ r_2\ \dots\ r_M]$$

where each r_i is distributed uniformly between -1 and 1, so that

$$E[r_i]=0, \qquad i=1,2,\dots,M$$

$$E[r_i r_j]=\tfrac{1}{3}\delta_{ij} \tag{18}$$

The rounded floating-point product of an $M\times M$ matrix A and an M-vector x, also shown in Appendix A, is given as follows:

$$\mathrm{FL}[Ax]=A(I+\Delta H)x \tag{19}$$

with

$$H=\mathrm{diag}[h_1\ h_2\ \dots\ h_M]$$

where h_i has zero mean and the variances are approximately given as

$$E[h_i^2]=[(i+1)/3], \qquad \text{for} \qquad i=1,2,\dots,M-1 \tag{20a}$$

$$E[h_M^2]=\mathrm{M}/3 \tag{20b}$$

$$E[h_ih_j]=i/3, \qquad \text{for} \qquad j>i \tag{20c}$$

Using the representations of Eqs. (17) and (19), Eq. (13) becomes

$$\hat{x}^*(k+1)=(I+\Delta R_2)\{A(I+\Delta H_1)\hat{x}^*(k)+K(I+\Delta H_3) \times (I+\Delta R_1)[y^*(k)-C(I+\Delta H_2)\hat{x}^*(k)]\} \tag{21}$$

Upon substitution of $Cx(k)+e(k)$ for $y^*(k)$, we have

$$\hat{x}^*(k+1)=(A-KC)\hat{x}^*(k)+\Delta(R_2A+AH_1-KH_3C-KR_1 \times C-KCH_2-R_2KC)\hat{x}^*(k)+KCx(k)+\Delta(KH_3C +KR_1C+R_2KC)x(k)+Ke(k)+\Delta(KH_3+KR_1 +R_2K)e(k) \tag{22}$$

+ terms of higher-order Δ^2, Δ^3, Δ^4.

Combining Eqs. (8a) and (22), it follows that

$$\begin{bmatrix} x(k+1) \\ \hat{x}^*(k+1) \end{bmatrix} \cong \begin{bmatrix} A & 0 \\ KC & A-KC \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \left[\begin{array}{c|c} 0 & 0 \\ \hline \Delta(KH_3C + KR_1C + R_2KC) & \Delta(R_2A + AH_1 - KH_3C - KR_1C - KCH_2 - R_2KC) \end{array}\right] \times \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \begin{bmatrix} 0 \\ K + \Delta(KH_3 + KR_1 + R_2K) \end{bmatrix} e(k) + \begin{bmatrix} I \\ 0 \end{bmatrix} v(k) \quad (23)$$

where 0 is the zero matrix, and I is the identity matrix with appropriate dimensions. In Eq. (23), because these terms of higher-order Δ^2, Δ^3, and Δ^4 are small, they are ignored. Let

$$\bar{A} = \begin{bmatrix} A & 0 \\ KC & A-KC \end{bmatrix}$$

$$\bar{B} = \left[\begin{array}{c|c} 0 & 0 \\ \hline \Delta(KH_3C + KR_1C + R_2KC) & \Delta(R_2A + AH_1 - KH_3C - KR_1C - KCH_2 - R_2KC) \end{array}\right]$$

$$\bar{C} = \begin{bmatrix} 0 \\ K + \Delta(KH_3 + KR_1 + R_2K) \end{bmatrix}$$

$$\bar{D} = \begin{bmatrix} I \\ 0 \end{bmatrix}$$

Then Eq. (23) can be rewritten as

$$\begin{bmatrix} x(k+1) \\ \hat{x}^*(k+1) \end{bmatrix} \cong \bar{A} \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \bar{B} \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \bar{C}e(k) + \bar{D}v(k) \quad (24)$$

Next, we shall analyze the estimation error due to finite-wordlength implementation of digital Kalman filters. Although the analysis gives precisely the error due to finite-wordlength effects, the calculation is slightly cumbersome. Thus, by ignoring higher quantities in Eq. (22), the actual estimation error of the digital Kalman filter can be expressed by

$$\begin{aligned} \hat{x}^*(k+1) - x(k+1) &\cong [-(A-KC) \quad (A-KC)] \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} \\ &+ [\Delta(KH_3C + KR_1C + R_2KC) \quad \Delta(R_2A + AH_1 \\ &- KH_3C - KR_1C - KCH_2 - R_2KC)] \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} \\ &+ [K + \Delta(KH_3 + KR_1 + R_2K)]e(k) - v(k) \\ &= A^{\#} \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + B^{\#} \begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + C^{\#}e(k) - D^{\#}v(k) \end{aligned} \quad (25)$$

