Optimal Realizations of Floating-Point Implemented Digital Controllers with Finite Word Length Considerations

Jun Wu [†], Sheng Chen [‡ 0], James F. Whidborne [§] and Jian Chu [†]

† National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control
Zhejiang University, Hangzhou, 310027, P. R. China

‡ Department of Electronics and Computer Science
University of Southampton, Highfield
Southampton SO17 1BJ, U.K.

§ Department of Mechanical Engineering
King's College London
Strand, London WC2R 2LS, U.K.

Abstract

The closed-loop stability issue of finite-precision realizations is investigated for digital controllers implemented in floating-point arithmetic. Unlike the existing methods which only address the effect of the mantissa bits in floating-point implementation to the sensitivity of closed-loop stability, the sensitivity of closed-loop stability is analyzed with respect to both the mantissa and exponent bits of floating-point implementation. A computationally tractable FWL closed-loop stability measure is then defined, and the method of computing the value of this measure is given. The optimal controller realization problem is posed as searching for a floating-point realization that maximizes the proposed FWL closed-loop stability measure, and a numerical optimization technique is adopted to solve for the resulting optimization problem. Simulation results show that the proposed design procedure yields computationally efficient controller realizations with enhanced FWL closed-loop stability performance.

Index Terms — digital controller, finite word length, floating-point, closed-loop stability, optimization.

1 Introduction

The classical digital controller design methodology often assumes that the controller is implemented exactly, even though in reality a control law can only be realized in finite precision. It may seem that the uncertainty resulting from finite-precision computing of the digital controller is so small, compared to

[0]Contact author. Tel/Fax: +44 (0)23 8059 6660/4508; Email: sqc@ecs.soton.ac.uk

the uncertainty within the plant, such that this controller "uncertainty" can simply be ignored. Increasingly, however, researchers have realized that this is not necessarily the case. Due to the FWL effect, a casual controller implementation may degrade the designed closed-loop performance or even destabilize the designed stable closed-loop system, if the controller implementation structure is not carefully chosen. The effects of finite-precision computation have become more critical with the growing popularity of robust controller design methods which focus only on dealing with large plant uncertainty (Keel & Bhattacharryya, 1997; Mäkilä, 1999). It is well known that a control law can be implemented with different realizations, and these different realizations are all equivalent if they are implemented in infinite precision. However, different controller realizations possess different degrees of "robustness" to FWL errors. This property can be utilized to select "optimal" realizations that optimize some given criteria, and several works (Williamson, 1991; Gevers & Li, 1993; Istepanian & Whidborne, 2001) have studied many aspects of finite-precision digital controller design.

Generally speaking, there are two types of FWL errors in the digital controller. The first one is perturbation of controller parameters implemented with FWL and the second one is the rounding errors that occur in arithmetic operations of signals (Li, Wu, Chen & Zhao, 2000). Typically, effects of these two types of errors are investigated separately for the reason of mathematical tractability. The first type of FWL errors directly concerns with the critical issue of closed-loop stability, and many studies have investigated some closed-loop stability robustness measures, especially for fixed-point implementation (Fialho & Georgiou, 1994, 1999; Madievski, Anderson & Gevers, 1995; Li, 1998; Chen, Wu, Istepanian & Chu, 1999; Whidborne, Wu & Istepanian, 2000; Whidborne, Istepanian & Wu, 2001; Wu, Chen, Li, Istepanian & Chu, 2001). The second type of FWL errors is usually measured with the so-called roundoff noise gain (Moroney, Willsky & Houpt, 1980; Williamson & Kadiman, 1989; Li & Gevers, 1990; Liu, Skelton & Grigoriadis, 1992; Li *et al.*, 2000), assuming an FWL implementation of controller coefficients. This paper addresses the closed-loop stability with respect to FWL implementation and in the remaining of the paper, without further pointing out, the FWL effect is taken to mean the first type of FWL errors.

In real-time applications where computational efficiency is critical, a digital controller implemented with fixed-point arithmetic has some advantages over floating-point implementation. However, the detrimental FWL effects are markedly increased in fixed-point implementation due to a reduced precision. It is therefore not surprising that previous works have focused on finding optimal controller realizations using fixed-point arithmetic by maximizing some closed-loop stability measures (Li, 1998; Fialho & Georgiou, 1999; Chen *et al.*, 1999; Chen, Istepanian, Wu & Chu, 2000; Whidborne *et al.*, 2000, 2001;

Wu, Chen, Li, Istepanian & Chu, 2000; Wu *et al.*, 2001). In all the previous works using fixed-point arithmetic, various FWL closed-loop stability measures, which can be shown to directly link to the bits required to implement the fractional part of fixed-point representation (Li, 1998; Chen *et al.*, 1999), are maximized to produce optimal realizations. However, the dynamic range of fixed-point representation is determined by its integer part. Overflow occurs when there are not enough bits for the integer part. Maximizing these measures, while minimizing the bits required for the fractional part, may actually increase the bits required for the integer part (Chen *et al.*, 2000; Wu *et al.*, 2000). Arguably, a better approach would be to consider some measure which has a direct link to the total bit length required.

With decreasing in price and increasing in availability, the use of floating-point processors in controller implementations has increased dramatically. Floating-point representation has quite different characteristics from fixed-point representation. The dynamic range of floating-point representation is determined by its exponent part. Overflow or underflow occurs when the bits for the exponent part are not sufficient. The effects of finite-precision floating-point implementation have been well studied in digital filter designs (Rao, 1996; Kalliojärvi & Astola, 1996; Ralev & Bauer, 1999). However, there has been relatively little work studying explicitly floating point digital controller implementations. Some exceptions include Rink & Chong (1979), Molchanov & Bauer (1995), Istepanian, Whidborne & Bauer (2000), Whidborne & Gu (2001). In the work by Istepanian *et al.* (2000), a block floating-point arithmetic was used, in which control coefficients were forced to have a common exponent and the problem was converted into a fixed-point one. The work by Whidborne & Gu (2001) represents a case of true floating-point implementation. In this work, a weighted closed-loop eigenvalue sensitivity index was defined for floating-point digital controller realizations. This index, however, only considers the mantissa part of floating-point arithmetic, under an assumption that the exponent bits are unlimited.

The generic contribution of this paper is to derive a new FWL closed-loop stability measure that explicitly considers both the mantissa and exponent parts of floating-point arithmetic. The remainder of this paper is organized as follows. Section 2 briefly summarizes the floating-point representation and highlights the multiplicative nature of perturbations resulting from FWL floating-point arithmetic. Section 3 analyses the FWL effect of floating-point arithmetic on closed-loop stability and addresses how to measure such an effect on floating-point implemented digital controllers. Section 4 defines a computationally tractable FWL closed-loop stability measure for floating-point controller realizations and provides the method of computing its value. In section 5, the optimal floating-point controller realization problem is formulated, and a numerical optimization technique is adopted to solve for the resulting optimization problem. Two examples are given in section 6 to demonstrate the effectiveness of the proposed design method, and the paper concludes at section 7.

2 Floating-Point Representation

It is well known that any real number $x \in \mathcal{R}$ can be represented uniquely by

$$x = (-1)^s \times w \times 2^e, \tag{1}$$

where $s \in \{0, 1\}$ is for the sign of x, $w \in [0.5,\ 1)$ is the mantissa of x, $e \in \mathcal{Z}$ is the exponent of x, and $\mathcal{Z}$ denotes the set of integers. When x is stored in a digital computer of finite β bits in a floating-point format, the bits consist of three parts: one bit for s, β_w bits for w and β_e bits for e. Obviously,

$$\beta = 1 + \beta_w + \beta_e. \tag{2}$$

Since β_w and β_e are finite, the set of numbers that is represented by a particular floating-point scheme is not dense on the real line. Thus the set of possible floating-point numbers, $\mathcal{F}$, is given by

$$\mathcal{F} \triangleq \left\{ (-1)^s \left(0.5 + \sum_{i=1}^{\beta_w} b_i 2^{-(i+1)} \right) \times 2^e : s \in \{0,1\}, b_i \in \{0,1\}, e \in \mathcal{Z}, \underline{e} \le e \le \overline{e} \right\} \cup \{0\}, \tag{3}$$

where $\underline{e}$ and $\overline{e}$ represent the lower and upper limits of the exponent, respectively, and $\overline{e} - \underline{e} = 2^{\beta_e} - 1$. Note that unlike fixed-point representation, underflow can occur in floating-point arithmetic.

