JOINTLY OPTIMIZED ERROR-FEEDBACK AND REALIZATION FOR ROUNDOFF NOISE MINIMIZATION IN STATE-ESTIMATE FEEDBACK DIGITAL CONTROLLERS

Takao Hinamoto, Keijiro Kawai and Masayoshi Nakamoto

Graduate School of Engineering, Hiroshima University
Higashi-Hiroshima 739-8527, Japan
phone:+81-82-424-7672, fax:+81-82-422-7195
email:{hinamoto, may}@hiroshima-u.ac.jp

Wu-Sheng Lu

Dept. of Elect. & Comput. Engineering,
University of Victoria
Victoria, B.C., V8W 3P6, Canada
email: wslu@ece.uvic.ca

ABSTRACT

The joint optimization problem of error-feedback and realization for the closed-loop system with a state-estimate feedback digital controller is investigated where the main objective is to minimize the effects of roundoff noise at the closed-loop system output subject to $l_2$-norm dynamic-range scaling constraints. It is shown that the problem can be converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem at hand is then solved iteratively by employing an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. Analytical details are given as to how the proposed technique can be applied to the cases where the error-feedback matrix is a general, diagonal, or scalar matrix. A numerical example is presented to illustrate the utility of the proposed technique.

1. INTRODUCTION

Due to the finite precision nature of computer arithmetic, the output roundoff noise of a fixed-point IIR digital filter usually arises. This noise is critically dependent on the internal structure of an IIR digital filter [1],[2]. Error feedback (EF) is known as an effective technique for reducing the output roundoff noise in an IIR digital filter [3]-[5]. Williamson [6] has reduced the output roundoff noise more effectively by choosing the filter structure and applying EF to the filter. Lu and Hinamoto [7] have developed a jointly optimized technique of EF and realization to minimize the effects of roundoff noise at the filter output subject to $l_2$-scaling constraints. Li and Gevers [8] have analyzed the output roundoff noise of the closed-loop system with a state-estimate feedback controller, and presented an algorithm for realizing the state-estimate feedback controller with minimum output roundoff noise under $l_2$-norm dynamic-range scaling constraints. Hinamoto and Yamamoto [9] have proposed a method for applying EF to a given closed-loop system with a state-estimate feedback controller.

This paper investigates the problem of jointly optimizing EF and realization for the closed-loop system with a state-estimate feedback controller so as to minimize the output roundoff noise subject to $l_2$-norm dynamic-range scaling constraints. To this end, an iterative technique which relies on an efficient quasi-Newton algorithm [10] is developed. Our computer simulation results demonstrate the validity and effectiveness of the proposed technique.

2. ROUNDOFF NOISE ANALYSIS

Let a stable, controllable and observable linear discrete-time system be described by

$$x(k+1) = A_o x(k) + b_o u(k)$$
$$y(k) = c_o x(k)$$

(1)

where $x(k)$ is an $n \times 1$ state-variable vector, $u(k)$ is a scalar input, $y(k)$ is a scalar output, and $A_o$, $b_o$ and $c_o$ are real constant matrices of appropriate dimensions. The transfer function of the linear system in (1) is given by

$$H_o(z) = c_o (z I_n - A_o)^{-1} b_o.$$  

(2)

If a regulator is designed by using the full-order state observer, we obtain a state-estimate feedback controller as

$$\hat{x}(k+1) = F_o \hat{x}(k) + b_o u(k) + g_o y(k)$$
$$= R_o \hat{x}(k) + b_r r(k) + g_o y(k)$$

(3)

$$u(k) = -k_o \hat{x}(k) + r(k)$$

where $\hat{x}(k)$ is an $n \times 1$ state-variable vector in the full-order state observer, $g_o$ is an $n \times 1$ gain vector chosen so that all the eigenvalues of $F_o = A_o - g_o c_o$ are inside the unit circle in the complex plane, $k_o$ is a $1 \times n$ state-feedback gain vector chosen so that each of the eigenvalues of $A_o - b_o k_o$ is at a desirable location within the unit circle, $r(k)$ is a scalar reference signal, and $R_o = F_o - b_o k_o$.