where

$$A^{\#} = [-(A-KC) \quad (A-KC)]$$

$$B^{\#} = [\Delta(KH_3C + KR_1C + R_2KC) \; \Delta(R_2A + AH_1 - KH_3C - KR_1C - KCH_2 - R_2KC)]$$

$$C^{\#} = [K + \Delta(KH_3 + KR_1 + R_2K)]$$

$$D^{\#} = I$$

Because $\bar{B}$, $\bar{C}$, $B^{\#}$, and $C^{\#}$ are stochastic, $\|\bar{B}\|$, $\|\bar{C}\|$, $\|B^{\#}\|$, and $\|C^{\#}\|$ can be evaluated as follows:

$$\|\bar{B}\| = \sqrt{\lambda_{\max}(E[\bar{B}^T\bar{B}])} = \sqrt{\lambda_{\max}\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}} \quad (26)$$

where

$$F_{11} = 2^{-2W}\{C^TE[H_3K^TKH_3]C + C^TE[R_1K^TKR_1]C + C^TK^TE[R_2R_2]KC\}$$

$$F_{12} = 2^{-2W}\{C^TK^TE[R_2R_2]A\} - F_{11}$$

$$F_{21} = 2^{-2W}\{A^TE[R_2R_2]KC\} - F_{11}$$

$$F_{22} = 2^{-2W}\{A^TE[R_2R_2]A + E[H_1A^TAH_1] + E[H_2C^TK^TKCH_2]\} - F_{11} - F_{12} - F_{21}$$

Referring to Appendix A, the values of $E[H_3H_3]$, $E[R_1K^TKR_1]$, $E[R_2R_2]$, etc., can be obtained easily.

$$\|\bar{C}\| = \sqrt{\lambda_{\max}(E[\bar{C}^T\bar{C}])} = \sqrt{\lambda_{\max}(Q)} \quad (27)$$

where

$$Q = K^TK + 2^{-2W}\{E[H_3K^TKH_3] + E[R_1K^TKR_1] + K^TE[R_2R_2]K\}$$

Doing a simple calculation, it is found that the values of $\|B^{\#}\|$ and $\|C^{\#}\|$ are the same as the values of $\|\bar{B}\|$ and $\|\bar{C}\|$, respectively.

Since $\bar{D}$, $A^{\#}$, and $D^{\#}$ are deterministic, from Eqs. (16), the values of $\|\bar{D}\|$, $\|A^{\#}\|$, and $\|D^{\#}\|$ are obtained and listed below

$$\|\bar{D}\| = \sqrt{\lambda_{\max}(\bar{D}^T\bar{D})} = 1 \quad (28)$$

$$\|A^{\#}\| = \sqrt{\lambda_{\max}\begin{bmatrix} (A^TA - C^TK^TA - A^TKC + C^TK^TKC) & -(A^TA - C^TK^TA - A^TKC + C^TK^TKC) \\ -(A^TA - C^TK^TA - A^TKC + C^TK^TKC) & (A^TA - C^TK^TA - A^TKC + C^TK^TKC) \end{bmatrix}} \quad (29)$$

$$\|D^{\#}\| = \sqrt{\lambda_{\max}[(D^{\#})^TD^{\#}]} = 1 \quad (30)$$

It is easy to see that, if $\bar{A}$ is a stable transition matrix for digital filters in infinite precision, then

$$\|\bar{A}^k\| \le mr^k, \qquad k > 0 \quad (31)$$

for some constant $m > 0$ and $0 \le r < 1$. Simply choose

$$r = \max_i |\lambda_i(\bar{A})|$$

where $\lambda_i(\bar{A})$, for $i = 1, 2, \ldots, n$, denotes the eigenvalues of A. That is, r is the absolute value of the eigenvalue of $\bar{A}$ (or the pole of the digital filter) nearest the unit circle. An estimate of m can be made from $\|\bar{A}^k\|/r^k \le m$ for all k. How to get m is

sometimes very difficult. Fortunately, m can be obtained with the aid of a computer.

To derive the stability condition and the actual estimation error bound under the finite-wordlength effects, the Bellman-Gronwall lemma in discrete form is employed. The lemma is listed as follows:

Lemma 1.[15] Let $(u(k))_0^\infty$, $(f(k))_0^\infty$, and $(h(k))_0^\infty$ be real-valued sequences on the set of the positive integer Z_+. Let

$$h(k) \geq 0, \qquad \forall k \in Z_+ \tag{32}$$

Under these conditions, if

$$u(k) \leq f(k) + \sum_{i=0}^{k-1} h(i)u(i), \qquad k = 0, 1, 2, \ldots \tag{33}$$

then

$$u(k) \leq f(k) + \sum_{i=0}^{k-1} \left\{ \prod_{j=i+1}^{k-1} [1 + h(j)]h(i)f(i) \right\}$$

$$k = 0, 1, 2, \ldots \tag{34}$$

where

$$\prod_{j=i+1}^{k-1} [1 + h(j)]$$

is set equal to 1 when $i = k - 1$.