Denote the set of integers $\underline{e} \le e \le \overline{e}$ as $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$. When no underflow or overflow occurs, that is, the exponent of x is within $\mathcal{Z}_{[\underline{e},\ \overline{e}]}$, the floating-point quantization operator $\mathcal{Q} : \mathcal{R} \to \mathcal{F}$ can be defined as

$$\mathcal{Q}(x) \triangleq \begin{cases} \mathrm{sgn}(x) 2^{(e-\beta_w-1)} \lfloor 2^{(\beta_w-e+1)} |x| + 0.5 \rfloor, & \text{for } x \neq 0 \\ 0, & \text{for } x = 0 \end{cases} \tag{4}$$

where the exponent $e = \lfloor \log_2 |x| \rfloor + 1$ and the floor function $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. The quantization error, ε, is defined as

$$\varepsilon \triangleq |x - \mathcal{Q}(x)|. \tag{5}$$

It can easily be shown that the quantization error is bounded by

$$\varepsilon < |x| 2^{-(\beta_w+1)}. \tag{6}$$

Thus, when x is implemented in floating-point format of β_w mantissa bits, assuming no underflow or overflow, it is perturbed to

$$\mathcal{Q}(x) = x(1+\delta), \quad |\delta| < 2^{-(\beta_w+1)}. \tag{7}$$

It can be seen that the perturbation resulting from finite-precision floating-point arithmetic is multiplicative, unlike the perturbation resulting from finite-precision fixed-point arithmetic, which is additive.

3 Problem Statement

Consider the discrete-time closed-loop control system, consisting of a linear time invariant plant $P(z)$ and a digital controller $C(z)$. The plant model $P(z)$ is assumed to be strictly proper with a state-space description $(\mathbf{A}_P, \mathbf{B}_P, \mathbf{C}_P)$, where $\mathbf{A}_P \in \mathcal{R}^{m\times m}$, $\mathbf{B}_P \in \mathcal{R}^{m\times l}$ and $\mathbf{C}_P \in \mathcal{R}^{q\times m}$. Let $(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C)$ be a state-space description of the controller $C(z)$, with $\mathbf{A}_C \in \mathcal{R}^{n\times n}$, $\mathbf{B}_C \in \mathcal{R}^{n\times q}$, $\mathbf{C}_C \in \mathcal{R}^{l\times n}$ and $\mathbf{D}_C \in \mathcal{R}^{l\times q}$. A linear system with a given transfer function matrix has an infinite number of state-space descriptions. In fact, if $(\mathbf{A}_C^0, \mathbf{B}_C^0, \mathbf{C}_C^0, \mathbf{D}_C^0)$ is a state-space description of $C(z)$, all the state-space descriptions of $C(z)$ form a *realization* set

$$\mathcal{S}_C \triangleq \left\{(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C) | \mathbf{A}_C = \mathbf{T}^{-1}\mathbf{A}_C^0\mathbf{T}, \mathbf{B}_C = \mathbf{T}^{-1}\mathbf{B}_C^0, \mathbf{C}_C = \mathbf{C}_C^0\mathbf{T}, \mathbf{D}_C = \mathbf{D}_C^0\right\} \tag{8}$$

where the transformation matrix $\mathbf{T} \in \mathcal{R}^{n\times n}$ is an arbitrary non-singular matrix. Denote

$$\mathbf{X} = [x_{j,k}] \triangleq \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix}. \tag{9}$$

The stability of the closed-loop control system depends on the eigenvalues of the closed-loop transition matrix

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{A}_P + \mathbf{B}_P\mathbf{D}_C\mathbf{C}_P & \mathbf{B}_P\mathbf{C}_C \\ \mathbf{B}_C\mathbf{C}_P & \mathbf{A}_C \end{bmatrix} &= \begin{bmatrix} \mathbf{A}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \\ &\triangleq \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2 \end{aligned} \tag{10}$$

where $\mathbf{0}$ denotes the zero matrix of appropriate dimension and $\mathbf{I}_n$ the $n \times n$ identity matrix. All the different realizations $\mathbf{X}$ in $\mathcal{S}_C$ have exactly the same set of closed-loop poles if they are implemented with infinite precision. Since the closed-loop system has been designed to be stable, all the eigenvalues $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le i \le m+n$, are within the unit disk. Define

$$\|\mathbf{X}\|_{\max} \triangleq \max_{j,k} |x_{j,k}| \tag{11}$$

and

$$g(\mathbf{X}) \triangleq \min_{j,k}\{|x_{j,k}| : x_{j,k} \neq 0\}. \tag{12}$$

The controller $\mathbf{X}$ is implemented with a floating-point processor of β_e exponent bits, β_w mantissa bits and one sign bit.

Firstly, in order to avoid underflow and/or overflow, both the exponent of $\|\mathbf{X}\|_{\max}$ and the exponent of $g(\mathbf{X})$ should be within $\mathcal{Z}_{[\underline{e},\, \overline{e}]}$ supported by the β_e exponent bits. We define an exponent measure for the floating-point controller realization $\mathbf{X}$ as

$$\gamma(\mathbf{X}) \triangleq \log_2\left(\frac{4\|\mathbf{X}\|_{\max}}{g(\mathbf{X})}\right). \tag{13}$$

The rationale of this exponent measure becomes clear in the following (obvious) proposition.

Proposition 1 $\mathbf{X}$ can be represented in the floating-point format of β_e exponent bits without underflow or overflow, if $2^{\beta_e} \geq \log_2\left(\frac{\|\mathbf{X}\|_{\max}}{g(\mathbf{X})}\right) + 2$.

Let β_e^{min} be the smallest exponent bit length that, when used to implement $\mathbf{X}$, can avoid underflow and overflow. It can be computed as

$$\beta_e^{min} = -\lfloor -\log_2(\lfloor \log_2 \|\mathbf{X}\|_{\max} \rfloor - \lfloor \log_2 g(\mathbf{X}) \rfloor + 1) \rfloor . \tag{14}$$

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

$$\hat{\beta}_e^{min} \stackrel{\triangle}{=} -\lfloor -\log_2 \gamma(\mathbf{X}) \rfloor . \tag{15}$$

It is clear that $\hat{\beta}_e^{min} \geq \beta_e^{min}$.

Secondly, when there is no underflow or overflow, according to the results of section 2, $\mathbf{X}$ is perturbed to $\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}$ due to the effect of finite β_w where

$$\mathbf{X} \circ \mathbf{\Delta} \stackrel{\triangle}{=} [x_{j,k}\delta_{j,k}] \tag{16}$$

represents the Hadamard product of $\mathbf{X}$ and $\mathbf{\Delta} = [\delta_{j,k}]$. Each element of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_w+1)}$, that is,

$$\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_w+1)} . \tag{17}$$

With the perturbation $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision floating-point implemented $\mathbf{X}$.

It is therefore critical to know when the FWL error will cause closed-loop instability. This means that we would like to know the largest open "cube" in the perturbation space, within which the closed-loop system remains stable. Based on this consideration, a mantissa measure for the floating-point realization $\mathbf{X}$ can be defined as

$$\mu_0(\mathbf{X}) \stackrel{\triangle}{=} \inf\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}) \text{ is unstable}\} . \tag{18}$$

From the above definition, the following proposition is obvious.

Proposition 2 $\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta})$ is stable if $\|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{X})$.

Let β_w^{min} be the mantissa bit length such that $\forall \beta_w \geq \beta_w^{min}$, $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is stable for the floating-point implemented $\mathbf{X}$ with β_w mantissa bits and $\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta})$ is unstable for the floating-point implemented $\mathbf{X}$ with $\beta_w^{min}-1$ mantissa bits. Except through simulation, β_w^{min} is generally unknown. It should be pointed out that due to the complex nonlinear relationship between β_w and closed-loop stability, there may exist some odd cases of smaller mantissa bit length $\beta_w < \beta_w^{min}-1$ which regain closed-loop stability. For example, consider the following hypothetical system. When $\beta_w \geq 9$, the closed-loop system is stable, but the closed-loop system becomes unstable with $\beta_w = 8$. However, with $\beta_w = 7$, the closed-loop regains stability. The system becomes unstable again for $\beta_w \leq 6$. For this system, β_w^{min} is 9 rather than 7. The mantissa measure $\mu_0(\mathbf{X})$ provides an estimate of β_w^{min} as

$$\hat{\beta}_{w0}^{min} \triangleq -\lfloor \log_2 \mu_0(\mathbf{X}) \rfloor - 1\,. \tag{19}$$

It can be seen that $\hat{\beta}_{w0}^{min} \geq \beta_w^{min}$.

