Finite Word Length Implementation for Digital Reduced Order Observer Based Controllers

J. Wu*, S. Chen¹, G. Li² and J. Chu
National Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, P. R. China
¹Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
²School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Implementation issues for digital reduced-order observer-based controllers with Finite Word Length (FWL) considerations is studied. A tractable FWL stability related measure is derived, and the optimal FWL realization problem for digital reduced-order observer-based controller is to find those realizations that maximize this related measure. This optimization problem is formulated as an unconstrained nonlinear programming problem which can be solved using the simplex search algorithm. A numerical example is given to illustrate the design procedure and the effectiveness of the proposed method.

Introduction
The recent advances in fixed-point implementation of digital controllers such as the design of dedicated fixed-point Digital Signal Processors (DSP) and new Digital Control Processors (DCP) architectures have made Finite Word Length (FWL) implementation an important issue in modern digital control engineering design applications. Improved control performance and increased levels of integration are especially important in many areas. This is because hardware controller implementation with fixed-point arithmetic offers the advantages of speed, memory space, cost and simplicity over floating-point arithmetic.

The FWL effects have been studied in digital control systems using different approaches: the effects of FWL implemented digital controller on the degradation of an LQG cost function was studied [1] using a statistical point of view; the effects of FWL on the stability and performance of sampled data systems was analyzed and an FWL stability measure was presented [2], but computing explicitly this measure seems very hard and is still an open problem; based on the first order approximation, a tractable FWL stability related measures was developed [3].

*Author for correspondence (e-mail: jwu@ipc.zju.edu.cn).
In all the above studies of FWL effects of digital controllers, the controllers are output feedback controllers. It is well known that there are another type of controllers, i.e. observer-based controllers. Because that state-space methods and observer theory form a direct multivariable approach to linear control system synthesis and design, the design of observer-based controllers is more apparent and simpler than the design of output feedback controllers. Specifically, reduced-order observers are preferred as they reduce the redundancy of full-order observers and have the simplest construction. Hence this paper intends to study the FWL implementation issues for digital reduced-order observer-based controllers which were not discussed in the previous FWL study work. One contribution of this paper is to compute the FWL stability related measure for any realization of a reduced-order observer-based controller. Another is to develop an algorithm for searching for the optimal reduced-order observer-based controller realization providing the maximal FWL stability related measure.

**Notation and Problem Statement**

Consider the discrete-time plant $P(z)$ represented as

\[ \begin{align*}
  x(k+1) &= A_s x(k) + B_s e(k) \\
  y(k) &= C_s x(k)
\end{align*} \quad \ldots (1) \]

which is assumed to be strictly proper, completely state controllable, completely state observable, with $A_s \in \mathbb{R}^{n \times n}, B_s \in \mathbb{R}^{n \times p}, C_s \in \mathbb{R}^{q \times n}, q < n$ and $\text{rank} C_s = q$. Given the digital $(n-q)$-order observer-based controller $C(z)$ as

\[ \begin{align*}
  v(k+1) &= F v(k) + G y(k) + H e(k) \\
  u(k) &= J v(k) + M y(k)
\end{align*} \quad \ldots (2) \]

where $F \in \mathbb{R}^{(n-q) \times (n-q)}, G \in \mathbb{R}^{(n-q) \times q}, J \in \mathbb{R}^{p \times (n-q)}, M \in \mathbb{R}^{p \times q}$ and $H \in \mathbb{R}^{(n-q) \times p}$. The realizations $(F, G, J, M, H)$ of $C(z)$ are not unique. In fact, through the reduced-order observer-based controller design procedure, a realization $(F_0, G_0, J_0, M_0, H_0)$ has been determined. Any realization of $C(z)$ can be described as $(F = T^{-1} F_0 T, G = T^{-1} G_0, J = J_0 T, M = M_0, H = T^{-1} H_0)$, where $T \in \mathbb{R}^{(n-q) \times (n-q)}$ is any (real-valued) non-singular matrix, called a similarity transformation. Denote $S_C$ be the set including all realizations of $C(z)$. Denote $U(\cdot)$ be the column stacking operator. Denote

\[ w \triangleq \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}, \quad w_0 \triangleq \begin{bmatrix} U(F) \\ U(G) \\ U(J) \\ U(M) \\ U(H) \end{bmatrix} \quad \ldots (3) \]

where $N = n^2 + 2np - nq - pq$. Obviously, we can call $w$ a realization of $C(z)$. Since the input of $P(z)$

\[ e(k) = r(k) - u(k) \quad \ldots (4) \]
Digital Reduced Order Observer Based Controllers