Remarks.

1) If for some constant h, $h(i) \leq h$, $\forall_i$, then Eq. (34) becomes

$$u(k) \leq f(k) + h \sum_{i=0}^{k-1} (1 + h)^{k-1-i} f(i) \tag{35a}$$

2) If for some constant f, $f(i) \leq f$, $\forall_i$, then Eq. (34) becomes

$$u(k) \leq f \prod_{i=1}^{k-1} [1 + h(i)] \tag{35b}$$

We make the observation that Eq. (24) has the perturbation term related to

$$\begin{bmatrix} x(k) \\ x^*(k) \end{bmatrix}$$

It is feasible for us to apply the Bellman-Gronwall lemma to obtain the stability criterion. Based on Lemma 1 and the preceding definitions, we can relate a sufficient condition of stability on the estimator [Eq. (24)] to roundoff errors in the following theorem.

Theorem 1. Consider the state-estimator system [Eq. (24)] with the induced norms [Eq. (26)], $\|e(k)\| = g_1$ and $\|v(k)\| = g_2$, and suppose the transition matrix $\bar{A}$ fulfills the requirement of Eq. (31). If the stability inequality

$$r + m\|\bar{B}\| < 1 \tag{36}$$

is satisfied, then the deteriorated state estimator due to roundoff errors is still stable.

Proof. See Appendix B.

Remarks.

1) The relationship between the location of the pole nearest to the unit circle and the computational roundoff errors is revealed. From the stability inequality [Eq. (36)], it is seen that the smaller r is, the stronger the stability will be.

2) From Eqs. (26), (31), and (36), the wordlength W can be determined to guarantee the stability of the deteriorated system under the finite-wordlength effect.

3) For a given wordlength W, the stability inequality [Eq. (36)] can be used as a criterion to test the stability of the estimator system deteriorated by the roundoff noise.

4) From Theorem 1, it is assumed that $m\|\bar{B}\|$ is evaluated, and all of the eigenvalues of the Kalman filter must be inside the disk with radius $r < 1 - m\|\bar{B}\|$ to guarantee the stability of the Kalman filter with floating-point computation. However, if the eigenvalues of $A - K(k)C$ in Eq. (9) are not all within the disk with radius r, a scheme proposed by Anderson[17] is employed to treat this problem. Suppose we artificially multiply the covariances S_1 and S_2 in the system of Eqs. (5) with $(1/r)^{2k}$ and $r < 1$; i.e., $S_1' = S_1\,(1/r)^{2k}$ and $S_2' = S_2\,(1/r)^{2k}$. Anderson has shown that the system of Eq. (9), the computation of Kalman gain K [Eq. (11)], and the P in the Riccati equation (12) can be changed to

$$\hat{x}(k+1) = [A - K_r(k)C]\hat{x}(k) + K_r(k)y^*(k) \tag{37}$$

$$K_r(k) = AP(k)C^T[S_2(1/r)^{2k} + CP(k)C^T]^{-1} \tag{38}$$

$$P(k+1) = AP(k)A^T + S_1(1/r)^{2k} - AP(k)C^T \times [S_2(1/r)^{2k} + CP(k)C^T]^{-1}CP^T(k)A^T \tag{39}$$

Then the estimation error $\epsilon(k)$ of the filter in Eq. (10) will converge at least as fast as r^k when k increases; i.e., all of the eigenvalues of $(A - K_r(k)C)$ are inside the disk with radius r.

After deriving the sufficient condition under which the deteriorated Kalman filter is stable, we can find one form of bound by the estimate given in Theorem 2.

Theorem 2. Consider the state-estimator system of Theorem 1. If the stability criterion of Eq. (36) is satisfied when $k \to \infty$, then the upper bound of

$$\left\| \begin{matrix} x(k) \\ \hat{x}^*(k) \end{matrix} \right\|$$

can be evaluated as

$$\frac{m(g_1\|\bar{C}\| + g_2\|\bar{D}\|)}{1 - (r + m\|\bar{B}\|)} \tag{40}$$

and the actual estimation error $\|\hat{x}^*(k+1) - x(k+1)\|$ is bounded by

$$(\|A^\#\| + \|B^\#\|)\frac{m(g_1\|\bar{C}\| + g_2\|\bar{D}\|)}{1 - (r + m\|\bar{B}\|)} + g_1\|C^\#\| + g_2\|D^\#\| \tag{41}$$

where $g_1 = \|e(k)\|$ and $g_2 = \|v(k)\|$.

Proof. See Appendix C.

We do not claim that the bound given here is the ultimate tool for solving the problem of the minimal wordlength of implementing digital Kalman filters; however, it can be a reference for a conservative design. This paper does permit an approximate analysis of the performance/cost tradeoff for wordlength design choices.