Define the minimum total bit length required in floating point implementation as

$$\beta^{min} \triangleq \beta_e^{min} + \beta_w^{min} + 1\,. \tag{20}$$

Clearly, a floating-point implemented $\mathbf{X}$ with a bit length $\beta \geq \beta^{min}$ can guarantee no underflow, no overflow and closed-loop stability. Combining the measures $\gamma(\mathbf{X})$ and $\mu_0(\mathbf{X})$ results in the following true FWL closed-loop stability measure for the floating-point realization $\mathbf{X}$

$$\rho_0(\mathbf{X}) \triangleq \mu_0(\mathbf{X})/\gamma(\mathbf{X})\,. \tag{21}$$

An estimate of β^{min} is given by $\rho_0(\mathbf{X})$ as

$$\hat{\beta}_0^{min} \triangleq -\lfloor \log_2 \rho_0(\mathbf{X}) \rfloor + 1\,. \tag{22}$$

It is clear that $\hat{\beta}_0^{min} \geq \beta^{min}$. The following proposition summarizes the usefulness of $\rho_0(\mathbf{X})$ as a measure for the FWL characteristics of $\mathbf{X}$.

Proposition 3 A floating-point implemented $\mathbf{X}$ with a bit length β can guarantee no underflow, no overflow and closed-loop stability, if

$$2^{\beta-1} \geq \rho_0(\mathbf{X})\,. \tag{23}$$

Since the closed-loop stability measure $\rho_0(\mathbf{X})$ is a function of the controller realization $\mathbf{X}$ and $\hat{\beta}_0^{min}$ decreases with the increase of $\rho_0(\mathbf{X})$, an optimal realization can in theory be found by maximizing $\rho_0(\mathbf{X})$, leading to the following optimal controller realization problem

$$\upsilon_{\text{true}} \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} \rho_0(\mathbf{X})\,. \tag{24}$$

However, the difficulty with this approach is that computing the value of $\mu_0(\mathbf{X})$ is an unsolved open problem. Thus, the true FWL closed-loop stability measure $\rho_0(\mathbf{X})$ and the optimal realization problem (24) have limited practical significance. In the next section, we will seek an alternative measure that not only can quantify FWL characteristics of $\mathbf{X}$ but also is computationally tractable.

4 A Tractable FWL Closed-Loop Stability Measure

When the FWL error $\mathbf{\Delta}$ is small, from a first-order approximation, $\forall i \in \{1, \cdots, m+n\}$

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \approx \sum_{j=1}^{l+n}\sum_{k=1}^{q+n} \left.\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right|_{\mathbf{\Delta}=\mathbf{0}} \delta_{j,k}\,. \tag{25}$$

For the derivative matrix $\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}} = \left[\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right]$, define

$$\left\|\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right\|_{\text{sum}} \triangleq \sum_{j,k}\left|\frac{\partial|\lambda_i|}{\partial\delta_{j,k}}\right|\,. \tag{26}$$

Then

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max}\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}\,. \tag{27}$$

This leads to the following mantissa measure for the floating-point realization $\mathbf{X}$

$$\mu_1(\mathbf{X}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))|}{\left\|\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}}\right\|_{\text{sum}}}\,. \tag{28}$$

Obviously, if $\|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})$, then $|\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\mathbf{X}\circ\mathbf{\Delta}))| < 1$ which means that the closed-loop remains stable under the FWL error $\mathbf{\Delta}$. In other words, for a given $\mathbf{X}$, the closed-loop can tolerate those FWL perturbations $\mathbf{\Delta}$ whose norms $\|\mathbf{\Delta}\|_{\max}$ are less than $\mu_1(\mathbf{X})$. The larger $\mu_1(\mathbf{X})$ is, the larger FWL errors the closed-loop system can tolerate. Similar to (19), from the mantissa measure $\mu_1(\mathbf{X})$, an estimate of β_w^{min} is given as

$$\hat{\beta}_{w1}^{min} \triangleq -\lfloor\log_2\mu_1(\mathbf{X})\rfloor - 1\,. \tag{29}$$

The assumption of small $\mathbf{\Delta}$ is usually valid in floating-point implementation. Generally speaking, there is no rigorous relationship between $\mu_0(\mathbf{X})$ and $\mu_1(\mathbf{X})$, but $\mu_1(\mathbf{X})$ is connected with a lower bound of $\mu_0(\mathbf{X})$ in some manners: there are "stable perturbation cubes" larger than $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_1(\mathbf{X})\}$ while there is no "stable perturbation cube" larger than $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{X})\}$ (Wu *et al.*, 2000, 2001). Hence, in most cases, it is reasonable to take that $\mu_1(\mathbf{X}) \le \mu_0(\mathbf{X})$ and $\hat{\beta}_{w1}^{min} \ge \hat{\beta}_{w0}^{min}$. More importantly, unlike the measure $\mu_0(\mathbf{X})$, the value of $\mu_1(\mathbf{X})$ can be computed explicitly. It is easy to see that

$$\left.\frac{\partial|\lambda_i|}{\partial\mathbf{\Delta}}\right|_{\mathbf{\Delta}=\mathbf{0}} = \frac{\partial|\lambda_i|}{\partial\mathbf{X}}\circ\mathbf{X}\,. \tag{30}$$

Let $\mathbf{p}_i$ be a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X})$ corresponding to the eigenvalue λ_i. Define

$$\mathbf{M}_{\mathbf{p}} \stackrel{\triangle}{=} [\,\mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_{m+n}\,] \tag{31}$$

and

$$\mathbf{M}_{\mathbf{y}} \stackrel{\triangle}{=} [\,\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_{m+n}\,] = \mathbf{M}_{\mathbf{p}}^{-H} \tag{32}$$

where the superscript H denotes the conjugate transpose operator and $\mathbf{y}_i$ is called the reciprocal left eigenvector related to $\mathbf{p}_i$. The following lemma is due to Li (1998).

Lemma 1 Let $\overline{\mathbf{A}}(\mathbf{X}) = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2$ given in (10) be diagonalizable. Then

$$\frac{\partial \lambda_i}{\partial \mathbf{X}} = \mathbf{M}_1^T \mathbf{y}_i^* \mathbf{p}_i^T \mathbf{M}_2^T \tag{33}$$

where the superscript $*$ denotes the conjugate operation and T the transpose operator.

Comments: The necessary and sufficient condition for $\overline{\mathbf{A}}(\mathbf{X})$ being diagonalizable is that it has $m+n$ linearly independent eigenvectors. This includes two cases. Firstly, $\overline{\mathbf{A}}(\mathbf{X})$ has $m+n$ distinct eigenvalues. In this case, we can differentiate eigenvalues simply by their values. Secondly, the eigenvalues of $\overline{\mathbf{A}}(\mathbf{X})$ are not all distinct but there are $m+n$ linearly independent eigenvectors. In this case, we can differentiate eigenvalues by their corresponding eigenvectors.

The following proposition shows that, given a $\mathbf{X}$, the value of $\mu_1(\mathbf{X})$ can easily be calculated.

Proposition 4 Let $\overline{\mathbf{A}}(\mathbf{X})$ be diagonalizable. Then

$$\mu_1(\mathbf{X}) = \min_{i \in \{1,\cdots,m+n\}} \frac{|\lambda_i|(1-|\lambda_i|)}{\|(\mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T] \mathbf{M}_2^T) \circ \mathbf{X}\|_{\mathrm{sum}}}\,. \tag{34}$$

Proof: Noting $|\lambda_i| = \sqrt{\lambda_i^* \lambda_i}$ leads to

$$\frac{\partial |\lambda_i|}{\partial \mathbf{X}} = \frac{1}{2\sqrt{\lambda_i^*\lambda_i}} \left(\frac{\partial \lambda_i^*}{\partial \mathbf{X}} \lambda_i + \lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right) = \frac{1}{2|\lambda_i|} \left(\left(\frac{\partial \lambda_i}{\partial \mathbf{X}} \right)^* \lambda_i + \lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right) = \frac{1}{|\lambda_i|} \mathrm{Re}\left[\lambda_i^* \frac{\partial \lambda_i}{\partial \mathbf{X}} \right]. \tag{35}$$

Combining (28), (30), (35) and Lemma 1 results in this proposition.

