Quantization in Control Systems and Forward Error Analysis of Iterative Numerical Algorithms

Ammar Hasan [*] **Eric C. Kerrigan** [*,**]
George A. Constantinides [*]

[*] *Department of Electrical and Electronic Engineering, Imperial College London, U.K.*
e-mail: {ammar.hasan07, e.kerrigan, g.constantinides}@imperial.ac.uk
[**] *Department of Aeronautics, Imperial College London, U.K.*

Abstract:
The use of control theory to study iterative algorithms, which can be considered as dynamical systems, opens many opportunities to find new tools for analysis of algorithms. In this paper we show that results from the study of quantization effects in control systems can be used to find systematic ways for forward error analysis of iterative algorithms. The proposed schemes are applied to the classical iterative methods for solving a system of linear equations. The obtained bounds are compared with bounds given in the numerical analysis literature.

Keywords: algorithms as dynamical systems, iterative algorithms, forward error analysis, quantization, finite precision, numerical analysis.

1. INTRODUCTION

It is clear that many iterative numerical algorithms can be considered as dynamical systems, which can be studied using control systems theory. Although having known this fact for many years (Hurt (1967), Ortega (1973)) and the possible potential of this approach to provide us with new tools for analysis of algorithms, this idea has not been very popular in the control systems literature. However, in recent years, interest has been shown in this application of control theory, e.g. see Bhaya and Kaszkurewicz (2007); Chehab and Laminie (2005); Chu (2008). Most of this work is focused on design of new iterative methods and analysis of algorithms in exact precision. Kashima and Yamamoto (2007) have done some work on algorithms in finite precision and demonstrated the role of the internal model principle in representing iterative algorithms as closed-loop feedback systems. They have also shown that in Newton's method the sum of solution error is bounded if the sum of finite precision errors in each iteration is bounded. However, they have not given a bound for either the finite precision errors or the solution error. Hasan et al. (2010) have used the control tools of l_∞-stability and input-to-state stability (ISS) to present a scheme for forward error analysis of iterative algorithms; forward error analysis is the process of finding bounds for the error in solutions obtained from an algorithm in finite precision.

In this paper we explore other ways of forward error analysis based on the control-theoretic point of view of algorithms. It is advantageous to have more than one scheme for forward error analysis as the requirements of a particular scheme might not be satisfied by an algorithm. Moreover, if applicable, each forward error analysis scheme could result in a different bound, lending more insight into the dependence of solution error on problem data and number representation.

We utilize results from the study of quantization in control systems to present two systematic ways for forward error analysis of iterative algorithms. Quantization refers to the restriction of a variable to a discrete set of values and can arise in control systems due to digital implementation, analog-to-digital converters, digital-to-analog converters, etc. A more detailed discussion on the sources and effects of quantization in control systems can be found in Haimovich (2006) and references therein. Miller et al. (1988) and Kofman et al. (2007) have looked at the problem of finding the ultimate bounds for control systems in the presence of perturbations that arise due to quantization. In this paper we use their results to present new schemes for forward error analysis of algorithms. The proposed techniques are applied on the classical iterative methods for solving a system of linear equations. The obtained bounds are compared with bounds given in the numerical analysis literature and are numerically shown to be tighter.

Section 2 discusses how algorithms can be represented as dynamical systems. In Sections 3 and 4 we present new ways for forward error analysis of algorithms by utilizing the results in Miller et al. (1988) and Kofman et al. (2007), respectively. The proposed methods of forward error analysis are applied on the classical iterative methods in Section 5, where we also compare the obtained forward error bounds with bounds in the numerical analysis literature. The concluding remarks are given in Section 6.

Notation: We denote vectors, scalars and functions with small letters and matrices with capital letters. We denote

[⋆] This work was funded by the EPSRC under grant number EP/G031576/1.

the vector Euclidean norm and the matrix spectral norm (or the induced matrix 2-norm) with $\|\cdot\|_2$. We denote the vector infinity norm and the induced matrix infinity norm with $\|\cdot\|_\infty$. We use $|\cdot|$ to denote the componentwise absolute value of a matrix or a vector. We will often consider a system of linear equations $Ax^* = b$. We will always assume that $A \in \mathbb{R}^{n\times n}$, $x^* \in \mathbb{R}^n$ and $b \in \mathbb{R}^n$.

2. ALGORITHMS AS DYNAMICAL SYSTEMS

In this section we discuss how iterative algorithms in exact and finite precision can be represented as dynamical systems. This section follows the discussion in Hasan et al. (2010) and is included here for completeness.