Both $P(z)$ and $C(z)$ form a discrete-time closed-loop system. Denote $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ be the state-space description of the closed-loop system, it can be shown that

$$
\tilde{A}(w) = \begin{bmatrix}
A_s - B_sMC_s & -B_sJ \\
GC_s - HMC_s & F - HJ
\end{bmatrix} = \begin{bmatrix}
I_n & 0 \\
0 & T^{-1}
\end{bmatrix} A(w_0) \begin{bmatrix}
I_n & 0 \\
0 & T
\end{bmatrix}
$$

(5)

where $I_n$ denotes the $n \times n$ identity matrix. Denote $\lambda_i(\cdot)$ as the $i$th eigenvalue of matrix. It follows from the fact that the closed loop system is stable that

$$
|\lambda_i(\tilde{A}(w_0))| < 1, \forall i \in \{1, \ldots, 2n - q\}
$$

(6)

which implies that all different realization $w$ achieve exactly the same closed-loop poles if $C(z)$ is implemented by an infinite precision DCP. In practice, however, $C(z)$ can only be implemented by an DCP with FWL. Due to the FWL effect, $w$ is perturbed into $w + \Delta w$ and each element of $\Delta w$ is bounded by $\epsilon/2$, i.e.

$$
\mu(\Delta w) \overset{\Delta}{=} \max_{i \in \{1, \ldots, N\}} |\Delta w_i| \leq \epsilon/2
$$

(7)

For a fixed point processor of $B_s$ bits,

$$
\epsilon = 2^{-(B_s-B_X)}
$$

(8)

where $2^{B_X}$ is the biggest normalization factor such that each parameter of $2^{-B_X} w$ is absolutely not bigger than 1. With the perturbation $\Delta w$, $\lambda_i(\tilde{A}(w))$ is moved to $\lambda_i(\tilde{A}(w + \Delta w))$ which may be outside the open unit disk. Thus, the closed-loop system designed to be stable may be unstable with an FWL implementation of the controller realization $w$.

Obviously, for a realization $w$, there is the smallest word length $B_s^{\text{min}}(w)$ that ensures stability. Define the FWL stability measure

$$
\mu_0(w) \overset{\Delta}{=} \inf\{\mu(\Delta w) : \tilde{A}(w + \Delta w) \text{ is unstable}\}
$$

(9)

It follows from $\epsilon/2 \leq \mu_0(w)$ that $B_s^{\text{min}}(w)$ is not less than $-\log_2 \mu_0(w) - 1 + B_X$. Hence we can define

$$
\hat{B}_s^{\text{min}}(w) \overset{\Delta}{=} \text{Int}(-\log_2 \mu_0(w)) - 1 + B_X
$$

(10)

as the estimate of $B_s^{\text{min}}(w)$, where $\text{Int}(x)$ rounds $x$ to the nearest integer towards $+\infty$. Noting that $\mu_0(w)$ is a function of the controller realization $w$, the interesting problem is to find out those realizations such that $\mu_0(w)$ is maximized

$$
\max_{w \in S_C} \mu_0(w)
$$

(11)

These realizations need less word length to ensure stability. It should be pointed out that computing explicitly the value for $\mu_0(w)$ and solving problem (11) seem very hard and are still open problems. In order to overcome the difficulty of $\mu_0(w)$, a tractable FWL stability related measure will be discussed as follows.
A Tractable FWL Stability Related Measure

