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 SO17 1BJ, U.K.
E-mail: sqc@ecs.soton.ac.uk

Presented at CACSCUK'2003, Luton, U,K., 20 September, 2002

The authors wish to thank the supports of the U.K. Royal Society (KC Wong fellowship RL/ART/CN/XFI/KCW/11949)

Motivation

◯ Finite word length effects

degrade 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

Number Formats

◯ Fixed-point of bit length $\beta = 1 + \beta_g + \beta_f$: 1 sign bit, β_g bits integer part, β_f bits fractional part. If no overflow

$$\mathcal{Q}_1(x) = x + \delta_1, \quad |\delta_1| < 2^{-(\beta_f+1)}$$

◯ Floating point of bit length $\beta = 1 + \beta_e + \beta_w$: 1 sign bit, β_e bits exponent, β_w bits mantissa. If no overflow/underflow

$$\mathcal{Q}_2(x) = x + x\delta_2, \quad |\delta_2| < 2^{-(\beta_w+1)}$$

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

$$\mathcal{Q}_3(x) = x + r(x)\delta_3, \quad |\delta_3| < 2^{-(\beta_u+1)}$$

$$r(x) = 2\eta_i, \quad \text{if} \quad x \in \mathcal{S}_i \quad \text{and} \quad \eta_i = \max_{y \in \mathcal{S}_i}\{|y|\}$$

Dynamic range bit length β_r (β_g, β_e or β_h); **Precision bit length** β_p (β_f, β_w or β_u)

Closed-Loop

Plant

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases}$$

Controller

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{H}\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases}$$

◯ Controller realizations $(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H})$ infinite many. Let $(\mathbf{F}_0, \mathbf{G}_0, \mathbf{J}_0, \mathbf{M}_0, \mathbf{H}_0)$ be a realization designed by some standard procedure, all realizations form set:

$$\mathcal{S}_C \triangleq \{(\mathbf{F}, \mathbf{G}, \mathbf{J}, \mathbf{M}, \mathbf{H}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0,$$

$$\mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0, \mathbf{H} = \mathbf{T}^{-1}\mathbf{H}_0\}$$

$\mathbf{T}$ being nonsingular. All are equivalent if implemented in infinite precision

◯ Different realizations have different degrees of robustness against FWL effect

Alternatively, realization presented as $\mathbf{w} = [w_1 \cdots w_N]^T \triangleq [\mathbf{w}_F^T \ \mathbf{w}_G^T \ \mathbf{w}_J^T \ \mathbf{w}_M^T \ \mathbf{w}_H^T]^T$ with $\mathbf{w}_F = \mathrm{Vec}(\mathbf{F}), \cdots, \mathbf{w}_H = \mathrm{Vec}(\mathbf{H})$

Dynamic Range Consideration

◯ Dynamic range measure

$$\gamma(\mathbf{w},\alpha) \triangleq \begin{cases} \|\mathbf{w}\|_{\max}, & \alpha = 1 \text{ (fixed point)} \\ \log_2 \frac{4\|\mathbf{w}\|_{\max}}{\pi(\mathbf{w})}, & \alpha = 2 \text{ (floating point)} \\ \log_2 \frac{4\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}, & \alpha = 3 \text{ (block floating point)} \end{cases}$$

with $\|\mathbf{w}\|_{\max} \triangleq \max_{j\in\{1,\cdots,N\}} |w_j|, \quad \pi(\mathbf{w}) \triangleq \min_{j\in\{1,\cdots,N\}} \{|w_j| : w_j \neq 0\},$

$$\mathbf{z}(\mathbf{w}) \triangleq [\eta_F \ \ \eta_G \ \ \eta_J \ \ \eta_M \ \ \eta_H]^T$$

Proposition: Realization **w** can be represented in format α of β_r dynamic-range bit length without overflow and/or underflow, if $2^{\beta_r} \geq \gamma(\mathbf{w},\alpha)$

◯ Let $\beta_r^{min}(\mathbf{w},\alpha)$ be minimum dynamic range bit length that guarantees no overflow and/or underflow. $\gamma(\mathbf{w},\alpha)$ provides an estimate of $\beta_r^{min}(\mathbf{w},\alpha)$:

$$\hat{\beta}_r^{min}(\mathbf{w},\alpha) \triangleq \lceil \log_2 \gamma(\mathbf{w},\alpha) \rceil \quad \text{with} \quad \hat{\beta}_r^{min}(\mathbf{w},\alpha) \geq \beta_r^{min}(\mathbf{w},\alpha)$$

where $\lceil \cdot \rceil$ is ceiling function

Precision Consideration

◯ By design, closed-loop eigenvalues

$$|\lambda_i(\mathbf{w})| < 1, \quad \forall i$$

But $\mathbf{w}$ cannot be implemented exactly (infinite precision)