IV. Numerical Example

To illustrate the stability criterion proposed herein, we consider an inertial navigation system (INS) due to wind-induced bending. The state-space representation of the system is[19]

$$\begin{bmatrix} \frac{dp_b(t)}{dt} \\ \frac{dv_b(t)}{dt} \\ \frac{da_b(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha & \beta & \gamma \end{bmatrix} \begin{bmatrix} p_b(t) \\ v_b(t) \\ a_b(t) \end{bmatrix} + v(t) \tag{42}$$

where $p_b(t)$, $v_b(t)$, and $a_b(t)$ are the horizontal displacement, velocity, and acceleration, respectively. The white Gaussian

noise processes $v(\cdot,\cdot)$ is of appropriate strength to yield the desired root-mean-square (rms) value of wind, σ_{wind}, with correlation time $1/\lambda$. If the bending dynamics model is second order with undamped natural frequency ω_n and damping ratio ζ, then the three parameters α, β, and γ are specified by $\alpha=-\lambda\omega_n^2$, $\beta=-\omega_n^2-2\zeta\omega_n$, and $\gamma=-2\zeta\omega_n-\lambda$.

The discrete state-space representation of the INS with $\omega_n=3$ rad/s, $\lambda=0.6667$ 1/s, and $\zeta=0.5$ can be expressed by

$$\begin{bmatrix} p_b(k+1) \\ v_b(k+1) \\ a_b(k+1) \end{bmatrix} = \begin{bmatrix} 0.999985 & 0.024972 & 0.000303 \\ -0.001818 & 0.996652 & 0.023861 \\ -0.143166 & -0.264288 & 0.909161 \end{bmatrix} \times \begin{bmatrix} p_b(k) \\ v_b(k) \\ a_b(k) \end{bmatrix} + v(k) \tag{43a}$$

$$y(k)=[-1 \quad 0 \quad 0]\begin{bmatrix} p_b(k) \\ v_b(k) \\ a_b(k) \end{bmatrix} + \bar{e}(k) \tag{43b}$$

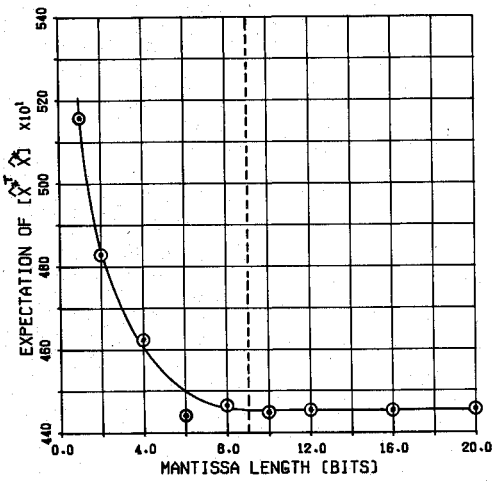


Fig. 1 Expectation of $[x^{*T}\hat{x}^*]$ (solid curve) with the wordlength bound (dashed line).

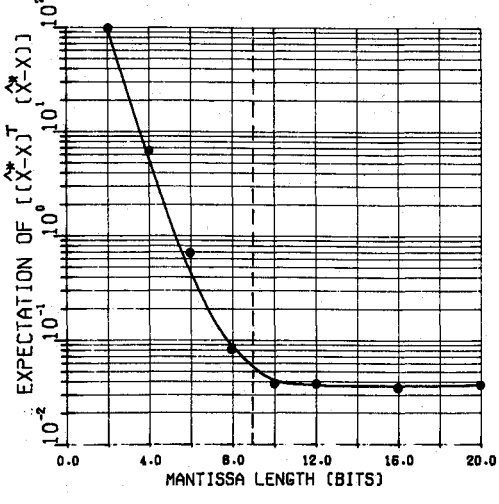


Fig. 2 Expectation of $[(\hat{x}^*-x)^T(\hat{x}^*-x)]$ (solid curve) with the wordlength bound (dashed line).

where $v(k)$ and $\bar{e}$ are white noise processes with zero-mean values and the covariances

$$S_1=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.01 & 0.002 \\ 0 & 0.002 & 0.0004 \end{bmatrix}$$

$$S_2=0.25$$

Hence,

$$A=\begin{bmatrix} 0.999985 & 0.024972 & 0.000303 \\ -0.001818 & 0.996652 & 0.023861 \\ -0.143166 & -0.264288 & 0.909161 \end{bmatrix}$$

$$C=[-1 \quad 0 \quad 0]$$

From Eqs. (26) and (31), we have

$$\|\bar{B}\|=1.2644\cdot 2^{-W} \tag{44}$$

$$m=4.6564 \tag{45}$$

$$r=0.985 \tag{46}$$

Using the robust stability criterion of Eq. (36), we obtain

$$W>8.479 \tag{47}$$

From the preceding analysis, we suggest requiring the wordlength to be greater than or equal to 9 bits; otherwise, it may lead to an unstable response. Substituting these values of $\|\bar{C}\|$, $\|\bar{D}\|$, $\|A^{\#}\|$, $\|B^{\#}\|$, $\|C^{\#}\|$, and $\|D^{\#}\|$, we can estimate the upper bounds of

$$\left\| \begin{matrix} x(k) \\ \hat{x}^*(k) \end{matrix} \right\|$$

and $\|\hat{x}^*(k+1)-x(k+1)\|$.