Consider an iterative algorithm in exact precision and a dynamical system $x_{k+1} = f(x_k)$, where $x_k \in \mathbb{R}^n$ is the state vector. If, for some initial state $x_0 = \xi$, the state vector x_k of the dynamical system $x_{k+1} = f(x_k)$ is equal to the k^{th} iterate of the algorithm for all k, then we will say that the dynamical system represents the algorithm in exact precision when the initial state is $x_0 = \xi$. The 'k^{th} iterate' refers to the approximate solution obtained at iteration k of an algorithm.

To represent algorithms in finite precision as dynamical systems, we also have to consider finite precision errors in the dynamical model. One way of achieving this is to consider the finite precision errors as disturbance inputs w_k in the dynamical model.

Example 1. Consider the classical iterative methods for solving a system of linear equations $Ax^* = b$ (Golub and Van Loan, 1996, Sec. 10.1). The algorithm in exact precision can be represented by the following dynamics and initial state:

$$x_0 := 0, \qquad x_{k+1} := f(x_k) := M^{-1}Nx_k + M^{-1}b,$$

where $M \in \mathbb{R}^{n\times n}$ and $N \in \mathbb{R}^{n\times n}$ are matrices such that M is invertible and $A = M - N$.

The errors in the calculation of single iterations of a classical iterative method in finite precision floating point can be considered as additive errors (Higham, 2002, Sec. 17.2). Therefore, the classical iterative methods in finite precision can be represented by

$$x_0 := 0, \qquad x_{k+1} := f(x_k, w_k) := M^{-1}Nx_k + M^{-1}b + w_k,$$

where the disturbance w_k in the dynamical system represents the finite precision errors in a single iteration.

3. FORWARD ERROR ANLAYSIS USING MILLER'S WORK

In this section we propose a new forward error analysis scheme for iterative numerical algorithms based on the results given by Miller et al. (1988).

3.1 Miller's work

Miller et al. (1988) have used a Lyapunov-based approach to find ultimate bounds on solutions of perturbed linear control systems, where the perturbation arises due to quantization. They have considered a discrete-time linear system

$$x_{k+1} = Gx_k \tag{1}$$

and a nonlinear and possibly discontinuous perturbation of (1) given by

$$x_{k+1} = Gx_k + p(x_k) + \alpha_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$, $G \in \mathbb{R}^{n\times n}$, $p : \mathbb{R}^n \to \mathbb{R}^n$ and $\alpha_k \in \mathbb{R}^n$ for $k = 1, 2, \cdots$. By comparing (1) and (2) we note that the perturbation is assumed to be additive and is represented by $(p(x_k) + \alpha_k)$. The term α_k represents the part of the perturbation that approaches zero as the time index k approaches infinity and the term $p(x_k)$ represents any perturbation that depends on the state and is possibly non-zero at the origin.

The following theorem, which is given by Miller et al. (1988), can be used to find error bounds for iterative algorithms:

Theorem 1. Suppose that the magnitude of all the eigenvalues of matrix G of system (1) are less than one, $\alpha_k \to 0$ as $k \to \infty$ and that p satisfies $\|p(x_k)\|_2 \le \epsilon_1 \|x_k\|_2 + \epsilon_2$ for some small positive constants ϵ_1 and ϵ_2. If ϵ_1 is sufficiently small, then there exists a ψ (depending only on G, ϵ_1 and ϵ_2) such that system (2) is uniformly ultimately bounded with bound $\psi\epsilon_2$.

The definition of a uniformly ultimately bounded system given in Section 6.1 of Michel et al. (2008) is

Definition 1. A nonlinear system $x_{k+1} = f(x_k, k)$ is said to be uniformly ultimately bounded with bound β if for any $a > 0$ and for every $K \in \mathbb{Z}_{\ge 0}$, there exists an $l(a) > 0$, independent of K, such that $\|x_{k+K}\| \le \beta$ for all $\|x_K\| \le a$ and $k \ge l(a)$, where $\|\cdot\|$ denotes any norm on $\mathbb{R}^n$.

By an analysis similar to the one given in Miller et al. (1988), it can be shown that a possible value for ψ is

$$\psi = \frac{\rho_1\rho_2}{1-\gamma_1}, \tag{3}$$

where $\gamma_1 := (\gamma + \rho_1\rho_2\epsilon_1)$, γ is a constant such that $0 < \gamma < 1$ and such that for a Lyapunov function $V(\cdot)$ of system (1) it holds that $V(x_{k+1}) - V(x_k) \le (\gamma - 1)V(x_k)$ for all x_k, ρ_1 is a constant such that $V(x_k) \le \rho_1 \|x_k\|_2$ for all x_k, ρ_2 is a constant such that $\|x_k\|_2 \le \rho_2 V(x_k)$ for all x_k and ϵ_1 is sufficiently small such that $\gamma_1 < 1$. Therefore, by Theorem 1 we have the following bound:

$$\lim_{k\to\infty} \|x_k\|_2 \le \frac{\rho_1\rho_2}{1-\gamma_1}\epsilon_2. \tag{4}$$