First, when the FWL error $\Delta w$ is small we have $\forall i \in \{1, \ldots, 2n - q\}$,

$$\Delta \lambda_i = \lambda_i(\bar{A}(w + \Delta w)) - \lambda_i(\bar{A}(w)) \approx \sum_{j=1}^{N} \frac{\partial \lambda_i}{\partial w_j} \Delta w_j \quad \ldots (12)$$

It follows that

$$|\Delta \lambda_i| \leq \sum_{j=1}^{N} |\frac{\partial \lambda_i}{\partial w_j}| |\Delta w_j| \leq \mu(\Delta w) \sum_{j=1}^{N} |\frac{\partial \lambda_i}{\partial w_j}|, \forall i \quad \ldots (13)$$

Defining

$$\mu_1(w) \triangleq \min_{i \in \{1, \ldots, 2n-q\}} \frac{1-|\lambda_i(\bar{A}(w))|}{\sum_{j=1}^{N} |\frac{\partial \lambda_i}{\partial w_j}|} \quad \ldots (14)$$

Therefore

$$\mu_1(w) \leq \mu_0(w) \quad \ldots (15)$$

holds if $\mu_0(w)$ is small enough, and the system is stable if $\mu(\Delta w) \leq \mu_1(w)$ when (15) is true. Hence $\mu_1(w)$ can be viewed as a FWL stability related measure.

For computation of $\mu_1(w)$, the following theorem is important.

Theorem 1: Let $A = M_0 + M_1X M_2 \in R^{m \times m}$ be diagonalizable with $X \in R^{l \times r}$ and $M_0$, $M_1$, and $M_2$ independent of $X$ and having a proper dimension. Denote $\{\lambda_i\} = \{\lambda_i(A)\}$ as its eigenvalues, Let $x_i$ be a right eigenvector of $A$ corresponding to the eigenvalue $\lambda_i$. Denote $M_x = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$ and $M_y = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix} = M_x^{-H}$, where $y_i$ is called the reciprocal left eigenvector corresponding to $\lambda_i$. Then

$$\frac{\partial \lambda_i}{\partial X} = \begin{bmatrix} \frac{\partial \lambda_i}{\partial x_{i1}} & \cdots & \frac{\partial \lambda_i}{\partial x_{ir}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \lambda_i}{\partial x_{i1}} & \cdots & \frac{\partial \lambda_i}{\partial x_{ir}} \end{bmatrix} = M_1^T y_i x_i^T M_2^T \quad \ldots (16)$$

where superscript ‘$H$’ denotes the transpose and conjugate operation, ‘$y_i$’ is conjugate to $y_i$.

Proof: Let $a$ be a element of $X$. It follows from $y_i^H x_i = 1$ that

$$\frac{\partial y_i^H}{\partial a} x_i + y_i^H \frac{\partial x_i}{\partial a} = 0 \quad \ldots (17)$$

Noting $Ax_i = \lambda_i x_i$, one steadily has $\lambda_i = y_i^H A x_i$ and hence

$$\frac{\partial \lambda_i}{\partial a} = y_i^H A x_i + y_i^H \frac{\partial A}{\partial a} x_i + y_i^H A \frac{\partial x_i}{\partial a} \quad \ldots (18)$$

It follows from (17) and $y_i^H A = \lambda_i y_i^H$ that

$$\frac{\partial \lambda_i}{\partial a} = (\frac{\partial y_i^H}{\partial a} \lambda_i x_i + \lambda_i y_i^H \frac{\partial x_i}{\partial a}) + y_i^H \frac{\partial A}{\partial a} x_i = y_i^H M_1 \frac{\partial x}{\partial a} M_2 x_i \quad \ldots (19)$$

For the $(k, j)th$ element of $X$, i.e., $\alpha = x_{k,j}$, one has

$$\frac{\partial \lambda_i}{\partial a} = (y_i^H M_1)(k)(M_2 x_i)(j) \quad \ldots (20)$$
Digital Reduced Order Observer Based Controllers

where \((y_i^HM_1)(k)\) and \((M_2x_i)(j)\) are the \(k\)th and \(j\)th element of \(y_i^HM_1\) and \(M_2x_i\), respectively. This leads to (16).