The estimator system using the Kalman filter was simulated on an IBM-AT computer system with a very long wordlength, and the ADC for the output measurement was 10 bits. The wind input was simulated by use of a pseudorandom Gaussian number generator. Simulations were done for wordlengths from 1 to 20 bits, and the results are shown in Figs. 1 and 2. Notice in Fig. 1 that the estimator is not well behaved and in Fig. 2 that the estimation error is very large when the wordlength is small.

V. Conclusions

A sufficient condition has been presented to ensure the stability of the state estimator with a Kalman filter subjected to finite-wordlength effects. If the sufficiency condition is not satisfied, it does not necessarily imply system instability, but instability may really occur if the analog-to-digital converter (ADC) length and the mantissa length are too short in application. This criterion shows that a state estimator, which may be unstable with respect to roundoff errors of quantization, can be stable by choosing the appropriate ADC and mantissa lengths.

In fact, because the ADC noise and the computational roundoff errors will appear continuously, it is clear that the estimation error will not approach zero. Notice that, if the sufficient condition of Eq. (36) is satisfied, $\hat{x}^*(k)$ and $x(k)$ in Eq. (24) will be bounded, and the estimation error will be also bounded. This point agrees with the result mentioned in Theorem 1.

The results of the paper have been applied to the stability analysis of finite-wordlength Kalman filters with precomputed

steady-state coefficients. However, the results have not been applied to the design of time-varying filters. This issue is under investigation.

Appendix A: Error Representations of Floating-Point Computation

The error representations of addition and multiplication with floating-point computation are given here.[11,13]

Wilkinson[13] provided the method of analyzing the errors of floating-point arithmetical results. The rounded floating-point sum of two numbers, a and b, can be expressed by

$$\text{fl}(a+b) = (a+b)(1+r) \tag{A1}$$

where $\text{fl}(\cdot)$ is the floating-point operator, and r is a random variable uniformly distributed between -2^{-W} and 2^{-W}, W being the wordlength of the mantissa. If a and b are two M-vectors, then the floating-point sum is expressed by

$$\begin{aligned}\text{fl}(a+b) &= \begin{bmatrix} (a_1+b_1)(1+r_1) \\ (a_2+b_2)(1+r_2) \\ \vdots \\ (a_M+b_M)(1+r_M) \end{bmatrix} \\ &= [I+R][a+b] \end{aligned} \tag{A2}$$

where a_i and b_i ($i = 1,2,3, \ldots, M$) are elements of M-vectors a and b, respectively, I is an identity, and

$$R = \text{diag}[r_1\ r_2\ r_3 \ldots r_M]$$

Here r_i, $i = 1, 2, \ldots, M$, are mutually independent and are distributed between -2^{-W} and 2^{-W}.

The rounded floating-point product of two numbers, a and b, is represented by

$$\text{FL}(ab) = ab(1+\delta) \tag{A3}$$

where $\text{FL}(\cdot)$ is the floating-point operator and δ is also distributed between -2^{-W} and 2^{-W}. Similarly, the rounded inner product of two M-vectors is

$$\begin{aligned}\text{fl}\left(\sum_{i=1}^{M} a_i b_i\right) &= \text{fl}(\text{FL}(a_1b_1) + \cdots \\ &+ \text{fl}(\text{FL}(a_{M-2}b_{M-2}) + \text{fl}(\text{FL}(a_{M-1}b_{M-1}) \\ &+ \text{FL}(a_Mb_M))\cdots) \\ &= a_1b_1(1+\delta_1)(1+r_1) + \cdots \\ &+ a_ib_i(1+\delta_i)\prod_{k=1}^{i}(1+r_k) + \cdots \\ &+ a_Mb_M(1+\delta_M)\prod_{k=1}^{M-1}(1+r_k) \end{aligned} \tag{A4}$$

