Optimal Floating-Point Realizations of Finite-Precision Digital Controllers

J. Wu [†], S. Chen [‡], J.F. Whidborne [§] and J. Chu [†]

[†] National Key Laboratory of Industrial Control Technology
Zhejiang University, Hangzhou, 310027, P. R. China
[‡] Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, U.K.
[§] Department of Mechanical Engineering
King's College London, Strand, London WC2R 2LS, U.K.

Presented at 41st IEEE CDC, December 10-13, 2002

Las Vegas, Nevada, USA

Support of U.K. Royal Society is gratefully acknowledged

Motivations and Background

Finite precision controller implementation can seriously influence closed-loop performance.

- Two types of finite word length errors: roundoff errors in arithmetic operations – controller signal errors, and controller coefficient representation errors – controller parameter errors. This work is concerned with the latter, which has critical influence on closed-loop stability.
- Two strategies: direct and indirect. This work adopts an indirect approach.
- Most works deal with fix-point implementation. This work is for floating-point implemented controllers.
- A main contribution of this work is dealing with not only precision but also dynamic range of a numerical representation scheme.

Floating-Point Representation

- Floating-point processor of bit length $\beta = 1 + \beta_w + \beta_e$ represents $x \in \mathcal{R}$: one bit for sign, β_w bits for mantissa, and β_e bits for exponent of x.
- Given β_e bits, the lower and upper limits of exponents are $\underline{e}$ and $\overline{e}$, with $\overline{e} - \underline{e} = 2^{\beta_e} - 1$. Denote the set of integers $\underline{e} \le e \le \overline{e}$ as $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$.
- If the exponent of x, $e = \lfloor \log_2 |x| \rfloor + 1$, is within $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$, there is no underflow or overflow. In such a case, x is perturbed to

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

  The perturbation is multiplicative, unlike the additive perturbation resulting from fixed-point arithmetic.
- β_e determines the dynamic range, and β_w the precision of representation.

Problem Definition

- Plant: $P(z) \sim (\mathbf{A}_P, \mathbf{B}_P, \mathbf{C}_P)$; $\mathbf{A}_P \in \mathcal{R}^{m\times m}$, $\mathbf{B}_P \in \mathcal{R}^{m\times l}$, $\mathbf{C}_P \in \mathcal{R}^{q\times m}$.
- Controller: $C(z) \sim (\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C)$; $\mathbf{A}_C \in \mathcal{R}^{n\times n}$, $\mathbf{B}_C \in \mathcal{R}^{n\times q}$, $\mathbf{C}_C \in \mathcal{R}^{l\times n}$, $\mathbf{D}_C \in \mathcal{R}^{l\times q}$.

  Denote an initially designed controller realization as $\mathbf{X}_0$ and a generic realization $\mathbf{X}$. Let $\overline{\mathbf{A}}(\mathbf{X})$ be the closed-loop transition matrix with $\mathbf{X}$.
- Controller realization set

$$\mathcal{S}_C \triangleq \{\mathbf{X} : \mathbf{A}_C = \mathbf{T}^{-1}\mathbf{A}_C^0\mathbf{T}, \mathbf{B}_C = \mathbf{T}^{-1}\mathbf{B}_C^0, \mathbf{C}_C = \mathbf{C}_C^0\mathbf{T}, \mathbf{D}_C = \mathbf{D}_C^0\}$$

  where $\mathbf{T} \in \mathcal{R}^{n\times n}$ is an arbitrary non-singular matrix
- All $\mathbf{X} \in \mathcal{S}_C$ are equivalent in infinite precision implementation: an identical set of closed-loop eigenvalues $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le i \le m+n$, which are all within the unit disk.

Dynamic Range Measure

- An dynamic range (exponent) measure for floating-point realization $\mathbf{X}$:

$$\gamma(\mathbf{X}) \triangleq \log_2\left(\frac{4\|\mathbf{X}\|_{\max}}{g(\mathbf{X})}\right)$$

where $\|\mathbf{X}\|_{\max} \triangleq \max_{j,k} |x_{j,k}|$ and $g(\mathbf{X}) \triangleq \min_{j,k}\{|x_{j,k}| : x_{j,k} \neq 0\}$.

- $\mathbf{X}$ can be represented in floating-point format of β_e exponent bits without underflow or overflow, if $2^{\beta_e} \geq \gamma(\mathbf{X})$.

- Let β_e^{min} be the smallest exponent bit length for $\mathbf{X}$ without underflow and overflow. Then, $\beta_e^{min} = -\lfloor -\log_2(\lfloor \log_2 \|\mathbf{X}\|_{\max}\rfloor - \lfloor \log_2 g(\mathbf{X})\rfloor + 1)\rfloor$.

- $\gamma(\mathbf{X})$ provides an estimate of β_e^{min} as:

$$\hat{\beta}_e^{min} \triangleq -\lfloor -\log_2 \gamma(\mathbf{X})\rfloor$$

Finite Precision Stability Consideration

- Even without underflow or overflow, due to finite β_w, $\mathbf{X} \Rightarrow \mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}$, with perturbation matrix $\mathbf{\Delta}$ satisfying $\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_w+1)}$.

- With $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{X})) \Rightarrow \lambda_i(\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}))$: Will any of which become outside the unit disk? Or how robust closed-loop stability is to $\mathbf{\Delta}$?

- It is critical to know how large $\mathbf{\Delta}$ will cause closed-loop instability for realization $\mathbf{X}$. Or we would like to know the largest open hypercube in perturbation space, within which closed-loop system remains stable.