From (5), we know that

\[
\begin{align*}
\bar{A}(w) &= \begin{bmatrix} A_s - B_sMC_s & -B_sJ \\ GC_s - HMC_s & -HJ \end{bmatrix} + \begin{bmatrix} 0 \\ I_{n-q} \end{bmatrix} F \begin{bmatrix} 0 & I_{n-q} \end{bmatrix} \\
\end{align*}
\]

\( \ldots (21) \)

Using Theorem 1, we have

\[
\begin{align*}
\frac{\partial \lambda}{\partial F} &= \begin{bmatrix} 0 & I_{n-q} \end{bmatrix} y_i^T x_i^T \begin{bmatrix} 0 \\ I_{n-q} \end{bmatrix} \\
\frac{\partial \lambda}{\partial G} &= \begin{bmatrix} 0 & I_{n-q} \end{bmatrix} y_i^T x_i^T \begin{bmatrix} C_s^T \\ 0 \end{bmatrix} \\
\frac{\partial \lambda}{\partial J} &= \begin{bmatrix} -B_s^T & -H^T \end{bmatrix} y_i^T x_i^T \begin{bmatrix} 0 \\ I_{n-q} \end{bmatrix} \\
\frac{\partial \lambda}{\partial M} &= \begin{bmatrix} -B_s^T & -H^T \end{bmatrix} y_i^T x_i^T \begin{bmatrix} C_s^T \\ 0 \end{bmatrix} \\
\frac{\partial \lambda}{\partial H} &= \begin{bmatrix} 0 & I_{n-q} \end{bmatrix} y_i^T x_i^T \begin{bmatrix} -C_sM^T \\ -J^T \end{bmatrix} \\
\end{align*}
\]

\( \ldots (26-30) \)

With \(\frac{\partial \lambda}{\partial F}, \frac{\partial \lambda}{\partial G}, \frac{\partial \lambda}{\partial J}, \frac{\partial \lambda}{\partial M}, \text{ and } \frac{\partial \lambda}{\partial H}\), \(\mu_1(w)\) can be computed easily using (14). Based on \(\mu_1(w)\), we can compute

\[
\hat{B}_{s1}^{\min}(w) = \text{Int}(-\log_2 \mu_1(w)) - 1 + B_X
\]

\( \ldots (31) \)

as the estimate of the minimum word length \(\hat{B}_{s1}^{\min}(w)\) that ensures stability of the closed-loop system.

Optimal realization

Let \(x_{i0} = \begin{bmatrix} x_{i0}(1) \\ x_{i0}(2) \end{bmatrix} \in \mathbb{C}^{2n-q}\) be a right eigenvector of \(\bar{A}(w_0)\) corresponding to the eigenvalue \(\lambda_0 = \lambda_i(\bar{A}(w_0)) = \lambda_i(\bar{A}(w))\), \(y_{i0} = \begin{bmatrix} y_{i0}(1) \\ y_{i0}(2) \end{bmatrix} \in \mathbb{C}^{2n-q}\) be the reciprocal left eigenvector corresponding to \(x_{i0}\), where \(x_{i0}(1), y_{i0}(1) \in \mathbb{C}^n, x_{i0}(2), y_{i0}(2) \in \mathbb{C}^{n-q}\). It is easy to see from (5) that

\[
x_i = \begin{bmatrix} x_{i0}(1) \\ T^{-1}x_{i0}(2) \end{bmatrix}
\]