Performing quantization before matrix-vector multiplication, we can express the finite-word-length (FWL) implementation of (3) with error feedback as

$$\hat{x}(k+1) = R Q [\hat{x}(k)] + b_r r(k) + g_y y(k) + D e(k)$$
$$u(k) = -k Q [\hat{x}(k)] + r(k)$$

(4)

where $e(k) = \hat{x}(k) - Q [\hat{x}(k)]$ and $D$ is an $n \times n$ error feedback matrix. All coefficient matrices $R$, $b$, $g$ and $k$...
are assumed to have an exact fractional $B$, bit representation. The FWL state-variable vector $\hat{x}(k)$ and signal $u(k)$ all have a $B$ bit fractional representation, while the reference input $r(k)$ is a $(B_0 - B_x)$ bit fraction. The vector quantizer $Q[.]$ in (4) rounds the $B$ bit fraction $\hat{x}(k)$ to $(B_0 - B_x)$ bits after completing the multiplications and additions, where the sign bit is not counted. It is assumed that the roundoff error $e(k)$ can be modeled as a zero-mean noise process with covariance $\sigma^2 I_n$.

The closed-loop control system consisting of the linear system in (1) and the state-estimate feedback controller in (4) is illustrated in Fig. 1. This closed-loop system is described by

$$
\begin{bmatrix}
    x(k+1) \\
    \hat{x}(k+1)
\end{bmatrix} = \begin{bmatrix} A & -b_{o}k \\
    gc_{o} & R \end{bmatrix} \begin{bmatrix}
    x(k) \\
    \hat{x}(k)
\end{bmatrix} + \begin{bmatrix}
    b_{o} \\
    b
\end{bmatrix} r(k) + \begin{bmatrix}
    b_{o}k \\
    D-R
\end{bmatrix} e(k)
$$

(5)

$$
y(k) = \tau \begin{bmatrix}
    x(k) \\
    \hat{x}(k)
\end{bmatrix}
$$

where

$$
\bar{A} = \begin{bmatrix} A_o & -b_{o}k \\
    gc_{o} & R \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} b_{o} \\
    b \end{bmatrix}
$$

$$
\bar{B} = \begin{bmatrix} b_{o}k \\
    D-R
\end{bmatrix}, \quad \tau = [c_{o} \ 0].
$$

Let the transfer function from the roundoff noise $e(k)$ to the output $y(k)$ in (5) be defined by $G_D(z)$. Then

$$
G_D(z) = \tau (zI_{2n} - \bar{A})^{-1}\bar{B}.
$$

(6)

The noise gain $J(D) = \sigma_{out}^2/\sigma^2$ is then computed as

$$
J(D) = \text{tr}[W_D]
$$

(7)

with

$$
W_D = \frac{1}{2\pi j} \oint_{|z|=1} G_p(z)G_D(z)\frac{dz}{z}
$$

(8)

where $\sigma_{out}^2$ stands for the noise variance of the output. For tractability, we evaluate $J(D)$ in (7) by replacing $R$, $b$, $g$ and $k$ by $R_o$, $b_o$, $g_o$ and $k_o$, respectively. Define

$$
S = \begin{bmatrix} I_n & 0 \\
    I_n & -I_n \end{bmatrix}
$$

(9)

the transfer function $G_D(z)$ in (6) can be expressed as

$$
G_D(z) = \tau S(zI_{2n} - S^{-1}\bar{A}S)^{-1}S^{-1}\bar{B}
$$

$$
= \tau (zI_{2n} - \Phi)^{-1} \begin{bmatrix} b_{o}k_o \\
    F_o - D \end{bmatrix}
$$

$$
= c_o(zI_{n} - A_o + b_{o}k_o)^{-1}b_{o}k_0(zI_{n} - F_o)^{-1} - (zI_{n} - D)
$$

$$
= \tau (zI_{2n} - \Phi)^{-1}U(zI_{n} - D)
$$

(10)

where

$$
\Phi = \begin{bmatrix} A_o - b_{o}k_o & b_{o}k_o \\
    0 & F_o \end{bmatrix}, \quad U = \begin{bmatrix} 0 \\
    I_n \end{bmatrix}
$$

$$
$$

It is noted that the stability of the closed-loop control system is determined by the eigenvalues of matrix $\bar{A}$ in (5), or equivalently, those of matrix $\Phi$ in (10). This means that neither of the quantization error $e(k)$ and the error-feedback matrix $D$ affects the stability.