Although the theorem by Miller et al. (1988) is for the ultimate bound, the analysis given in Miller et al. (1988) can be used to find a bound on the norm of the state for all $k \ge 0$, which is as follows:

$$\|x_k\|_2 \le \gamma_1^k V(x_0) + \frac{1-\gamma_1^k}{1-\gamma_1}\rho_1\rho_2\epsilon_2 + \sum_{j=0}^{k} \gamma_1^{k-j} \|\alpha_j\|_2. \tag{5}$$

3.2 A forward error analysis scheme for iterative algorithms

Let us define the solution error e_k as the difference between the k^{th} iterate x_k of an algorithm and the solution x^* of a numerical problem, i.e. $e_k := x_k - x^*$. If we can represent an iterative algorithm in exact precision as a dynamical system

$$x_{k+1} := f(x_k), \tag{6}$$

or in finite precision as

$$x_{k+1} := f(x_k, w_k), \tag{7}$$

where w_k represents the finite precision error, then by subtracting x^* from both sides of (6) or (7) we can also find the dynamics of the solution error e_k

$$e_{k+1} := f_e(e_k) := f(e_k + x^*) - x^*, \tag{8}$$

or

$$e_{k+1} := f_e(e_k, w_k) := f(e_k + x^*, w_k) - x^*. \tag{9}$$

The forward error for a numerical problem and an algorithm is defined as the norm or the component-wise absolute value of the solution error e_k. The process of bounding the forward error is called the forward error analysis. If systems (8)–(9) have the same structure as systems (1)–(2), then we can apply Theorem 1 and (5) to find a bound on the norm of e_k. Therefore, we can use the work of Miller et al. (1988) to have a systematic way of forward error analysis of algorithms.

Although Miller et al. (1988) have stated the theorem for linear systems, the basic tool of Lyapunov functions that they have utilized is also applicable for non linear systems. Hence, it is straightforward to extend their technique to algorithms that can only be represented by non linear dynamical systems.

4. FORWARD ERROR ANALYSIS USING KOFMAN'S WORK

In this section we briefly review the results given by Kofman et al. (2007) on ultimate boundedness of perturbed systems and present a new scheme for forward error analysis based on these results.

4.1 Kofman's work

Kofman et al. (2007) have given a method to obtain ultimate bounds for linear discrete-time systems with state-dependent perturbations that do not disappear as the state approaches the equilibrium point. These kind of perturbations arise in quantized systems. The following theorem stated by Kofman et al. (2007) can be used to find forward error bounds for iterative algorithms:

Theorem 2. Consider the system

$$x_{k+1} = Gx_k + u_k, \tag{10}$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^n$ and $G \in \mathbb{R}^{n\times n}$ has all its eigenvalues strictly inside the unit circle and Jordan canonical form

$$\Lambda = W^{-1}GW.$$

Suppose that $|u_k| \le \delta(|x_k|)$ for all $k \ge 0$, where $\delta : \mathbb{R}^n_{\ge 0} \to \mathbb{R}^n_{\ge 0}$ is a continuous map satisfying

$$|z_1| \le |z_2| \Rightarrow \delta(|z_1|) \le \delta(|z_2|), \tag{11}$$

for all $z_1 \in \mathbb{R}^n$ and $z_2 \in \mathbb{R}^n$.

Consider a map $T : \mathbb{R}^n_{\ge 0} \to \mathbb{R}^n_{\ge 0}$ defined by

$$T(y) := |\Lambda|\, y + \left|W^{-1}\right| \delta(|W|\, y). \tag{12}$$

Suppose that a point $\zeta \in \mathbb{R}^n$ satisfying $\zeta = T(\zeta)$ exists. Let $\tilde{x} \in \mathbb{R}^n$ denote any point satisfying $\lim_{k\to\infty} T^k(\left|W^{-1}\tilde{x}\right|) = \zeta$ (note that $W\zeta$ is one such point). If the initial condition x_0 satisfies

$$\left|W^{-1}x_0\right| \le \left|W^{-1}\tilde{x}\right|, \tag{13}$$

then for any $\epsilon \in \mathbb{R}^n_{>0}$ there exists an $l = l(\epsilon, \tilde{x})$ such that for all $k \ge l$

$$|x_k| \le |W|\,\zeta + |W|\,\epsilon.$$

We note that although Theorem 2 states an important result about ultimate boundedness, Kofman et al. (2007) do not discuss how $l = l(\epsilon, \tilde{x})$ or ϵ can be calculated. Therefore, we can only calculate a bound for $k \to \infty$, in which case $\epsilon = 0$ and the bound becomes

$$\lim_{k\to\infty} |x_k| \le |W|\,\zeta.$$

4.2 A forward error analysis scheme for iterative algorithms

Following the discussion in Section 3.2, if (9) has the same structure as (10), then we can apply Theorem 2 to find a bound on the forward error as $k \to \infty$. Therefore, we have a systematic way for forward error analysis of iterative algorithms.

