Comparative Study on Finite-Precision Controller Realizations in Different Representation Schemes

Jun Wu†, Sheng Chen† and Jian Chu†
† National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou, 310027, China
† School of Electronics and Computer Science
University of Southampton, Southampton S017 1BJ, U.K.
E-mail: sqc@ecs.soton.ac.uk


The authors wish to thank the support of the U.K. Royal Society (KC Wong fellowship E/ART/CN/XH1/HCW/1/1949)

Number Formats

- Fixed-point of bit length $\beta = 1 + \beta y + \beta f$: 1 sign bit, $\beta y$ bits integer part, $\beta f$ bits fractional part. If no overflow

  \[ Q_1(x) = x + \delta_1, \quad |\delta_1| < 2^{-(\beta f + 1)} \]

- Floating point of bit length $\beta = 1 + \beta x + \beta u$: 1 sign bit, $\beta x$ bits exponent, $\beta u$ bits mantissa. If no overflow/underflow

  \[ Q_2(x) = x + x\delta_2, \quad |\delta_2| < 2^{-(\beta u + 1)} \]

- Block floating point of bit length $\beta = 1 + \beta h + \beta u$: 1 sign bit, $\beta h$ bits exponent, $\beta u$ bits block mantissa (in fixed-point). If no overflow/underflow

  \[ Q_3(x) = x + r(x)\delta_3, \quad |\delta_3| < 2^{-(\beta h + 1)} \]

  \[ r(x) = 2\eta_x, \quad x \in S_i \quad \text{and} \quad \eta_x = \max_{y \in \delta_i} \{|y|\} \]

Dynamic range bit length $\beta y, (\beta y, \beta x, \text{or} \beta h)$; Precision bit length $\beta u, (\beta f, \beta u, \text{or} \beta u)$

Motivation

- Finite word length effects
degraded designed closed-loop performance, even cause loss of closed-loop stability

- Unified approach to different representation formats
fixed point, floating point, block floating point

- Dynamic range and precision considerations
closed-loop stability robustness with respect to total bit length

Closed-Loop

Plant

\[ \begin{align*}
x(k + 1) &= Ax(k) + Be(k) \\
y(k) &= Cx(k)
\end{align*} \]

Controller

\[ \begin{align*}
v(k + 1) &= Fv(k) + Gy(k) + He(k) \\
u(k) &= Jv(k) + My(k)
\end{align*} \]

- Controller realizations $(F, G, J, M, H)$ infinite many. Let $(F_0, G_0, J_0, M_0, H_0)$ be a realization designed by some standard procedure, all realizations form set:

  \[ S_C \triangleq \{ (F, G, J, M, H) : F = T^{-1}F_0T, G = T^{-1}G_0, \]

  \[ J = J_0T, M = M_0, H = T^{-1}H_0 \} \]

$T$ being nonsingular. All are equivalent if implemented in infinite precision

- Different realizations have different degrees of robustness against FWL effect
Alternatively, realization presented as $w = [w_1 \cdots w_N]^T \triangleq [w_F \ w_G \ w_J \ w_M \ w_H]^T$
with $w_F = \text{Vec}(F), \cdots, w_H = \text{Vec}(H)$
Dynamic Range Consideration

- Dynamic range measure
  \[ \gamma(w, \alpha) = \begin{cases} \|w\|_{\text{max}}, & \alpha = 1 \text{ (fixed point)} \\ \log_2 \frac{\|w\|_{\text{max}}}{\|w\|_{\text{min}}}, & \alpha = 2 \text{ (floating point)} \\ \log_2 \frac{\|w\|_{\text{min}}}{\|w\|_{\text{max}}}, & \alpha = 3 \text{ (block floating point)} \end{cases} \]

- with \( \|w\|_{\text{max}} = \max_{j \in \{1, \ldots, N\}} |w_j| \), \( \pi(w) = \min_{j \in \{1, \ldots, N\}} \{|w_j| : w_j \neq 0\} \), and \( z(w) = \left[ \begin{array}{c} \eta_F \\ \eta_G \end{array} \right] \).