◯ Assume sufficient large β_r (no overflow and/or underflow). Since β_p is finite

$$\mathbf{w} \Rightarrow \mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \boldsymbol{\Delta}$$

where $\mathbf{x} \circ \mathbf{y} \triangleq [x_j y_j]$ is Hadamard product of two same-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{r}(\mathbf{w},1) = [1\ 1 \cdots 1]^T$, $\mathbf{r}(\mathbf{w},2) = \mathbf{w}$, $\mathbf{r}(\mathbf{w},3) = 2[\eta_F \cdots \eta_F\ \eta_G \cdots \eta_H \cdots \eta_H]^T$, and perturbation vector $\boldsymbol{\Delta}$ is bounded: $\|\boldsymbol{\Delta}\|_{\max} < 2^{-(\beta_p+1)}$

◯ With $\boldsymbol{\Delta}$, closed-loop eigenvalues

$$\lambda_i(\mathbf{w}) \longrightarrow \lambda_i(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \boldsymbol{\Delta})$$

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

Robustness of Closed-Loop Stability

◯ Assuming sufficient β_r, precision or stability measure:

$$\mu(\mathbf{w},\alpha) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1 - |\lambda_i(\mathbf{w})|}{\left\| \frac{\partial|\lambda_i|}{\partial \boldsymbol{\Delta}} \Big|_{\boldsymbol{\Delta}=\mathbf{0}} \right\|_1}$$

where $\left\| \frac{\partial|\lambda_i|}{\partial \boldsymbol{\Delta}} \right\|_1 \triangleq \sum_{j=1}^{N} \left| \frac{\partial|\lambda_i|}{\partial \delta_j} \right|$ and $\frac{\partial|\lambda_i|}{\partial \boldsymbol{\Delta}} \Big|_{\boldsymbol{\Delta}=\mathbf{0}} = \mathbf{r}(\mathbf{w},\alpha) \circ \frac{\partial|\lambda_i|}{\partial \mathbf{w}}$

Proposition: Under mild conditions, if $\|\boldsymbol{\Delta}\|_{\max} < \mu(\mathbf{w},\alpha)$, then

$$|\lambda_i(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \boldsymbol{\Delta})| < 1, \quad \forall i$$

◯ Let $\beta_p^{min}(\mathbf{w},\alpha)$ be minimum precision bit length that guarantees closed-loop stability. $\mu(\mathbf{w},\alpha)$ provides an estimate of $\beta_p^{min}(\mathbf{w},\alpha)$:

$$\hat{\beta}_p^{min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \mu(\mathbf{w},\alpha) \rfloor - 1 \quad \text{with} \quad \hat{\beta}_p^{min}(\mathbf{w},\alpha) \geq \beta_p^{min}(\mathbf{w},\alpha)$$

where $\lfloor \cdot \rfloor$ is floor function

Optimal Realization Problem

◯ Combined FWL measure:

$$\rho(\mathbf{w},\alpha) \triangleq \mu(\mathbf{w},\alpha)/\gamma(\mathbf{w},\alpha)$$

Let $\beta^{min}(\mathbf{w},\alpha) \triangleq \beta_r^{min}(\mathbf{w},\alpha) + \beta_p^{min}(\mathbf{w},\alpha) + 1$ be minimum required total bit length. $\rho(\mathbf{w},\alpha)$ provides an estimate of $\beta^{min}(\mathbf{w},\alpha)$:

$$\hat{\beta}^{min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \rho(\mathbf{w},\alpha) \rfloor + 1$$

◯ Given $\mathbf{w}_0$, optimal realization problem:

$$\max_{\mathbf{w}\in\mathcal{S}_C} \rho(\mathbf{w},\alpha) = \max_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m} \\ \det(\mathbf{T})\neq 0}} \left(\min_{i\in\{1,\cdots,m+n\}} \frac{1 - |\lambda_i(\mathbf{w}_0)|}{\left\| \mathbf{r}(\mathbf{w},\alpha) \circ \frac{\partial|\lambda_i|}{\partial \mathbf{w}} \right\|_1 \gamma(\mathbf{w},\alpha)} \right)$$