\( \ldots (32) \)
is a right eigenvector and
\[ y_i = \begin{bmatrix} y_i(1) \\ T^Ty_i(2) \end{bmatrix} \quad \ldots (33) \]
is the reciprocal left eigenvector of \( \bar{A}(w) \) corresponding to \( \lambda_{i0}, \forall i \in \{1, \ldots, 2n-q\} \). Applying (26)–(30), we have
\[ \frac{\partial \lambda_{i0}}{\partial T} = T^Ty_i(2)x_i(1)C_s^T \quad \ldots (34) \]
\[ \frac{\partial \lambda_{i0}}{\partial G} = -B_s^Ty_i(1) + H_s^Ty_i(2)x_i(2)T^{-T} \quad \ldots (35) \]
\[ \frac{\partial \lambda_{i0}}{\partial M} = -B_s^Ty_i(1) + H_s^Ty_i(2)x_i(1)C_s^T \quad \ldots (36) \]
\[ \frac{\partial \lambda_{i0}}{\partial H} = -TT^Ty_i(2)(x_i(1)C_s^T + x_i(2)J_s^T) \quad \ldots (37) \]
\[ \frac{\partial \lambda_{i0}}{\partial F} = T^Ty_i(2)x_i(2)T^{-T} \quad \ldots (38) \]

Defining
\[ \|X\|_S = \sum_{i=1}^{r} \sum_{j=1}^{r} |x_{ij}|, \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1r} \\ \vdots & \ddots & \vdots \\ x_{11} & \cdots & x_{rr} \end{bmatrix} \in \mathbb{C}^{r \times r} \quad \ldots (39) \]

From (34)–(38), we define the following function of the similarity matrix \( T \):
\[ f(T) = \max_{i \in \{1, \ldots, 2n-q\}} \| \frac{\partial \lambda_{i0}}{\partial T} \|_S + \| \frac{\partial \lambda_{i0}}{\partial G} \|_S + \| \frac{\partial \lambda_{i0}}{\partial M} \|_S + \| \frac{\partial \lambda_{i0}}{\partial H} \|_S \frac{1}{1-|\lambda_{i0}|} \quad \ldots (40) \]

We can describe the optimal FWL realization problem of reduced-order observer-based controller as the optimization problem:
\[ v = \frac{1}{\max_{w \in C} \mu_1(w)} = \min_{T \in \mathbb{R}^{(n-q) \times (n-q)}} f(T) \quad \ldots (41) \]

The above problem is a nonconvex nonlinear programming problem. Denote \( T_{opt} \) as the solutions to (41). We intend to search for \( T_{opt} \) with an iterative optimization method, in which a sequence \( \{T_0, T_1, T_2, \ldots\} \) which converges to \( T_{opt} \) is generated. In this iterative procedure, we can neglect the constraint \( \det T \neq 0 \), i.e. we solve the problem
\[ \min_{T \in \mathbb{R}^{(n-q) \times (n-q)}} f(T) \quad \ldots (42) \]

There are reasons for us to do so: \( \Omega = \{ T \mid \det T = 0, T \in \mathbb{R}^{(n-q) \times (n-q)} \} \) is only a manifold in space \( \mathbb{R}^{(n-q) \times (n-q)} \). Hence the situation when \( T_i \) moves into \( \Omega \) is rare when we search the space \( \mathbb{R}^{(n-q) \times (n-q)} \) for \( T_{opt} \in \Omega \) by an iterative sequence from a start point \( T_0 \notin \Omega \). Even if it happens that \( T_i \) moves into \( \Omega \) in the iterative procedure, we can add a small perturbation \( \tau I_n \) to \( T_i \) such that \( T_i + \tau I_n \notin \Omega \). This small perturbation would not affect the convergence of the iterative sequence to \( T_{opt} \).

In this paper, the simplex search algorithm is applied to solve problem (42) which is a unconstrained convex nonlinear programming problem. There are
Digital Reduced Order Observer Based Controllers

many existing optimization software which uses the simplex search algorithm, for example, the \textit{fmins.m} function in MATLAB Ver5.1 optimization toolbox.

**Illustrative Example**

In this section, we present a design example to show how the optimization approach presented in this paper can be used efficiently for searching for an optimal transformation and hence the optimal controller realization.