Each δ_i and r_i can be considered to be mutually independent and uniformly distributed between -2^{-W} and 2^{-W}. Let us define

$$1+h_i = (1+\delta_i)\prod_{k=1}^{i}(1+r_k), \quad \text{for } i = 1, 2, \ldots, M-1 \tag{A5a}$$

$$1+h_M = (1+\delta_M)\prod_{k=1}^{M-1}(1+r_k) \tag{A5b}$$

If the preceding equations are expanded directly in order to obtain exact h_i, the expression of h_i will be rather complicated. Therefore, we ignore the small higher-order terms in Eqs. (A5), and each h_i can be expressed approximately by

$$h_i = \delta_i + \sum_{k=1}^{i} r_k, \qquad \text{for } i = 1, 2, \ldots, M-1$$

$$h_M = \delta_M + \sum_{k=1}^{M-1} r_k$$

Then each h_i has zero mean and the variances are approximately

$$E[h_i^2] = (i+1)2^{-2W}/3, \qquad \text{for } i = 1, 2, \ldots, M-1$$

$$E[h_M^2] = M2^{-2W}/3$$

$$E[h_ih_j] = i2^{-2W}/3, \qquad \text{for } j > i$$

The rounded multiplication of an M-vector x by an $M \times M$ matrix A can be written by

$$\text{FL}(Ax) = Ax + \begin{bmatrix} a_{11}h_{11} & a_{12}h_{12} & \cdots & a_{1M}h_{1M} \\ a_{21}h_{21} & a_{22}h_{22} & \cdots & a_{2M}h_{2M} \\ \vdots & \vdots & & \vdots \\ a_{M1}h_{M1} & a_{M2}h_{M2} & \cdots & a_{MM}h_{MM} \end{bmatrix} x \tag{A6}$$

where a_{ij} is an element of A, and h's in different rows are statistically independent and uncorrelated. Hence, we substitute each M-tuple $(h_1, h_2, \ldots, h_M)$ for each M-tuple $(h_{i1}, h_{i2}, \ldots, h_{iM})$ and obtain the permitted representation

$$\text{FL}(Ax) = Ax + A\ \text{diag}[h_1\ h_2 \ldots h_M]x$$

which is convenient for manipulation in the derivation of our results.

Appendix B: Proof of Theorem 1

Consider the combined state Equation (24)

$$\begin{bmatrix} x(k+1) \\ \hat{x}^*(k+1) \end{bmatrix} \cong \bar{A}\begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \bar{B}\begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \bar{C}e(k) + \bar{D}v(k) \tag{B1}$$

where

$$\bar{A} = \begin{bmatrix} A & 0 \\ KC & A-KC \end{bmatrix}$$

$$\bar{B} = \left[\begin{array}{c|c} 0 & 0 \\ \hline \Delta(KH_3C & \Delta(R_2A + AH_1 - KH_3C \\ +KR_1C + R_2KC) & -KR_1C - KCH_2 - R_2KC) \end{array}\right]$$

$$\bar{C} = \begin{bmatrix} 0 \\ K + \Delta(KH_3 + KR_1 + R_2K) \end{bmatrix}$$

$$\bar{D} = \begin{bmatrix} I \\ 0 \end{bmatrix}$$

Solving the preceding difference equation, we obtain the solution

$$\begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} \cong \bar{A}^k\begin{bmatrix} x(0) \\ \hat{x}^*(0) \end{bmatrix} + \sum_{i=0}^{k-1}\bar{A}^{k-1-i}\bar{B}\begin{bmatrix} x(k) \\ \hat{x}^*(k) \end{bmatrix} + \sum_{i=0}^{k-1}\bar{A}^{k-1-i}\bar{C}e(k) + \sum_{i=0}^{k-1}\bar{A}^{k-1-i}\bar{D}v(k) \tag{B2}$$

Taking norms, we get

$$\left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| \le \|\bar{A}^k\| \left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + \sum_{i=0}^{k-1} \|\bar{A}^{k-1-i}\|\,\|\bar{B}\| \left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| + \sum_{i=0}^{k-1} \|\bar{A}^{k-1-i}\|\,\|\bar{C}\|\,\|e(k)\| + \sum_{i=0}^{k-1} \|\bar{A}^{k-1-i}\|\,\|\bar{D}\|\,\|v(k)\| \tag{B3}$$

Using Eq. (31), it is found that

$$\left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| \le mr^k \left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + \sum_{i=0}^{k-1} mr^{k-1-i}\|\bar{B}\| \left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| + \sum_{i=0}^{k-1} mr^{k-1-i}\|\bar{C}\|\,\|e(k)\| + \sum_{i=0}^{k-1} mr^{k-1-i}\|\bar{D}\|\,\|v(k)\| \tag{B4}$$

Dividing both sides of Eq. (B4) by r^k, we obtain

$$r^{-k}\left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| \le m\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + m\frac{r^{-k}-1}{1-r}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) + \frac{m\|\bar{B}\|}{r}\sum_{i=0}^{k-1} r^{-i}\left\|\begin{matrix}x(i)\\ \hat{x}^*(i)\end{matrix}\right\| \tag{B5}$$