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

With $\mathbf{T}_{\text{opt}}(\alpha) \Rightarrow$ optimal controller realization $\mathbf{w}_{\text{opt}}(\alpha)$

An Example

Plant

$$\mathbf{A} = \begin{bmatrix} 3.7156e+0 & -5.4143e+0 & 3.6525e+0 & -9.6420e-1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{B} = [1 \ \ 0 \ \ 0 \ \ 0]^T, \quad \mathbf{C} = [1.1160e-6 \ \ 4.3000e-8 \ \ 1.0880e-6 \ \ 1.4000e-8]$$

Initial designed controller

$$\mathbf{F}_0 = \begin{bmatrix} 2.6963e+2 & -4.2709e+1 & 2.2873e+1 & 2.6184e+2 \\ 2.5561e+2 & -4.0497e+1 & 2.1052e+1 & 2.4806e+2 \\ 5.6096e+1 & -8.5715e+0 & 5.2162e+0 & 5.4920e+1 \\ -2.3907e+2 & 3.7998e+1 & -2.0338e+1 & -2.3203e+2 \end{bmatrix}$$

$$\mathbf{G}_0 = \begin{bmatrix} -4.6765e+1 \\ -4.5625e+1 \\ -9.5195e+0 \\ 4.1609e+1 \end{bmatrix}, \quad \mathbf{J}_0 = [-2.5548e+2 \ \ -2.7185e+2 \ \ -2.7188e+2 \ \ 2.7188e+2],$$

$$\mathbf{M}_0 = [0], \quad \mathbf{H}_0 = [0 \ \ 0 \ \ 0 \ \ 0]^T.$$

MATLAB routine *fminsearch.m* used to solve optimization

		$\mathbf{w}_0$	$\mathbf{w}_{\rm opt}(1)$	$\mathbf{w}_{\rm opt}(2)$	$\mathbf{w}_{\rm opt}(3)$
Fixed point	$\rho(\mathbf{w},1)$	$\mathbf{1.2312e-10}$	$\mathbf{1.2003e-6}$	$1.0580e-7$	$1.1321e-6$
	$\hat{\beta}^{min}(\mathbf{w},1)$	$\mathbf{34}$	$\mathbf{21}$	25	21
	$\mu(\mathbf{w},1)$	$\mathbf{3.3474e-8}$	$\mathbf{2.3082e-4}$	$9.6673e-5$	$2.2287e-4$
	$\hat{\beta}_p^{min}(\mathbf{w},1)$	$\mathbf{24}$	$\mathbf{12}$	13	12
	$\gamma(\mathbf{w},1)$	$\mathbf{2.7188e+2}$	$\mathbf{1.9231e+2}$	$9.1370e+2$	$1.9687e+2$
	$\hat{\beta}_r^{min}(\mathbf{w},1)$	$\mathbf{9}$	$\mathbf{8}$	10	8
Floating point	$\rho(\mathbf{w},2)$	$\mathbf{2.9062e-11}$	$7.6826e-6$	$\mathbf{9.5931e-6}$	$8.5778e-6$
	$\hat{\beta}^{min}(\mathbf{w},2)$	$\mathbf{37}$	18	$\mathbf{18}$	18
	$\mu(\mathbf{w},2)$	$\mathbf{2.2389e-10}$	$9.5628e-5$	$\mathbf{1.5229e-4}$	$1.1822e-4$
	$\hat{\beta}_p^{min}(\mathbf{w},2)$	$\mathbf{32}$	13	$\mathbf{12}$	13
	$\gamma(\mathbf{w},2)$	$\mathbf{7.7038e+0}$	$1.2447e+1$	$\mathbf{1.5875e+1}$	$1.3782e+1$
	$\hat{\beta}_r^{min}(\mathbf{w},2)$	$\mathbf{3}$	4	$\mathbf{4}$	4
Block floating point	$\rho(\mathbf{w},3)$	$\mathbf{1.4347e-11}$	$3.2975e-6$	$3.6938e-7$	$\mathbf{3.5012e-6}$
	$\hat{\beta}^{min}(\mathbf{w},3)$	$\mathbf{38}$	20	23	$\mathbf{20}$
	$\mu(\mathbf{w},3)$	$\mathbf{6.5127e-11}$	$2.7666e-5$	$2.9985e-6$	$\mathbf{3.0083e-5}$
	$\hat{\beta}_p^{min}(\mathbf{w},3)$	$\mathbf{33}$	15	18	$\mathbf{15}$
	$\gamma(\mathbf{w},3)$	$\mathbf{4.5395e+0}$	$8.3902e+0$	$8.1176e+0$	$\mathbf{8.5923e+0}$
	$\hat{\beta}_r^{min}(\mathbf{w},3)$	$\mathbf{3}$	4	4	$\mathbf{4}$