Substituting (10) into matrix $W_D$ in (8) gives

$$
W_D = (b_{o}k_0)^T W_1 b_{o}k_0 + (b_{o}k_0)^T W_2 (F_o - D)
$$

$$
+ (F_o - D)^T W_3 b_{o}k_0
$$

$$
+ (b_{o}k_0)^T W_4 (F_o - D)
$$

(11)

where

$$
W = \Phi^T W\Phi + \tau^T \tau, \quad W = \begin{bmatrix} W_1 & W_2 \\
    W_3 & W_4 \end{bmatrix}
$$

Since $W$ is positive semidefinite, it can be shown that there exists an $n \times n$ matrix $P$ such that $W_3 = W_4 P$. In addition, (11) can be written by virtue of $W_2 = W_3^T$ as

$$
W_D = (F_o + P b_{o}k_0 - D)^T W_4 (F_o + P b_{o}k_0 - D)
$$

$$
+ (b_{o}k_0)^T (W_1 - P^T W_4 P) b_{o}k_0.
$$

(12)

Alternatively, applying $z$-transform to the first equation in (5) under the assumption that $e(k) = 0$, we obtain

$$
\begin{bmatrix} X(z) \\
    \hat{X}(z) \end{bmatrix} = (zI - \bar{A})^{-1}\bar{B}R(z)
$$

(13)

where $X(z)$, $\hat{X}(z)$ and $R(z)$ represent the $z$-transforms of $x(k)$, $\hat{x}(k)$ and $r(k)$, respectively. Replacing $R$, $b$, $k$ and $g$ by $R_o$, $b_o$, $k_o$ and $g_o$, respectively, and then using

$$
S^{-1} \begin{bmatrix} X(z) \\
    \hat{X}(z) \end{bmatrix} = (zI_{2n} - S^{-1}\bar{A}S)^{-1}S^{-1}\bar{B}
$$

yield

$$
\hat{X}(z) = X(z) = F(z) R(z)
$$

(14)

where

$$
F(z) = [zI_{n} - (A_o - b_{o}k_o)^{-1}b_{o}].
$$
The controllability Gramian $K$ defined by
\[ K = \frac{1}{2\pi j} \int_{|z|=1} F(z) F^*(z) \frac{dz}{z} \]  
(15)
can be obtained by solving the Lyapunov equation
\[ K = (A_0 - b_x k_0^r) K (A_0 - b_x k_0^r)^T + b_x b_x^T. \]  
(16)

### 3. Roundoff Noise Minimization

Consider the system in (4) with $D = 0$ and denote it by $(R, b, g, k)_n$. By applying a coordinate transformation $\tilde{x}(k) = T^{-1} x(k)$ to the above system $(R, b, g, k)_n$, we obtain a new realization characterized by $(\tilde{R}, \tilde{b}, \tilde{g}, \tilde{k})_n$ where
\[ \tilde{R} = T^{-1} R T, \quad \tilde{b} = T^{-1} b \]
\[ \tilde{g} = T^{-1} g, \quad \tilde{k} = k T. \]  
(17)

For the system in (17), the counterparts of $W_i$ for $i = 1, 2, 3, 4$ are given by
\[ \tilde{W}_i = T^T W_i T, \]  
(18)
and the corresponding noise gain is given by
\[ J(D, T) = \text{tr}[\tilde{W}_D] \]  
(19)
where $\tilde{W}_D$ can be obtained using (11) as
\[ \tilde{W}_D = [T^{-1}(F_0 + P b_0 k_0) T - D]^T T^T W_4 T [T^{-1}(F_0 + P b_0 k_0) T - D] + T^T (b_0 k_0) T (W_1 - P^T W_4 P) b_0 k_0 T. \]
In addition, (15) can be written as
\[ \tilde{K} = T^{-1} K T^{-T}. \]  
(20)

As a result, the output roundoff noise minimization problem amounts to obtaining matrices $D$ and $T$ which jointly minimize $J(D, T)$ in (19) subject to the $l_2$-scaling constraints specified by
\[ (\tilde{K})_{ii} = (T^{-1} K T^{-T})_{ii} = 1, \quad i = 1, 2, \ldots, n. \]  
(21)