Replacing $\mu_0(\mathbf{X})$ with $\mu_1(\mathbf{X})$ in (21) leads to a computationally tractable FWL closed-loop stability measure

$$\rho_1(\mathbf{X}) \stackrel{\triangle}{=} \mu_1(\mathbf{X})/\gamma(\mathbf{X})\,. \tag{36}$$

From the above measure, an estimate of β^{min} is given as

$$\hat{\beta}_1^{min} \stackrel{\triangle}{=} -\lfloor \log_2 \rho_1(\mathbf{X}) \rfloor + 1\,. \tag{37}$$

It is useful to compare the proposed measure with the previous results, especially the most recent one given by Whidborne & Gu (2001). For a complex-valued matrix $\mathbf{Y} = [y_{j,k}]$, define the Frobenius norm

$$\|\mathbf{Y}\|_{\mathrm{F}} \triangleq \left(\sum_{j,k} y_{j,k}^* y_{j,k} \right)^{1/2} . \tag{38}$$

Under an assumption that the exponent bits are unlimited, the computationally tractable weighted closed-loop eigenvalue sensitivity index addressed in (Whidborne & Gu, 2001) is given by

$$\Upsilon(\mathbf{X}) \triangleq \sum_{i=1}^{m+n} \alpha_i \Upsilon_i(\mathbf{X}) \tag{39}$$

where α_i are non-negative weighting scalars and $\Upsilon_i(\mathbf{X})$ are single-eigenvalue sensitivities defined by

$$\Upsilon_i(\mathbf{X}) \triangleq \|\mathbf{X}\|_{\mathrm{F}}^2 \left\| \frac{\partial \lambda_i}{\partial \mathbf{X}} \right\|_{\mathrm{F}}^2 . \tag{40}$$

The thinking behind the above definition is as follows. From a first-order approximation, it can easily be shown that

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta})) - \lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max} \|\mathbf{X}\|_{\mathrm{F}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{X}} \right\|_{\mathrm{F}} . \tag{41}$$

Therefore, for those multiplicative perturbations bounded by $\|\mathbf{\Delta}\|_{\max}$, a small $\Upsilon_i(\mathbf{X})$ will limit the resulting change of the corresponding eigenvalue within a small range.

The first obvious observation is that $\rho_1(\mathbf{X})$ considers both the mantissa and exponent of floating-point arithmetic and is therefore able to handle all the aspects of underflow, overflow and closed-loop stability, while $\Upsilon(\mathbf{X})$ only considers the mantissa part of floating-point arithmetic and is thus "incomplete". Secondly, it can be seen that $\Upsilon(\mathbf{X})$ deals with the sensitivity of λ_i while $\rho_1(\mathbf{X})$ ($\mu_1(\mathbf{X})$) considers the the sensitivity of $|\lambda_i|$. It is well-known that the stability of a discrete-time linear time-invariant system depends only on the module of its eigenvalues. As $\Upsilon(\mathbf{X})$ includes the unnecessary eigenvalue arguments in consideration, it is generally conservative in comparison with $\rho_1(\mathbf{X})$. Thirdly, $\rho_1(\mathbf{X})$ uses $\left\| \frac{\partial |\lambda_i|}{\partial \mathbf{X}} \circ \mathbf{X} \right\|_{\mathrm{sum}}$ while $\Upsilon(\mathbf{X})$ uses $\|\mathbf{X}\|_{\mathrm{F}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{X}} \right\|_{\mathrm{F}}$ in checking the change of an eigenvalue. It is easy to see that

$$|\lambda_i(\overline{\mathbf{A}}(\mathbf{X} + \mathbf{X} \circ \mathbf{\Delta}))| - |\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))| \le \|\mathbf{\Delta}\|_{\max} \left\| \frac{\partial |\lambda_i|}{\partial \mathbf{X}} \circ \mathbf{X} \right\|_{\mathrm{sum}} \le \|\mathbf{\Delta}\|_{\max} \|\mathbf{X}\|_{\mathrm{F}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{X}} \right\|_{\mathrm{F}} . \tag{42}$$

Obviously, $\left\| \frac{\partial |\lambda_i|}{\partial \mathbf{X}} \circ \mathbf{X} \right\|_{\mathrm{sum}}$ gives a more accurate limit than $\|\mathbf{X}\|_{\mathrm{F}} \left\| \frac{\partial \lambda_i}{\partial \mathbf{X}} \right\|_{\mathrm{F}}$ does on the change of the corresponding eigenvalue module due to the multiplicative perturbations. This again implies that $\rho_1(\mathbf{X})$ is less conservative than $\Upsilon(\mathbf{X})$ in estimating the robustness of closed-loop stability with respect to controller perturbations. The fourth observation is that $\rho_1(\mathbf{X})$ provides an estimate of β^{min}, $\hat{\beta}_1^{min}$ in (37), while $\Upsilon(\mathbf{X})$ cannot provide information on bit length to the designer. One reason is that the measure

$\rho_1(\mathbf{X})$ consists of two components, with $\mu_1(\mathbf{X})$ addressing the stability margin and eigenvalue sensitivity linked to the mantissa bits, and $\gamma(\mathbf{X})$ considering the exponent bits, while $\Upsilon(\mathbf{X})$ only focuses on the eigenvalue sensitivity partially linked to the mantissa part. The other reason is that, over all the closed-loop eigenvalues, $\mu_1(\mathbf{X})$ considers the "worst" one while $\Upsilon(\mathbf{X})$ considers a "weighted average".

5 Optimization Procedure

As different realizations $\mathbf{X}$ have different values of the FWL closed-loop stability measure $\rho_1(\mathbf{X})$, it is of practical importance to find an "optimal" realization $\mathbf{X}_{\text{opt}}$ that maximizes $\rho_1(\mathbf{X})$. The controller implemented with this optimal realization $\mathbf{X}_{\text{opt}}$ needs a minimum bit length and has a maximum tolerance to the FWL error. This optimal controller realization problem is formally defined as

$$\upsilon \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} \rho_1(\mathbf{X})\,. \tag{43}$$

Assume that a controller has been designed using some standard controller design method. This controller, denoted as

$$\mathbf{X}_0 \triangleq \begin{bmatrix} \mathbf{D}_C^0 & \mathbf{C}_C^0 \\ \mathbf{B}_C^0 & \mathbf{A}_C^0 \end{bmatrix}, \tag{44}$$

is used as the initial controller realization in the above optimal controller realization problem. Let $\mathbf{p}_{0i}$ be a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X}_0)$ corresponding to the eigenvalue λ_i, and $\mathbf{y}_{0i}$ be the reciprocal left eigenvector related to $\mathbf{p}_{0i}$. The definition of $\mathcal{S}_C$ in (8) means that

$$\mathbf{X} \triangleq \mathbf{X}(\mathbf{T}) = \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{45}$$

where $\det(\mathbf{T}) \neq 0$. It can then be shown that

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{X}_0) \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \tag{46}$$

which implies that

$$\mathbf{p}_i = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{p}_{0i}, \quad \mathbf{y}_i = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{y}_{0i}. \tag{47}$$

Hence

$$\begin{aligned} \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_i^* \mathbf{p}_i^T] \mathbf{M}_2^T &= \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_{0i}^* \mathbf{p}_{0i}^T] \mathbf{M}_2^T \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \\ &\triangleq \begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{\Phi}_i \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \end{aligned} \tag{48}$$

with $\mathbf{\Phi}_i = \mathbf{M}_1^T \mathrm{Re}[\lambda_i^* \mathbf{y}_{0i}^* \mathbf{p}_{0i}^T] \mathbf{M}_2^T$. Define the following cost function:

$$f(\mathbf{T}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \left(\frac{\left\| \left(\begin{bmatrix} \mathbf{I}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{T} \end{bmatrix} \mathbf{\Phi}_i \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T} \end{bmatrix} \right) \circ \mathbf{X}(\mathbf{T}) \right\|_{\text{sum}}}{|\lambda_i|(1-|\lambda_i|)} \log_2 \frac{4\|\mathbf{X}(\mathbf{T})\|_{\max}}{g(\mathbf{X}(\mathbf{T}))} \right)^{-1}. \tag{49}$$

Then the optimal controller realization problem (43) can be posed as the following optimization problem:

$$v = \max_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n} \\ \det \mathbf{T}\neq 0}} f(\mathbf{T})\,. \tag{50}$$

Efficient numerical optimization methods exist for solving for this optimization problem to provide an optimal transformation matrix $\mathbf{T}_{\rm opt}$. With $\mathbf{T}_{\rm opt}$, the optimal realization $\mathbf{X}_{\rm opt}$ can readily be computed.

6 Numerical Examples

Two examples are used to illustrate the proposed design procedure for obtaining optimal FWL floating-point controller realizations and to compare it with the method given in (Whidborne & Gu, 2001).