**Proposition:** Realization \( w \) can be represented in format \( \alpha \) of \( \beta \), dynamic-range bit length with overflow and/or underflow, if \( 2^{\beta_{\alpha}} \geq \gamma(w, \alpha) \).

- Let \( \beta_{\gamma}(w, \alpha) \) be minimum dynamic range bit length that guarantees no overflow and/or underflow. \( \gamma(w, \alpha) \) provides an estimate of \( \beta_{\gamma}(w, \alpha) \):
  \[ \beta_{\gamma}(w, \alpha) \triangleq \left\lceil \log_2 \gamma(w, \alpha) \right\rceil \quad \text{with} \quad \beta_{\gamma}(w, \alpha) \geq \beta_{\gamma}(w, \alpha), \]

where \( \lceil \cdot \rceil \) is ceiling function

Robustness of Closed-Loop Stability

- Assuming sufficient \( \beta \), precision or stability measure:
  \[ \mu(w, \alpha) \triangleq \min_{i \in \{1, \ldots, m+n\}} \frac{1 - |\lambda_i(w)|}{\|\partial w_0 / \partial x_i\|_{\Delta = 0}} \]

- with \( \|\partial w_0 / \partial x_i\|_{\Delta = 0} = r(w, \alpha) \circ \partial w_0 / \partial x_i \).

**Proposition:** Under mild conditions, if \( |\Delta|_{\text{max}} < \mu(w, \alpha) \), then
  \[ |\lambda_i(w + r(w, \alpha) \circ \Delta)| < 1, \quad \forall i \]

- Let \( \beta_{\mu}(w, \alpha) \) be minimum precision bit length that guarantees closed-loop stability. \( \mu(w, \alpha) \) provides an estimate of \( \beta_{\mu}(w, \alpha) \):
  \[ \beta_{\mu}(w, \alpha) \triangleq \left\lfloor \log_2 \mu(w, \alpha) \right\rfloor - 1 \quad \text{with} \quad \beta_{\mu}(w, \alpha) \geq \beta_{\mu}(w, \alpha), \]

where \( \lfloor \cdot \rfloor \) is floor function

Precision Consideration

- By design, closed-loop eigenvalues
  \[ |\lambda_i(w)| < 1, \quad \forall i \]

- But \( w \) cannot be implemented exactly [finite precision]

- Assume sufficient large \( \beta \) [no overflow and/or underflow]. Since \( \beta \) is finite
  \[ w \Rightarrow w + r(w, \alpha) \circ \Delta \]

where \( x \circ y = [x \cdot y] \) is Hadamard product of two same-dimensional vectors \( x \) and \( y \),
  \[ r(w, 1) = [1 \cdots 1]^T, \quad r(w, 2) = w, \quad r(w, 3) = 2[\eta_F \cdots \eta_F \eta_G \cdots \eta_G]^T, \]

- and perturbation vector \( \Delta \) is bounded: \( |\Delta|_{\text{max}} < 2^{-2 \beta} \).

- With \( \Delta \), closed-loop eigenvalues
  \[ \lambda_i(w) \rightarrow \lambda_i(w + r(w, \alpha) \circ \Delta) \]

If \( |\lambda_i(w + r(w, \alpha) \circ \Delta)| \geq 1 \) for some \( i \), closed-loop becomes unstable

Optimal Realization Problem

- Combined FWL measure:
  \[ \rho(w, \alpha) \triangleq \mu(w, \alpha) / \gamma(w, \alpha) \]

Let \( \beta_{\rho}(w, \alpha) \triangleq \beta_{\mu}(w, \alpha) + \beta_{\gamma}(w, \alpha) + 1 \) be minimum required total bit length. \( \rho(w, \alpha) \) provides an estimate of \( \beta_{\rho}(w, \alpha) \):
  \[ \beta_{\rho}(w, \alpha) \triangleq \left\lceil \log_2 \rho(w, \alpha) \right\rceil + 1 \]

- Given \( w_0 \), optimal realization problem:
  \[ \max_{w \in S_C} \rho(w, \alpha) = \max_{T \in \mathbb{R}_{m+n}} \left( \min_{i \in \{1, \ldots, m+n\}} \frac{1 - |\lambda_i(w)|}{\|r(w, \alpha) \circ \partial w_0 / \partial x_i\|_{\gamma(w, \alpha)}} \right) \]