To deal with (21), we define
\[ \hat{T} = T^T K^{-\frac{1}{2}}. \]  
(22)

Then the $l_2$-scaling constraints in (21) can be written as
\[ (\hat{T}^{-T} \hat{T})_{ii} = 1, \quad i = 1, 2, \ldots, n. \]  
(23)

These constraints are always satisfied if $\hat{T}^{-1}$ assumes the form
\[ \hat{T}^{-1} = \left[ t_1 \left\| t_1 \right\|, \; t_2 \left\| t_2 \right\|, \; \ldots, \; t_n \left\| t_n \right\| \right]^T. \]  
(24)

Substituting (22) into (19), we obtain
\[ J(D, T) = \text{tr} \left[ \hat{T} (A - T^T D T^{-T})^T \tilde{W}_4 \right] \]
\[ \left. (A - T^T D T^{-T})^T \right] T + \hat{T} \hat{C} T^T \]  
(25)
where
\[ \hat{A} = K^{-\frac{1}{2}} (F_0 + P b_0 k_0) K^{-\frac{1}{2}}, \quad \hat{W}_4 = K^{-\frac{1}{2}} W_4 K^{-\frac{1}{2}}, \quad \hat{C} = K^{-\frac{1}{2}} (b_0 k_0) (W_1 - P^T W_4 P) b_0 k_0 K^{-\frac{1}{2}}. \]

From the foregoing arguments, the problem of obtaining matrices $D$ and $T$ that jointly minimize (19) subject to the scaling constraints in (21) is now converted into an unconstrained optimization problem of obtaining $D$ and $T$ that jointly minimize $J(D, T)$ in (25).

Let $x$ be the column vector that collects the variables in matrices $D$ and $[t_1, t_2, \ldots, t_n]$. Then $J(D, T)$ is a function of $x$, denoted by $J(x)$. The proposed algorithm starts with an initial point $x_0$ obtained from an initial assignment $D = T = I_n$. In the $k$th iteration, a quasi-Newton algorithm updates the most recent point $x_k$ to point $x_{k+1}$ as [10]
\[ x_{k+1} = x_k + \alpha_k d_k, \]  
(26)
where
\[ d_k = -S_k \nabla J(x_k) \]
\[ \alpha_k = \arg \left( \min_{\alpha} J(x_k + \alpha d_k) \right) \]
\[ S_{k+1} = S_k + \left( 1 + \frac{\gamma_i^2 \delta_k \delta_k}{\gamma_i \delta_i} \right) \frac{\delta_i \delta_i}{\gamma_i \delta_i} \]
\[ S_0 = I, \quad \delta_k = x_{k+1} - x_k, \quad \gamma_k = \nabla J(x_{k+1}) - \nabla J(x_k). \]

Here, $\nabla J(x)$ is the gradient of $J(x)$ with respect to $x$, and $S_k$ is a positive-definite approximation of the inverse Hessian matrix of $J(x_k)$. This iteration process continues until
\[ |J(x_{k+1}) - J(x_k)| < \varepsilon \]  
(27)
where $\varepsilon > 0$ is a prescribed tolerance.

In what follows, we derive closed-form expressions of $\nabla J(x)$ for the cases where $D$ assumes the form of a general, diagonal, or scalar matrix.

1) Case 1: $D$ Is a General Matrix: From (25), the optimal choice of $D$ is given by
\[ D = \hat{T}^{-T} A T, \]  
(28)
which leads to
\[ J(\hat{T}^{-T} A T, T) = \text{tr} \left[ T \hat{C} T^T \right]. \]  
(29)

In this case, the number of elements in vector $x$ consisting of $T$ is equal to $n^2$ and the gradient of $J(x)$ is found to be
\[ \frac{\partial J(x)}{\partial t_{ij}} = \lim_{\Delta \to 0} \frac{J(T_{ij}) - J(T)}{\Delta} \]
\[ = 2 e_j^T T \hat{C} T^T T g_{ij}, \quad i, j = 1, 2, \ldots, n \]  
(30)
where $\hat{T}_{ij}$ is the matrix obtained from $\hat{T}$ with a perturbed $(i,j)$th component, which is given by 

$$
\hat{T}_{ij} = T + \frac{\Delta T g_i e_j^T}{1 - \Delta e_j^T T g_{ij}}
$$

and $g_{ij}$ is computed using 

$$
g_{ij} = \partial \left\{ \frac{1}{||t||} \right\} / \partial t_{ij} = \frac{1}{||t||} (t_{ij} - ||t||^2 e_i).
$$