Applying the Remarks of Lemma 1, we have

$$\begin{aligned} r^{-k}\left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| &\le m\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + m\frac{r^{-k}-1}{1-r}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \\ &+ \frac{m\|\bar{B}\|}{r}\sum_{i=0}^{k-1}\left(1+\frac{m\|\bar{B}\|}{r}\right)^{k-1-i}\left[m\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + m\frac{r^{-i}-1}{1-r}(g_1\|\bar{C}\| + g_2\|\bar{D}\|)\right] \\ &= m\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + m\frac{r^{-k}-1}{1-r}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \\ &+ \left(\frac{m^2\|\bar{B}\|}{r}\right)\left\{\frac{1-[1+(m\|\bar{B}\|/r)]^k}{1-[1+(m\|\bar{B}\|/r)]}\right\}\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| \\ &+ \left(\frac{-m^2\|\bar{B}\|}{r}\right)\left(\frac{1}{1-r}\right)\left\{\frac{1-[1+(m\|\bar{B}\|/r)]^{-k}}{1-[1+(m\|\bar{B}\|/r)]^{-1}}\right\} \\ &\times\left(1+\frac{m\|\bar{B}\|}{r}\right)^{k-1}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \\ &+ \left(\frac{m^2\|\bar{B}\|}{r}\right)\left(\frac{1}{1-r}\right)\left(\frac{1-(r+m\|\bar{B}\|)^{-k}}{1-(r+m\|\bar{B}\|)^{-1}}\right) \\ &\times\left(1+\frac{m\|\bar{B}\|}{r}\right)^{k-1}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \end{aligned} \tag{B6}$$

Next, multiplying either side of Eq. (B6), it follows that

$$\begin{aligned} \left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| &\le mr^k\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| + m\left(\frac{1-r^k}{1-r}\right)(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \\ &+ \left(\frac{m^2\|\bar{B}\|}{r}\right)\left\{\frac{r^k-(r+m\|\bar{B}\|)^k}{1-[1+(m\|\bar{B}\|/r)]}\right\}\left\|\begin{matrix}x(0)\\ \hat{x}^*(0)\end{matrix}\right\| \\ &+ \left(\frac{-m^2\|\bar{B}\|}{1-r}\right)\left\{\frac{1-[1+(m\|\bar{B}\|/r)]^{-k}}{1-[1+(m\|\bar{B}\|/r)]^{-1}}\right\}(r+m\|\bar{B}\|)^{k-1} \\ &\times(g_1\|\bar{C}\| + g_2\|\bar{D}\|) + \left(\frac{m^2\|\bar{B}\|}{1-r}\right)\left(\frac{1-(r+m\|\bar{B}\|)^{-k}}{1-(r+m\|\bar{B}\|)^{-1}}\right) \\ &\times(r+m\|\bar{B}\|)^{k-1}(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \end{aligned} \tag{B7}$$

When $k\to\infty$, if $r + m\|\bar{B}\| < 1$, then the state estimator will be robustly stable.

Q.E.D.

Appendix C: Proof of Theorem 2

As $k\to\infty$, Eq. (B7) becomes

$$\begin{aligned} \left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\| &\le \left(\frac{m}{1-r}\right)(g_1\|\bar{C}\| + g_2\|\bar{D}\|) + \left(\frac{-m^2\|\bar{B}\|}{1-r}\right) \\ &\times\left(\frac{(r+m\|\bar{B}\|)^{-1}}{1-(r+m\|\bar{B}\|)^{-1}}\right)(g_1\|\bar{C}\| + g_2\|\bar{D}\|) \\ &= \frac{m(g_1\|\bar{C}\| + g_2\|\bar{D}\|)}{1-(r+m\|\bar{B}\|)} \end{aligned} \tag{C1}$$

Hence, the bound of

$$\left\|\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right\|$$

is

$$\frac{m(g_1\|\bar{C}\| + g_2\|\bar{D}\|)}{1-(r+m\|\bar{B}\|)}$$

as $k\to\infty$.

From Eq. (25), the actual estimation error is

$$\begin{aligned} \hat{x}^*(k+1) - x(k+1) &\cong [-(A-KC) \quad (A-KC)]\left[\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right] \\ &+ [\Delta(KH_3C + KR_1C + R_2KC)\Delta(R_2A + AH_1 - KH_3C \\ &- KR_1C - KCH_2 - R_2KC)]\left[\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right] \\ &+ [K + \Delta(KH_3 + KR_1 + R_2K)]e(k) - v(k) \\ &= A^{\#}\left[\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right] + B^{\#}\left[\begin{matrix}x(k)\\ \hat{x}^*(k)\end{matrix}\right] + C^{\#}e(k) - D^{\#}v(k) \end{aligned} \tag{C2}$$