Optimization algorithms based on function values only can be used to solve this problem

With \( T_{\text{opt}}(\alpha) \Rightarrow \text{optimal controller realization } w_{\text{opt}}(\alpha) \).
An Example

Plant
\[
A = \begin{bmatrix}
3.7196e+0 & -5.4143e+0 & 3.0225e+0 & -0.6420e-1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
\end{bmatrix}
\]
\[
B = \begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
\end{bmatrix}^T, \quad C = \begin{bmatrix}
3.1100e+6 & 4.2000e-8 & 1.0880e-6 & 1.4000e-8 \\
\end{bmatrix}
\]

Initial designed controller
\[
F_0 = \begin{bmatrix}
2.6605e+2 & -1.2715e+1 & 2.8573e+1 & 2.0810e+2 \\
2.0506e+2 & -1.0597e+1 & 2.1852e+2 & 4.8000e+2 \\
5.0000e+1 & -8.5715e+0 & 5.2362e+0 & 5.6300e+1 \\
2.3507e+2 & 3.7000e+1 & -2.6020e+2 & -2.3850e+2 \\
\end{bmatrix}
\]

\[
G_0 = \begin{bmatrix}
-6.9700e+1 & -4.5600e+1 & 4.1000e+1 \\
-4.6700e+1 & -4.5600e+1 & 4.1000e+1 \\
\end{bmatrix} , \quad \Delta_0 = \begin{bmatrix}
-2.6048e+2 & -2.7146e+2 & -2.7188e+2 & -2.7188e+2 \\
\end{bmatrix}
\]

\[
M_0 = \begin{bmatrix}
0 & 0 & 0 & 0 \end{bmatrix}^T.
\]

MATLAB routine \texttt{fminsearch.m} used to solve optimization

<table>
<thead>
<tr>
<th>Realization</th>
<th>Representation scheme</th>
<th>measure</th>
<th>(\beta_{\text{min}})</th>
<th>(\beta_{p_{\text{max}}})</th>
<th>(\beta_{r_{\text{max}}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_0)</td>
<td>fixed-point</td>
<td>(1.2312e-10)</td>
<td>31</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>(w_{\text{osd}}) (1)</td>
<td>fixed-point</td>
<td>(1.2000e-6)</td>
<td>19</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>(w_0)</td>
<td>floating-point</td>
<td>(2.9062e-11)</td>
<td>33</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>(w_{\text{osd}}) (2)</td>
<td>floating-point</td>
<td>(9.5931e-6)</td>
<td>13</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>(w_0)</td>
<td>block-floating-point</td>
<td>(1.4347e-11)</td>
<td>33</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>(w_{\text{osd}}) (3)</td>
<td>block-floating-point</td>
<td>(3.5012e-6)</td>
<td>16</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

Comparison of true minimum required bit lengths for \(w_0\) in three representation schemes with those of fixed-point implemented \(w_{\text{osd}}\) (1), floating-point implemented \(w_{\text{osd}}\) (2) and block-floating-point implemented \(w_{\text{osd}}\) (3).

○ Any realization \(w \in SC\) implemented in infinite precision (unlimited \(\beta_{p}\) and infinite \(\beta_{r}\)) will achieve exact performance of infinite-precision implemented \(w_0\), which is designed controller performance.

Infinite-precision implemented \(w_0\) is referred to as ideal controller realization \(w_{\text{ideal}}\). Values of various measures and corresponding estimated bit lengths for four realizations in three different formats
Conclusions

- Unified true closed-loop stability measure for FWL implemented controllers in different representation formats
  Computationally tractable, taking into account both dynamic range and precision of arithmetic schemes
- Formulate and solve optimal controller realization problem
  Design provides useful quantitative information regarding finite precision computational properties, namely robustness to FWL errors and estimated minimum bit length for guaranteeing closed-loop stability
- Designer can choose an optimal controller realization in an appropriate representation scheme to achieve best computational efficiency and closed-loop performance