Values of various measures and corresponding estimated bit lengths for four realizations in three different formats

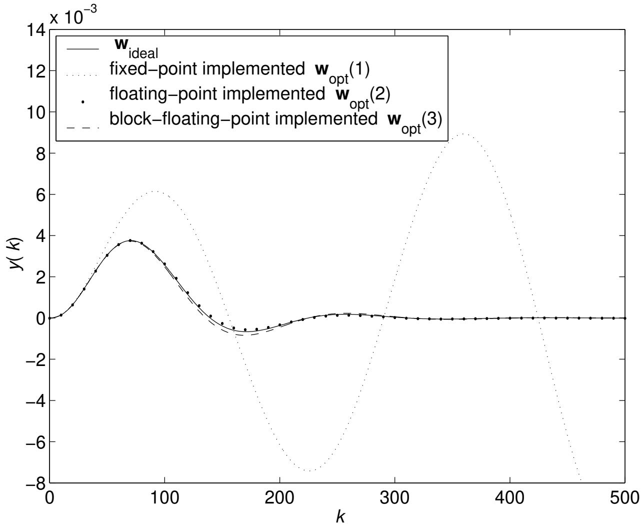
Realization	Representation scheme	measure ρ	β^{min}	β_p^{min}	β_r^{min}
$\mathbf{w}_0$	fixed-point	$\mathbf{1.2312e-10}$	31	21	9
$\mathbf{w}_{\rm opt}(1)$	fixed-point	$\mathbf{1.2003e-6}$	19	10	8
$\mathbf{w}_0$	floating-point	$\mathbf{2.9062e-11}$	33	29	3
$\mathbf{w}_{\rm opt}(2)$	floating-point	$\mathbf{9.5931e-6}$	13	8	4
$\mathbf{w}_0$	block-floating-point	$\mathbf{1.4347e-11}$	33	30	2
$\mathbf{w}_{\rm opt}(3)$	block-floating-point	$\mathbf{3.5012e-6}$	16	12	3

Comparison of true minimum required bit lengths for $\mathbf{w}_0$ in three representation schemes with those of fixed-point implemented $\mathbf{w}_{\rm opt}(1)$, floating-point implemented $\mathbf{w}_{\rm opt}(2)$ and block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$

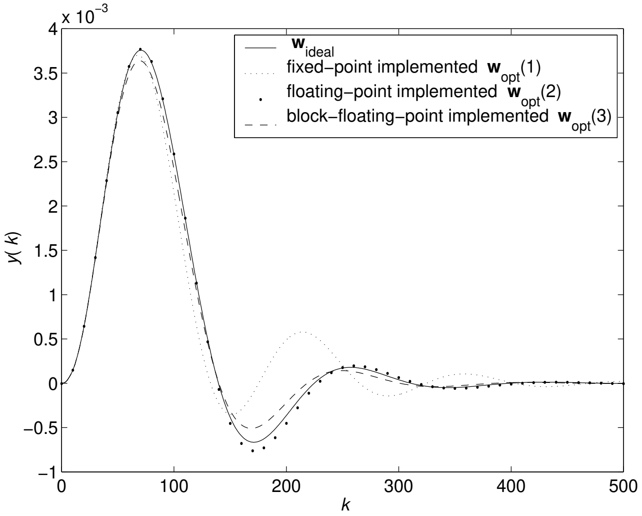
○ Any realization $\mathbf{w} \in \mathcal{S}_C$ implemented in infinite precision (unlimited β_r and infinite β_p) will achieve exact performance of infinite-precision implemented $\mathbf{w}_0$, which is **designed** controller performance

Infinite-precision implemented $\mathbf{w}_0$ is referred to as **ideal** controller realization $\mathbf{w}_{\rm ideal}$

Unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$, and 18-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$, floating-point implemented $\mathbf{w}_{\rm opt}(2)$ and block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$



Unit impulse response of $y(k)$ for $\mathbf{w}_{\text{ideal}}$, 19-bit fixed-point implemented $\mathbf{w}_{\text{opt}}(1)$, floating-point implemented $\mathbf{w}_{\text{opt}}(2)$ and block-floating-point implemented $\mathbf{w}_{\text{opt}}(3)$



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