Example 1. This example, taken from (Gevers & Li, 1993), has been studied by Whidborne & Gu (2001). The discrete-time plant is given by

$$\begin{aligned}
\mathbf{A}_P &= \begin{bmatrix} 3.7156e+0 & -5.4143e+0 & 3.6525e+0 & -9.6420e-1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \\
\mathbf{B}_P &= [1 \quad 0 \quad 0 \quad 0]^T\,, \\
\mathbf{C}_P &= [1.1160e-6 \quad 4.3000e-8 \quad 1.0880e-6 \quad 1.4000e-8]\,.
\end{aligned}$$

The initial realization of the digital controller is given by

$$\begin{aligned}
\mathbf{A}_C^0 &= \begin{bmatrix} 2.6743e+0 & -5.7446e+0 & 2.5101e+0 & -9.1782e-1 \\ 2.8769e-1 & -2.7446e-2 & -6.9444e-1 & -8.9358e-3 \\ -3.3773e-1 & 9.8699e-1 & -3.2925e-1 & -4.2367e-3 \\ -8.3021e-2 & -3.1988e-3 & 9.1906e-1 & -1.0415e-3 \end{bmatrix}, \\
\mathbf{B}_C^0 &= [1.0959e+6 \quad 6.3827e+5 \quad 3.0262e+5 \quad 7.4392e+4]^T\,, \\
\mathbf{C}_C^0 &= [1.8180e-1 \quad -2.8313e-1 \quad 5.0006e-2 \quad 6.1722e-2]\,, \quad \mathbf{D}_C^0 = 0\,.
\end{aligned}$$

Based on the proposed FWL closed-loop stability measure, the optimization problem (50) was formed and solved for using the MATLAB routine *fminsearch.m* to obtain an optimal transformation matrix

$$\mathbf{T}_{\rm opt} = \begin{bmatrix} 7.7275e+3 & -1.0904e+2 & -2.1292e+2 & 9.8042e+1 \\ 6.9729e+3 & 2.1370e+3 & 4.4988e+1 & 2.1812e+2 \\ 6.2844e+3 & 3.9092e+3 & 2.9303e+2 & 2.9240e+2 \\ 5.5879e+3 & 5.2862e+3 & 5.5027e+2 & 3.4367e+2 \end{bmatrix}$$

and the corresponding optimal realization of the digital controller $\mathbf{X}_{\rm opt}$ given by

$$\mathbf{A}_C^{\rm opt} = \begin{bmatrix} -1.4441e+0 & -1.0500e+0 & -6.0800e-2 & -1.0102e-1 \\ 3.8412e+0 & 2.4034e+0 & 6.7143e-2 & 1.7402e-1 \\ -1.3159e+1 & -4.5856e+0 & 5.3403e-1 & -6.8843e-1 \\ 3.2330e-1 & -2.1078e+0 & -6.6254e-2 & 8.2322e-1 \end{bmatrix},$$

$$\mathbf{B}_C^{\mathrm{opt}} = \begin{bmatrix} 1.6342e+2 & -2.5378e+2 & 9.1370e+2 & -6.1106e-2 \end{bmatrix}^T ,$$

$$\mathbf{C}_C^{\mathrm{opt}} = \begin{bmatrix} 8.9770e+1 & -1.0310e+2 & -2.8290e+0 & -8.0995e+0 \end{bmatrix}, \quad \mathbf{D}_C^{\mathrm{opt}} = 0 .$$

An "optimal" controller realization problem was given in (Whidborne & Gu, 2001) based on the weighted closed-loop eigenvalue sensitivity index (39). We will use the index "s", rather then "opt", to denote the solution of this "optimal" controller realization problem. For this example, the transformation matrix solution obtained using the MATLAB routine *fminsearch.m* given in (Whidborne & Gu, 2001) is

$$\mathbf{T}_{\mathrm{s}} = \begin{bmatrix} 8.1477e+3 & 0 & 0 & 0 \\ 7.0104e+3 & 2.2671e+3 & 0 & 0 \\ 6.1991e+3 & 3.9989e+3 & 1.1558e+2 & 0 \\ 5.6761e+3 & 5.2680e+3 & 3.5814e+2 & 1.5299e+1 \end{bmatrix}$$

with the corresponding controller realization $\mathbf{X}_{\mathrm{s}}$ given by

$$\mathbf{A}_C^{\mathrm{s}} = \begin{bmatrix} -9.9795e-1 & -9.5988e-1 & -4.7357e-3 & -1.7234e-3 \\ 2.1137e+0 & 1.6951e+0 & -2.2171e-2 & 5.2689e-3 \\ -1.4177e+0 & 6.1144e-1 & 6.7870e-1 & -9.0420e-2 \\ 1.9428e+0 & -2.4577e+0 & 4.2234e-1 & 9.4079e-1 \end{bmatrix},$$

$$\mathbf{B}_C^{\mathrm{s}} = \begin{bmatrix} 1.3451e+2 & -1.3439e+2 & 5.3833e+1 & -2.5633e+1 \end{bmatrix}^T ,$$

$$\mathbf{C}_C^{\mathrm{s}} = \begin{bmatrix} 1.5673e+2 & -1.1677e+2 & 2.7885e+1 & 9.4430e-1 \end{bmatrix}, \quad \mathbf{D}_C^{\mathrm{s}} = 0 .$$

It is obvious that the true minimum exponent bit length β_e^{min} for a realization $\mathbf{X}$ can directly be obtained by examining the elements of $\mathbf{X}$. The true minimum mantissa bit length β_w^{min} however can only be obtained through simulation. That is, starting from a very large β_w, reduce β_w by one bit and check the closed-loop stability. The process is repeated until there appears closed-loop instability at $\beta_w = \beta_{wu}$. Then $\beta_w^{min} = \beta_{wu} + 1$. Table 1 summarizes the various measures, the corresponding estimated minimum bit lengths and the true minimum bit lengths for the three controller realizations $\mathbf{X}_0$, $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{opt}}$, respectively. It can be seen that the floating-point implementation of $\mathbf{X}_0$ needs at least 26 bits (20 mantissa bits and 5 exponent bits) while the implementation of $\mathbf{X}_{\mathrm{opt}}$ needs at least 13 bits (8 mantissa bits and 4 exponent bits). The reduction in the bit length required is 13 (12-bit reduction for the mantissa part and 1-bit reduction for the exponent part). Comparing $\mathbf{X}_{\mathrm{opt}}$ with $\mathbf{X}_{\mathrm{s}}$, it can be seen that $\mathbf{X}_{\mathrm{opt}}$ needs one bit less in the exponent part and one bit less in the mantissa part.

Notice that any realization $\mathbf{X} \in \mathcal{S}_C$ implemented in infinite precision will achieve the exact performance of the infinite-precision implemented $\mathbf{X}_0$, which is the designed controller performance. For this reason, the infinite-precision implemented $\mathbf{X}_0$ is referred to as the ideal controller realization $\mathbf{X}_{\mathrm{ideal}}$. Figure 1 compares the unit impulse response of the plant output $y(k)$ for the ideal controller $\mathbf{X}_{\mathrm{ideal}}$ with those of the 8-mantissa-bit plus 5-exponent-bit implemented $\mathbf{X}_{\mathrm{s}}$ and 8-mantissa-bit plus 4-exponent-bit implemented $\mathbf{X}_{\text{opt}}$. The 8-mantissa-bit implemented $\mathbf{X}_0$ quickly becomes unstable and is not shown here. From Figure 1, it can be seen that the closed-loop system with the 13-bit implemented $\mathbf{X}_{\text{opt}}$ is stable while the system with the 14-bit implemented $\mathbf{X}_{\text{s}}$ is unstable. Figure 2 compares the unit impulse response of $y(k)$ for $\mathbf{X}_{\text{ideal}}$ with those of the 9-mantissa-bit plus 5-exponent-bit implemented $\mathbf{X}_{\text{s}}$ and the 9-mantissa-bit plus 4-exponent-bit implemented $\mathbf{X}_{\text{opt}}$. Again the 9-mantissa-bit implemented $\mathbf{X}_0$ is unstable and is not shown. It can be seen that the response with the 14-bit implemented $\mathbf{X}_{\text{opt}}$ is clearly closer to the ideal performance than that of the 15-bit implemented $\mathbf{X}_{\text{s}}$.