where

$$A^{\#} = [-(A-KC)\ (A-KC)]$$

$$B^{\#} = [\Delta(KH_3C + KR_1C + R_2KC)\Delta(R_2A + AH_1 - KH_3C - KR_1C - KCH_2 - R_2KC)]$$

$$C^{\#} = [K + \Delta(KH_3 + KR_1 + R_2K)]$$

$$D^{\#} = I$$

Taking norms, we have

$$\|\hat{x}^*(k+1) - x(k+1)\| \le \|A^{\#}\| \left\| \begin{matrix} x(k) \\ \hat{x}^*(k) \end{matrix} \right\| + \|B^{\#}\| \left\| \begin{matrix} x(k) \\ \hat{x}^*(k) \end{matrix} \right\| + \|C^{\#}\|\ \|e(k)\| + \|D^{\#}\|\ \|v(k)\| \qquad \text{(C3)}$$

Then Eqs. (C1) and (C3) together imply that

$$\|\hat{x}^*(k+1) - x(k+1)\| \le (\|A^{\#}\| + \|B^{\#}\|) \times \frac{m(g_1\|\bar{C}\| + g_2\|\bar{D}\|)}{1-(r+m\|\bar{B}\|)} + g_1\|C^{\#}\| + g_2\|D^{\#}\| \qquad \text{(C4)}$$

as $k \to \infty$.

Q.E.D.

Acknowledgments

The authors would like to thank Tatung Company, Taipei, Tawain, Republic of China, for its financial support, and the reviewers for their helpful suggestions to improve the manuscript.

References

[1]Amit, G., and Shaked, U., "Small Roundoff Noise Realization of Fixed-Point Digital Filters and Controller," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-36, June 1988, pp. 880–891.

[2]Rabiner, L. R., and Gold, B., *Theory and Application of Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, NJ, 1975.

[3]Oppenheim, A. V., and Schafer, R. W., *Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, NJ, 1975.

[4]Williamson, D., and Sridharan, S., "An Approach to Coefficient Wordlength Reduction in Digital Filters," *IEEE Transactions on Circuits and Systems,* Vol. CAS-32, Sept. 1985, pp. 893–903.

[5]Bomar, B., "Computationally Efficient Low Roundoff Noise Second-Order State-Space Structures," *IEEE Transactions on Circuits and Systems,* Vol. CAS-33, Jan. 1986, pp. 35–41.

[6]Bhaskar Rao, D. V., "Analysis of Coefficient Quantization Errors in State-Space Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-34, Feb. 1986, pp. 131–139.

[7]Stripad, A. B., "Performance Degradation in Digitally Implemented Kalman Filters," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. AES-17, Sept. 1981, pp. 626–634.

[8]Moroney, P., Willsky, A. S., and Houpt, P. K., "The Digital Implementation of Control Compensators: The Coefficient Wordlength Issue," *IEEE Transactions on Automatic Control,* Vol. AC-25, Aug. 1980, pp. 621–630.

[9]Moroney, P., Willsky, A. S., and Houpt, P. K., "Roundoff Noise and Scaling in the Digital Implementation of Control Compensators," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-31, Dec. 1983, pp. 1464–1477.

[10]Higgins, W. E., and Munson, D. C., Jr., "Noise Reduction Strategies for Digital Filters: Error Spectrum Shaping Versus the Optimal Linear State-Space Formulation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-30, Dec. 1982, pp. 963–973.

[11]Rink, R. E., and Chong, H. Y., "Performance of State Regulator Systems with Floating-Point Computation," *IEEE Transactions on Automatic Control,* Vol. AC-24, June 1979, pp. 411–421.

[12]Van Wingerden, A. J. M., and De Koning, W. L., "The Influence of Finite Wordlength on Digital Optimal Control," *IEEE Transactions on Automatic Control,* Vol. AC-29, May 1984, pp. 385–391.

[13]Wilkinson, J. H., *Rounding Errors in Algebraic Processes,* Prentice-Hall, Englewood Cliffs, NJ, 1963.

[14]Goodwin, G. C., and Sin, K. S., *Adaptive Filtering Prediction and Control,* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[15]Desoer, C. A., and Vidyasagar, M., *Feedback Systems: Input-Output Properties,* Academic, New York, 1975.

[16]Astrom, K. J., and Wittenmark, B., *Computer Controlled Systems: Theory and Design,* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[17]Anderson, B. D. O., "Exponential Data Weighting in the Kalman-Bucy Filter," *Information Sciences,* Vol. 5, 1973, pp. 217–230.

[18]Williamson, D., "Finite Wordlength Design of Digital Kalman Filters for State Estimation," *IEEE Transactions on Automatic Control,* Vol. AC-30, No. 10, 1985, pp. 930–939.

[19]Maybeck, P. S., *Stochastic Models, Estimation and Control,* Vol. 1, Academic, New York, 1979.