The discrete-time plant is given by

\[
A_s = \begin{bmatrix}
2.7582 & -2.5342 & 0.7756 \\
1.0000 & 0 & 0 \\
0 & 1.0000 & 0
\end{bmatrix},
B_s = \begin{bmatrix}
1 \\
0 \\
0
\end{bmatrix},
C_s = \begin{bmatrix}
0.0022 \\
0.0044 \\
0.0022
\end{bmatrix}^T
\]

The initial realization of the controller \( C(z) \) is given by

\[
F_0 = \begin{bmatrix}
-1.3384 & 1 \\
-4.1776 & 2.7582
\end{bmatrix},
G_0 = \begin{bmatrix}
-0.1479 \\
-0.2828
\end{bmatrix},
J_0 = \begin{bmatrix}
-87.896 & 51.537
\end{bmatrix},
M_0 = 4.3835, H_0 = \begin{bmatrix}
0.0075 \\
0.0219
\end{bmatrix}
\]

The corresponding transition matrix \( \tilde{A}(\omega_0) \) can then be formed using (5), from which the poles and the corresponding eigenvectors of the ideal closed loop system can be computed. The closed-loop poles are:

\[
[\lambda_{10} \lambda_{20} \lambda_{30} \lambda_{40} \lambda_{50}] = [0.9067 \ 0.8437 \ 0.7523 \ 0.5761 \ 0.6231]
\]

Hence problem (42) can be constructed. We use the simplex search algorithm to solve problem (42) which is an optimization problem on \( \mathbb{R}^{2 \times 2} \). Our solution is:

\[
T_{\text{opt}} = \begin{bmatrix}
0.0021 & 0.0031 \\
0.0038 & 0.0061
\end{bmatrix}
\]

and \( \nu = 514.66 \). The optimal realization corresponding to \( T_{\text{opt}} \) is

\[
F_{\text{opt}} = \begin{bmatrix}
0.7414 & -0.0785 \\
-0.2155 & 0.6785
\end{bmatrix},
G_{\text{opt}} = \begin{bmatrix}
-17.689 \\
-35.261
\end{bmatrix},
J_{\text{opt}} = \begin{bmatrix}
0.0068 & 0.0413
\end{bmatrix},
M_{\text{opt}} = 4.3835, H_{\text{opt}} = \begin{bmatrix}
-16.911 \\
14.055
\end{bmatrix}
\]

The results for the initial realization and optimal realization are summarized in Table 1. Obviously, \( \mu_1(w_{\text{opt}}) \) is nearly 50 times of \( \mu_1(w_0) \), and \( w_{\text{opt}} \) develops 6 bits in \( \tilde{B}_{s_{11}}^{\text{min}} \) comparing to \( w_0 \).

<table>
<thead>
<tr>
<th>Realization</th>
<th>( \mu_1(w) )</th>
<th>( \tilde{B}<em>{s</em>{11}}^{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_0 )</td>
<td>( 4.0612 \times 10^{-5} )</td>
<td>21</td>
</tr>
<tr>
<td>( w_{\text{opt}} )</td>
<td>( 1.9430 \times 10^{-3} )</td>
<td>15</td>
</tr>
</tbody>
</table>
Conclusions
In this paper, we have presented an approach to the implementation issues for
digital reduced-order observer-based controller with FWL considerations. A
tractable FWL stability related measure has been derived. Noting that this
related measure is a function of the controller realizations, the optimal realization
problem is to find those realizations that maximize this related measure. It
has been shown that the optimal realization problem can be interpreted as a
nonlinear programming problem. The computation of the relevant optimization
problem was solved using the simplex search algorithm. The theoretical results
were verified using a numerical example which illustrates that the optimum
realization, based on the optimization method presented in this paper, greatly
improves the stability robustness of the closed-loop system with minimum word-
length characteristics compared to non-optimal realizations.

Acknowledgments
J. Wu and J. Chu are supported by Zhejiang Provincial Natural Science Foun-
dation of China under Grant 699085. S. Chen wishes to thank the support of
the UK EPSRC under Grant GR/M16894.

References
compensators: The coefficient wordlength issue. IEEE Transactions on Automatic Control,
25(8), 621-630.
systems subject to word length constraint. IEEE Transactions on Automatic Control,
39(12), 2476-2481.
IEEE Transactions on Automatic Control, 43(5), 689-693.

Received: 28 October 1999; Accepted after revision: 12 May 2000.