Example 2. This example is taken from a continuous-time H_∞ robust control example studied in (Keel & Bhattacharryya, 1997; Whidborne *et al.*, 2001). The continuous-time plant model and H_∞ controller are sampled with a sampling period of 4 ms to obtain the discrete-time plant

$$\begin{aligned}
\mathbf{A}_P &= \begin{bmatrix} 1.9980e+0 & -9.9800e-1 \\ 1 & 0 \end{bmatrix}, \\
\mathbf{B}_P &= [1 \quad 0]^T, \quad \mathbf{C}_P = [3.9880e-3 \quad -4.0040e-3],
\end{aligned}$$

and the initial realization of the digital controller

$$\begin{aligned}
\mathbf{A}_C^0 &= \begin{bmatrix} 2.3985e+0 & -1.8017e+0 & 4.0317e-1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \\
\mathbf{B}_C^0 &= [1 \quad 0 \quad 0]^T, \\
\mathbf{C}_C^0 &= [-7.3591e+1 \quad 1.4661e+2 \quad -7.3018e+1], \quad \mathbf{D}_C^0 = 1.2450e+2.
\end{aligned}$$

The MATLAB routine *fminsearch.m* was used to solve for the optimization problem based on the FWL closed-loop stability measure presented in this paper to obtain an optimal transformation matrix

$$\mathbf{T}_{\text{opt}} = \begin{bmatrix} 1.8515e+2 & 7.2829e-1 & 9.7266e+0 \\ 1.8540e+2 & 1.6951e+1 & -2.3477e+0 \\ 1.8566e+2 & 3.3300e+1 & -1.4508e+1 \end{bmatrix}$$

and the corresponding optimal realization of the digital controller $\mathbf{X}_{\text{opt}}$ with

$$\begin{aligned}
\mathbf{A}_C^{\text{opt}} &= \begin{bmatrix} 1.0006e+0 & -8.8718e-2 & 9.9092e-2 \\ -2.7168e-2 & 1.0178e+0 & -4.5738e-1 \\ -3.6546e-2 & 3.2513e-2 & 3.8007e-1 \end{bmatrix}, \\
\mathbf{B}_C^{\text{opt}} &= [-6.8999e+0 \quad 9.2711e+1 \quad 1.2450e+2]^T, \\
\mathbf{C}_C^{\text{opt}} &= [-3.6469e-2 \quad 2.7168e-2 \quad -6.1334e-1], \quad \mathbf{D}_C^{\text{opt}} = 1.2450e+2.
\end{aligned}$$

Based on the method of the weighted closed-loop eigenvalue sensitivity index (Whidborne & Gu, 2001), the MATLAB routine *fminsearch.m* found a transformation matrix solution

$$\mathbf{T}_{\text{s}} = \begin{bmatrix} 1.8446e+2 & 0 & 0 \\ 1.8500e+2 & 2.9433e+0 & 0 \\ 1.8553e+2 & 5.9061e+0 & 8.3753e-3 \end{bmatrix}$$

with the corresponding controller realization $\mathbf{X}_{\rm s}$ given by

$$\begin{aligned}
\mathbf{A}_C^{\rm s} &= \begin{bmatrix} 9.9711e-1 & -1.5840e-2 & 1.8305e-5 \\ 3.2077e-5 & 9.9558e-1 & -1.1505e-3 \\ -2.8762e-2 & 2.5216e-1 & 4.0584e-1 \end{bmatrix}, \\
\mathbf{B}_C^{\rm s} &= \left[5.4211e-3 \quad -3.4074e-1 \quad 1.2019e+2\right]^T, \\
\mathbf{C}_C^{\rm s} &= \left[-2.9785e-2 \quad 2.6087e-1 \quad -6.1154e-1\right], \quad \mathbf{D}_C^{\rm s} = 1.2450e+02\,.
\end{aligned}$$

Table 2 summarizes the various measures, the corresponding estimated minimum bit lengths and the true minimum bit lengths for $\mathbf{X}_0$, $\mathbf{X}_{\rm s}$ and $\mathbf{X}_{\rm opt}$. Obviously, the implementation of $\mathbf{X}_0$ needs at least 30 bits (25 mantissa bits and 4 exponent bits) while the implementation of $\mathbf{X}_{\rm opt}$ requires at least 12 bits (7 mantissa bits and 4 exponent bits). It can be seen that the optimization results in a reduction of 18 bits for the mantissa part. It is interesting to note that the realization $\mathbf{X}_{\rm s}$, while reducing 16 bits in the required β_w^{min}, actually increases the required β_e^{min} by one bit, compared with $\mathbf{X}_0$. This is not surprising, since the measure $\Upsilon(\mathbf{X})$ completely neglects the exponent part. Figure 3 compares the unit impulse response of the plant output $y(k)$ for the ideal controller $\mathbf{X}_{\rm ideal}$ with those of the 14-bit implemented $\mathbf{X}_{\rm s}$ (8 mantissa bits and 5 exponent bits) and the 14-bit implemented $\mathbf{X}_{\rm opt}$ (9 mantissa bits and 4 exponent bits). It can be seen that the closed-loop system with the 14-bit implemented $\mathbf{X}_{\rm opt}$ is stable while the system with the 14-bit implemented $\mathbf{X}_{\rm s}$ is unstable. Figure 4 compares the unit impulse response of $y(k)$ for $\mathbf{X}_{\rm ideal}$ with those of the 15-bit implemented $\mathbf{X}_{\rm s}$ (9 mantissa bits and 5 exponent bits) and the 15-bit implemented $\mathbf{X}_{\rm opt}$ (10 mantissa bits and 4 exponent bits). The performance of the 15-bit implemented $\mathbf{X}_{\rm opt}$ is clearly closer to the ideal performance than that of the 15-bit implemented $\mathbf{X}_{\rm s}$.

7 Conclusions

The closed-loop stability issue of finite-precision realizations has been investigated for digital controller implemented in floating-point arithmetic. A new computationally tractable FWL closed-loop stability measure has been derived for floating-point controller realizations. Unlike the existing methods, which only consider the mantissa part of floating-point scheme, the proposed measure takes into account both the exponent and mantissa parts of floating-point format. It has been shown that this new measure yields a more accurate estimate for the FWL closed-loop stability. Based on this FWL closed-loop stability measure, the optimal controller realization problem has been formulated, which can easily be solved for using standard numerical optimization algorithms. Two numerical examples have demonstrated that the proposed design procedure yields computationally efficient controller realizations suitable for FWL float-point implementation in real-time applications.

Acknowledgements

J. Wu and S. Chen wish to thank the support of the UK Royal Society under a KC Wong fellowship (RL/ART/CN/XFI/KCW/11949). J. Wu and J. Chu wish to thank the support of the National Natural Science Foundation of China under Grant 60174026.

References

Chen, S., Istepanian, R.S.H., Wu, J., & Chu, J. (2000). Comparative study on optimizing closed-loop stability bounds of finite-precision controller structures with shift and delta operators. *Systems and Control Letters* 40(3), 153–163.

Chen, S., Wu, J., Istepanian, R.S.H., & Chu, J. (1999). Optimizing stability bounds of finite-precision PID controller structures. *IEEE Trans. Automatic Control* 44(11), 2149–2153.

Fialho, I.J., & Georgiou, T.T. (1994). On stability and performance of sampled-data systems subject to wordlength constraint. *IEEE Trans. Automatic Control* 39(12), 2476–2481.

Fialho, I.J., & Georgiou, T.T. (1999). Optimal finite wordlength digital controller realization. In *Proc. American Control Conf.* (San Diego, USA), June 2-4, 1999, 4326–4327.

Gevers, M., & Li, G. (1993). *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. London: Springer Verlag.

Istepanian, R.S.H., & Whidborne, J.F., eds. (2001). *Digital Controller Implementation and Fragility: A Modern Perspective*. Godalming, UK: Springer Verlag.

Istepanian, R.S.H., Whidborne, J.F., & Bauer, P. (2000). Stability analysis of block floating point digital controllers. In *Proc. UKACC Int. Conf. Control 2000* (Cambridge, UK), CD-ROM, 6 pages.

Kalliojärvi, K., & Astola, J. (1996). Roundoff errors in block-floating-point systems. *IEEE Trans. Signal Processing* 44(4), 783–790.

Keel, L.H., & Bhattacharryya, S.P. (1997). Robust, fragile, or optimal? *IEEE Trans. Automatic Control* 42(8), 1098–1105.

Li, G. (1998). On the structure of digital controllers with finite word length consideration. *IEEE Trans. Automatic Control* 43(5), 689–693.

Li, G., & Gevers, M. (1990). Optimal finite precision implementation of a state-estimate feedback controller. *IEEE Trans. Circuits and Systems* CAS-38(12), 1487–1499.

Li, G., Wu, J., Chen, S., & Zhao, K.Y. (2000). Optimum structures of digital controllers in sampled-data systems: a roundoff noise analysis. Submitted to *IEE Proc. Control Theory and Applications*.

Liu, K., Skelton, R., & Grigoriadis, K. (1992). Optimal controllers for finite wordlength implementation. *IEEE Trans. Automatic Control* 37(9), 1294–1304.

Madievski, A.G., Anderson, B.D.O., & Gevers, M. (1995). Optimum realizations of sampled-data controllers for FWL sensitivity minimization. *Automatica* 31(3), 367–379.

Mäkilä, P.M. (1999). Fragility and robustness puzzles. In *Proc. 1999 American Control Conf.* (San Diego, USA), pp.2914–2919.