- The size of this open hypercube is defined by

$$\mu_0(\mathbf{X}) \triangleq \inf\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}) \text{ is unstable}\}$$

However, we do not know how to calculate $\mu_0(\mathbf{X})$ given $\mathbf{X}$.

Tractable Precision Measure

- A tractable precision (mantissa) measure is:

$$\mu_1(\mathbf{X}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1 - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))|}{\left\|\left.\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}}$$

where $\left\|\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}}\right\|_{\text{sum}} \triangleq \sum_{j,k}\left|\frac{\partial|\lambda_i|}{\partial \delta_{j,k}}\right|$.

- Under some mild conditions, $|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| < 1$ if $\|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})$.

- Let β_w^{min} be the smallest mantissa bit length that guarantees closed-loop stability for floating-point implemented $\mathbf{X}$.

- $\mu_1(\mathbf{X})$ provides an estimate of β_w^{min} as: $\hat{\beta}_{w1}^{min} \triangleq -\lfloor \log_2 \mu_1(\mathbf{X})\rfloor - 1$.

FWL Closed-Loop Stability Measure

- Goodness of $\mathbf{X}$ can be measured by a large value of $\mu_1(\mathbf{X})$ and a small value of $\gamma(\mathbf{X})$ $\Rightarrow$ FWL closed-loop stability measure:

$$\rho_1(\mathbf{X}) \triangleq \mu_1(\mathbf{X})/\gamma(\mathbf{X})$$

- Define the minimum total bit length required in floating point implementation: $\beta^{min} = \beta_e^{min} + \beta_w^{min} + 1$. $\rho_1(\mathbf{X})$ provides an estimate of β^{min} as:

$$\hat{\beta}_1^{min} \triangleq -\lfloor \log_2 \rho_1(\mathbf{X})\rfloor + 1$$

- $\rho_1(\mathbf{X})$ takes into account both the dynamic range and precision considerations.

- Given a controller realization $\mathbf{X}$, the value of $\rho_1(\mathbf{X})$ can be computed.

Optimal Realization Problem

- An optimal controller realization problem is defined as

$$\upsilon \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} \rho_1(\mathbf{X})$$

- With respect to a given initial realization $\mathbf{X}_0$, $\mathbf{X} = \mathbf{X}(\mathbf{T})$. By defining

$$f(\mathbf{T}) \triangleq \rho_1(\mathbf{X}(\mathbf{T}))$$

  the optimal realization is posed as the optimization problem:

$$\upsilon = \max_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n}\\ \det\mathbf{T}\neq 0}} f(\mathbf{T})$$

- With an optimal transformation matrix $\mathbf{T}_{\text{opt}}$, the optimal realization $\mathbf{X}_{\text{opt}}$ can readily be computed.

Design Experiments

- MATLAB optimization routine *fminsearch.m* is used to solve the optimization problem numerically.

  The resulting optimal controller realization is denoted as $\mathbf{X}_{\text{opt}}$.

- Compare with an existing work (Whidborne and Gu 2002, IFAC World Congress), which minimizes a weighted closed-loop eigenvalue sensitivity index.

  This is the only existing work we can find that deals with FWL closed-loop stability of floating-point implemented controller.

  Note that this is effectively a precision measure only.

  The resulting "optimal" controller realization is denoted as $\mathbf{X}_{\text{s}}$.

Example One

Example from (Gevers and Li 1993): $m = 4$, $n = 4$ and $l = q = 1$ with an initially design controller $\mathbf{X}_0$.

Realization	$\mathbf{X}_0$	$\mathbf{X}_{\text{s}}$	$\mathbf{X}_{\text{opt}}$
ρ_1	2.6644e-9	4.7588e-6	9.5931e-6
$\hat{\beta}_1^{min}$	30	19	18
μ_1	8.5182e-8	8.7907e-5	1.5229e-4
$\hat{\beta}_{w1}^{min}$	23	13	12
γ	3.1971e+1	1.8473e+1	1.5875e+1
$\hat{\beta}_e^{min}$	5	5	4
β^{min}	26	15	13
β_w^{min}	20	9	8
β_e^{min}	5	5	4

Example Two

Example from (Whidborne *et al.* 2001, IEEE Trans. AC, Vol.46): $m = 2$, $n = 3$ and $l = q = 1$ with an initially design controller $\mathbf{X}_0$.

Realization	$\mathbf{X}_0$	$\mathbf{X}_{\text{s}}$	$\mathbf{X}_{\text{opt}}$
ρ_1	2.6767e-11	3.1047e-6	5.8446e-6
$\hat{\beta}_1^{min}$	37	20	19
μ_1	2.8122e-10	7.6679e-5	8.2771e-5
$\hat{\beta}_{w1}^{min}$	31	13	13
γ	1.0506e+1	2.4697e+1	1.4162e+1
$\hat{\beta}_e^{min}$	4	5	4
β^{min}	30	15	12
β_w^{min}	25	9	7
β_e^{min}	4	5	4

Conclusions

- A new computationally tractable FWL closed-loop stability measure has been derived for floating-point controller realizations, which takes into account both the exponent and mantissa of floating-point representation.

- This new measure yields a more accurate estimate for the FWL robustness of closed-loop stability for given controller realization.

- Based on this FWL closed-loop stability measure, the optimal controller realization problem has been formulated, which can be solved for using standard numerical optimization algorithms.