2) Case 2: $D$ Is a Diagonal Matrix: Here, matrix $D$ assumes the form

$$
D = \text{diag}\{d_1, d_2, \cdots, d_n\}. \quad (31)
$$

In this case, (25) becomes

$$
J(D, \hat{T}) = \text{tr} \left[ T M_d \hat{T}^T \right] \quad (32)
$$

where

$$
M_d = \hat{C} + \hat{A}^T \hat{W}_4 \hat{A} + \hat{W}_4 \hat{T}^T D^2 \hat{T}^T - \hat{A}^T \hat{W}_4 \hat{T}^T D \hat{T}^T - \hat{W}_4 \hat{A}^T D \hat{T}^T - \hat{W}_4 \hat{A} \hat{T}^T D \hat{T}^T - \hat{W}_4 \hat{A} \hat{T}^T D \hat{T}^T.
$$

It follows that

$$
\frac{\partial J(x)}{\partial t_{ij}} = 2e_i^T T M_d \hat{T}^T g_{ij}, \quad i, j = 1, 2, \cdots, n
$$

$$
\frac{\partial J(x)}{\partial d_i} = 2e_i^T (D \hat{T} - \hat{A}^T \hat{W}_4 \hat{A}^T) \hat{W}_4 \hat{T}^T e_i, \quad i = 1, 2, \cdots, n. \quad (33)
$$

3) Case 3: $D$ Is a Scalar Matrix: It is assumed here that $D = \alpha I_n$ with a scalar $\alpha$. The gradient of $J(x)$ can then be calculated as

$$
\frac{\partial J(x)}{\partial t_{ij}} = 2e_i^T T M_d \hat{T}^T g_{ij}, \quad i, j = 1, 2, \cdots, n
$$

$$
\frac{\partial J(x)}{\partial \alpha} = \text{tr} \left[ T(2\alpha \hat{W}_4 - \hat{A}^T \hat{W}_4 - \hat{W}_4 \hat{A}) \hat{T}^T \right] \quad (34)
$$

where

$$
M_s = (\hat{A} - \alpha I_n)^T \hat{W}_4 (\hat{A} - \alpha I_n) + \hat{C}.
$$

4. A NUMERICAL EXAMPLE

In this section we illustrate the proposed method by considering a linear discrete-time system specified by 

$$
A_0 = \begin{bmatrix}
0 & 1 & 1 & 0
0 & 339377 & -1.152652 & 1.520167
\end{bmatrix}, \quad b_0 = \begin{bmatrix}
0
0
\end{bmatrix},
$$

$$
c_0 = \begin{bmatrix}
0.093253
0.128620
0.314713
\end{bmatrix}.
$$

Suppose that the poles of the observer and regulator in the system are required to be located at $z = 0.1532$, $0.2861$, $0.1137$, and $z = 0.5067$, $0.6023$, $0.4331$, respectively. This can be achieved by choosing 

$$
k_o = \begin{bmatrix}
0.471552
-0.367158
3.062267
\end{bmatrix},
$$

$$
g_o = \begin{bmatrix}
-0.006436
3.683651
5.083920
\end{bmatrix}^T.
$$

Performing the $l_2$-scaling to the state-estimate feedback controller, we obtain $J(0) = 686.4121$ in (7) where $D = 0$. Next, the controller is transformed into the optimal realization that minimizes $J(0)$ in (7) under the $l_2$-scaling constraints. This leads to $J_{\text{min}}(0) = 28.6187$. Finally, EF and state-variable coordinate transformation are applied to the above optimal realization so as to jointly minimize the output roundoff noise. The profiles of $J(x)$ during the first 20 iteration for the cases of $D$ being a general, diagonal, and scalar matrix are depicted in Fig. 2.