Molchanov, A.P., & Bauer, P.H. (1995). Robust stability of digital feedback control systems with floating point arithmetic. In *Proc. 34th IEEE Int. Conf. Decision and Control* (New Orleans, USA), pp.4251–4258.

Moroney, P., Willsky, A.S., & Houpt, P.K. (1980). The digital implementation of control compensators: the coefficient wordlength issue. *IEEE Trans. Automatic Control* AC-25(4), 621–630.

Ralev, K.R., & Bauer, P.H. (1999). Realization of block floating-point digital filters and application to block implementations. *IEEE Trans. Signal Processing* 47(4), 1076–1086.

Rao, B.D. (1996). Roundoff noise in floating point digital filters. *Control and Dynamic Systems* 75, 79–103.

Rink, R.E., & Chong, H.Y. (1979). Performance of state regulator systems with floating point computation. *IEEE Trans. Automatic Control* 24, 411–421.

Whidborne, J.F., & Gu, D. (2001). Optimal finite-precision controller and filter realizations using floating-point arithmetic. *Research Report* EM2001/07, Department of Mechanical Engineering, King's College London.

Whidborne, J.F., Istepanian, R.S.H., & Wu, J. (2001). Reduction of controller fragility by pole sensitivity minimization. *IEEE Trans. Automatic Control* 46(2), 320–325.

Whidborne, J.F., Wu, J., & Istepanian, R.S.H. (2000). Finite word length stability issues in an l_1 framework. *Int. J. Control* 73(2), 166–176.

Williamson, D. (1991). *Digital Control and Implementation*. Englewood-Cliffs, NJ: Prentice-Hall.

Williamson, D., & Kadiman, K. (1989). Optimal finite wordlength linear quadratic regulation. *IEEE Trans. Automatic Control* 34(12), 1218–1228.

Wu, J., Chen, S., Li, G., Istepanian, R.S.H., & Chu, J. (2000). Shift and delta operator realizations for digital controllers with finite-word-length considerations. *IEE Proc. Control Theory and Applications* 147(6), 664–672.

Wu, J., Chen, S., Li, G., Istepanian, R.S.H., & Chu, J. (2001). An improved closed-loop stability related measure for finite-precision digital controller realizations. *IEEE Trans. Automatic Control* 46(7), 1162–1166.

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

Table 1: Various measures, corresponding estimated minimum bit lengths and true minimum bit lengths for three controller realizations $\mathbf{X}_0$, $\mathbf{X}_\mathrm{s}$ and $\mathbf{X}_\mathrm{opt}$ of Example 1.

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

Table 2: Various measures, corresponding estimated minimum bit lengths and true minimum bit lengths for three controller realizations $\mathbf{X}_0$, $\mathbf{X}_\mathrm{s}$ and $\mathbf{X}_\mathrm{opt}$ of Example 2.

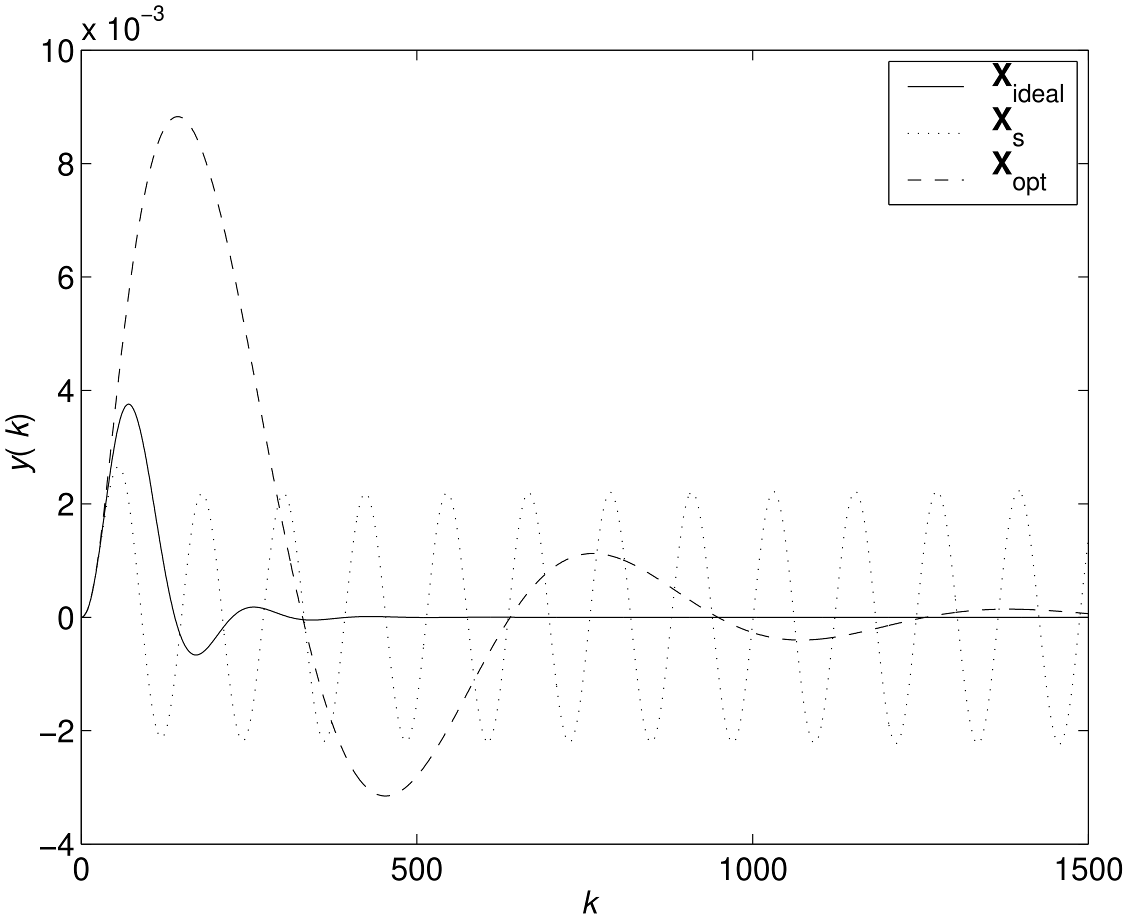

Figure 1: Unit impulse response $y(k)$ for $\mathbf{X}_{\rm ideal}$, 14-bit implemented $\mathbf{X}_{\rm s}$ (8 mantissa bits and 5 exponent bits) and 13-bit implemented $\mathbf{X}_{\rm opt}$ (8 mantissa bits and 4 exponent bits) of Example 1.