1) Case 1: $D$ Is a General Matrix: The quasi-Newton algorithm was applied to minimize (25). It took the algorithm 20 iterations to converge to the solution 

$$
D_{\text{3bit}} = \begin{bmatrix}
0.250 & -3.125 & -3.375
-1.375 & 1.875 & 3.250
1.875 & -1.875 & -3.750
\end{bmatrix},
$$

and a noise gain $J(D_{\text{3bit}}, \hat{T}) = 23.4873$. Furthermore, when the optimal EF matrix $D$ was rounded to a power-of-two representation with 3 bits after the binary point, which resulted in 

$$
D_{\text{int}} = \begin{bmatrix}
0 & -3 & -3
-1 & 2 & 3
2 & -2 & -4
\end{bmatrix},
$$

the noise gain was found to be $J(D_{\text{int}}, \hat{T}) = 293.0187$.

2) Case 2: $D$ Is a Diagonal Matrix: Again, the quasi-Newton algorithm was applied to minimize $J(D, \hat{T})$ in (25) for a diagonal EF matrix $D$. It took the algorithm 20 iterations to converge to the solution 

$$
D = \text{diag}\{0.050638, -0.608845, -0.951572\}
$$

$$
T = \begin{bmatrix}
3.58878 & 0.735966 & 0.010417
-2.457241 & 0.728171 & 0.556762
1.514232 & -2.058556 & 0.142204
\end{bmatrix}
$$

and the minimized noise gain was found to be $J(D_{\text{int}}, \hat{T}) = 12.7097$. Next, the above optimal diagonal EF matrix $D$ was rounded to a power-of-two representation with 3 bits after the binary point to yield $D_{\text{3bit}} = \text{diag}\{0.000, -0.625, -1.000\}$, which leads to a noise gain $J(D_{\text{3bit}}, \hat{T}) = 12.7722$. Furthermore, when the optimized diagonal EF matrix $D$ was rounded to the integer representation $D_{\text{int}} = \text{diag}\{0, -1, -1\}$, the noise gain was found to be $J(D_{\text{int}}, \hat{T}) = 13.7535$.

3) Case 3: $D$ Is a Scalar Matrix: In this case, the quasi-Newton algorithm was applied to minimize (25)
for $D = \alpha I_3$ with a scalar $\alpha$. The algorithm converges after 20 iterations to converge to the solution

$$D = -0.779678 I_3$$

and the minimized noise gain was found to be $J(D, \hat{T}) = 16.2006$. Next, the EF matrix $D = \alpha I_3$ was rounded to a power-of-two representation with 3 bits after the binary point as well as an integer representation. It was found that these representations were given by $D_{3\text{bit}} = \text{diag}(0.750, 0.750, 0.750)$ and $D_{\text{int}} = \text{diag}(1, 1, 1)$, respectively. The corresponding noise gains were obtained as $J(D_{3\text{bit}}, \hat{T}) = 16.2370$ and $J(D_{\text{int}}, \hat{T}) = 18.2063$, respectively.

The above simulation results in terms of noise gain $J(D, \hat{T})$ in (25) are summarized in Table 1. For comparison purpose, their counterparts obtained using the method in [9] are also included in the table, where the minimization of the roundoff noise was carried out using EF and state-variable coordinate transformation, but in a separate manner. From the table, it is observed that the proposed joint optimization offers improved reduction in roundoff noise gain for the cases of a scalar EF matrix and a diagonal EF matrix when compared with those obtained by using separate optimization. However, in the case of a general EF matrix, the optimal solution with infinite precision appears to be quite sensitive to the parameter perturbations.

More reduction of the noise gain might be possible by re-designing the coordinate transformation matrix $T$ for the optimally quantized $D$.

### 5. CONCLUSION

The joint optimization problem of EF and realization to minimize the effects of roundoff noise of the closed-loop system with a state-estimate feedback controller subject to $l_2$-scaling constraints has been investigated. The problem at hand has been converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm has been employed to solve the unconstrained optimization problem. The proposed technique has been applied to the cases where EF matrix is a general, diagonal, or scalar matrix. The effectiveness for the cases of a scalar EF matrix and a diagonal EF matrix compared with the existing method [9] has been illustrated by a numerical example.

### REFERENCES