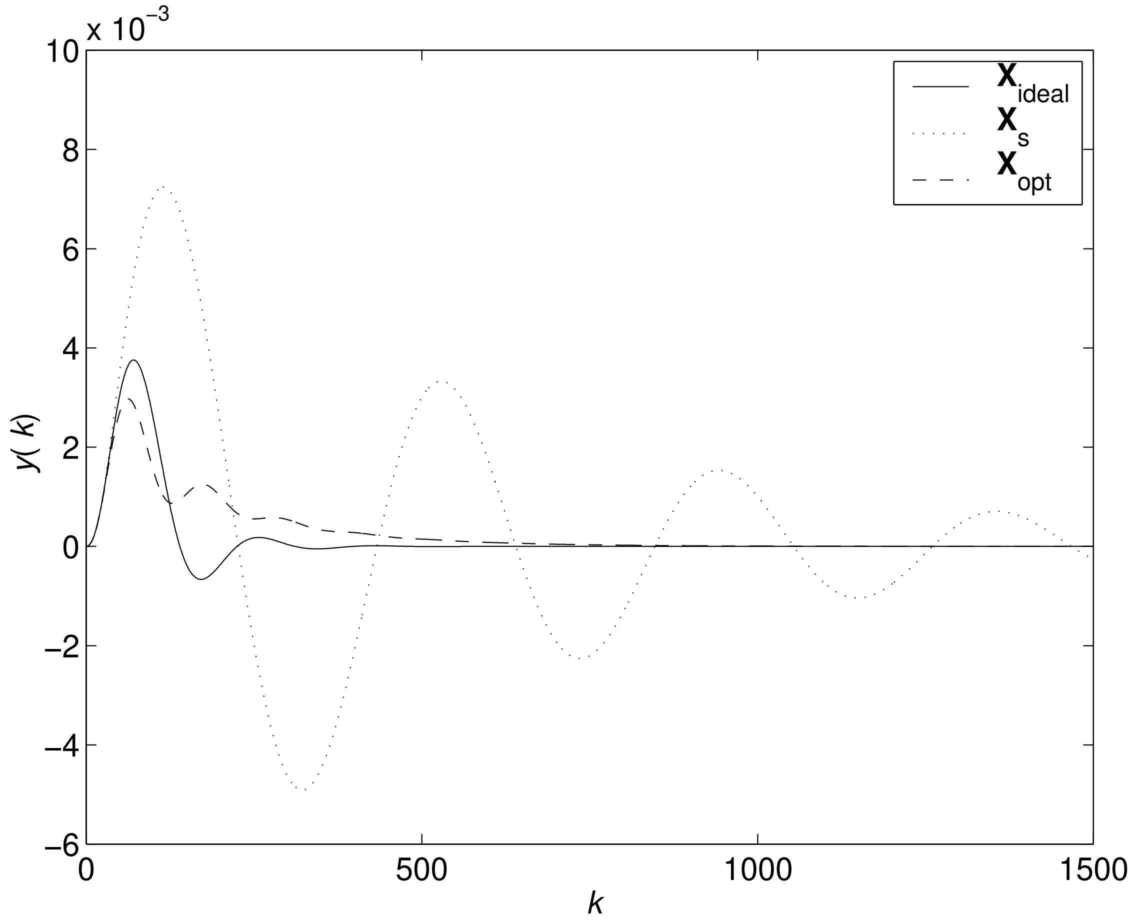

Figure 2: Unit impulse response $y(k)$ for $\mathbf{X}_{\rm ideal}$, 15-bit implemented $\mathbf{X}_{\rm s}$ (9 mantissa bits and 5 exponent bits) and 14-bit implemented $\mathbf{X}_{\rm opt}$ (9 mantissa bits and 4 exponent bits) of Example 1.

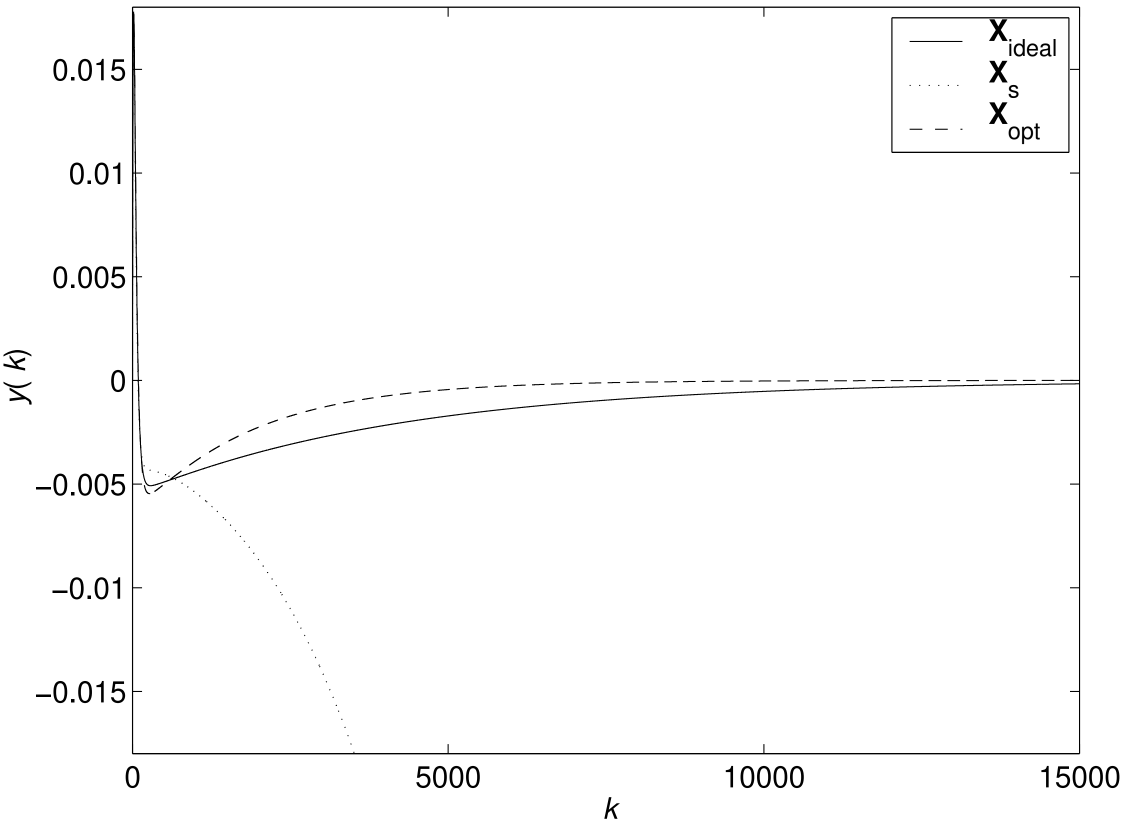

Figure 3: Unit impulse response $y(k)$ for $\mathbf{X}_{\text{ideal}}$, 14-bit implemented $\mathbf{X}_{\text{s}}$ (8 mantissa bits and 5 exponent bits) and 14-bit implemented $\mathbf{X}_{\text{opt}}$ (9 mantissa bits and 4 exponent bits) of Example 2.

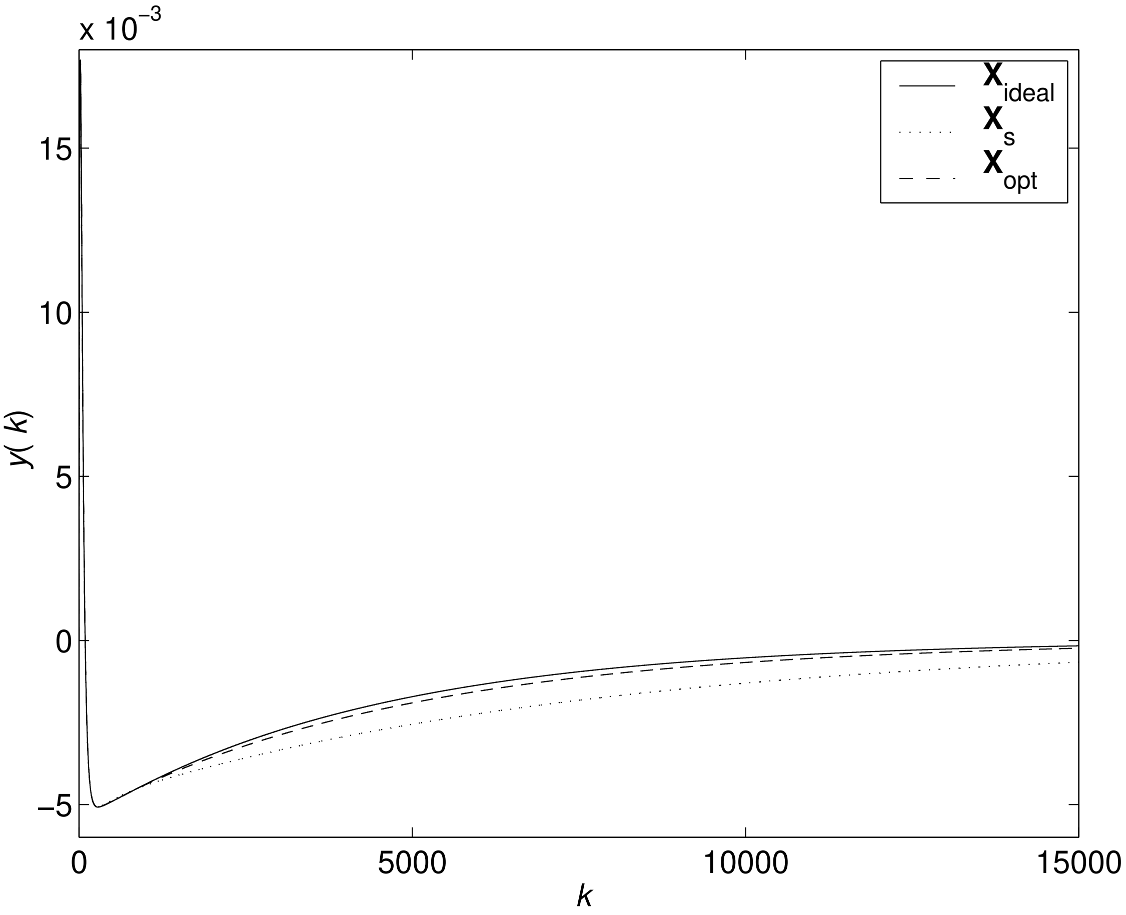

Figure 4: Unit impulse response $y(k)$ for $\mathbf{X}_{\rm ideal}$, 15-bit implemented $\mathbf{X}_{\rm s}$ (9 mantissa bits and 5 exponent bits) and 15-bit implemented $\mathbf{X}_{\rm opt}$ (10 mantissa bits and 4 exponent bits) of Example 2.