An advantage of this technique is that it gives us bounds on the component-wise absolute value of the error, which is desirable, because bounds on the 2-norm and ∞-norm can readily be calculated from it.

We note that due to the requirement of Jordan canonical form the work by Kofman et al. (2007) relies on the linearity of the system. Therefore, we can only use this approach for forward analysis of algorithms that can be represented by a linear dynamical system. Kofman et al. (2007) have also presented a way of extending their approach to non linear systems. They propose to linearize the system at the equilibrium point and consider the nonlinearity as part of the perturbation. However, this method could result in conservative bounds.

Inequality (13) imposes some requirements on the initial condition. This means that the bound is only guaranteed if we know that the initial error will lie in a certain set.

As noted earlier, Kofman et al. (2007) have not discussed how to calculate l or ϵ. Therefore, we can only find a bound on the forward error for $k \to \infty$. This is not desirable, because in practice only a finite number of iterations of an algorithm are computed. Moreover, having a bound on e_k for all k can help us to comment about the rate of convergence of an algorithm.

5. CLASSICAL ITERATIVE METHODS

In this section we will apply the proposed forward error analysis schemes on the classical iterative methods (Golub and Van Loan, 1996, Sec. 10.1) for solving a system of linear equations $Ax^* = b$. The iterations of the classical iterative methods are as follows:

$$x_0 = 0, \quad x_{k+1} = M^{-1}(Nx_k + b), \tag{14}$$

where $M \in \mathbb{R}^{n\times n}$ and $N \in \mathbb{R}^{n\times n}$ are matrices such that $A = M - N$. Table 1 lists some of the classical iterative methods (Golub and Van Loan, 1996, Sec. 10.1). In the table, D is the diagonal of the matrix A, L is the strictly lower triangle part of A and ω is a scalar. If the convergence conditions listed in the table are satisfied, then matrices A and M are invertible, the magnitudes of all the eigenvalues of the matrix $M^{-1}N$ are less than one and the algorithm converges to the solution $x^* = A^{-1}b$. In the sequel we will assume that the convergence conditions are satisfied.

Table 1. Classical iterative methods

Matrix M	**Method**	**Convergence conditions**
D	Jacobi	A is SDD
$(D+L)$	Gauss-Seidel	A is SDD or SPD
$(\omega^{-1}D+L)$	Successive over-relaxation	A is SPD and $0<\omega<2$

SDD = strictly diagonally dominant; SPD = symmetric positive definite

5.1 Finite precision error analysis for a single iteration

If the algorithm is implemented in floating point, then there are finite precision errors in the calculations in each iteration of the algorithm. These finite precision errors can be represented as an additive error (Higham, 2002, Sec. 17.2) and we have the expression

$$\begin{aligned} \mathrm{fl}(M^{-1}(Nx_k+b)) &= M^{-1}\otimes(N\otimes x_k\oplus b) && (15)\\ &= M^{-1}(Nx_k+b)+w_k, && (16)\end{aligned}$$

where operator $\mathrm{fl}(\cdot)$ represents calculation in floating point, w_k represents the finite precision errors in the whole calculation and $\otimes$ and $\oplus$ represent multiplication and addition in floating point, respectively. Floating point numbers in base 2, which is the most common one, are of the form

$$\pm m\times 2^{\eta-t},$$

where $m\in\mathbb{Z}$ is the significand, $t\in\mathbb{Z}$ is the precision and $\eta\in\mathbb{Z}$ is the exponent. The unit roundoff $\mu:=2^{-t}$ is the maximum relative error in approximating a real number (within the range of floating point representation) by a floating point number. For a tutorial on floating point number representation and arithmetic we refer the reader to Chapters 2 and 3 of Higham (2002).

We can think of the finite precision errors w_k as being produced because of two reasons: the calculation of (Nx_k+b) and the solution of a linear system of equations with matrix M. Consider a system of linear equations $Mz=y$, where $M\in\mathbb{R}^{n\times n}$ is a lower triangular matrix (which is the case for classical iterative methods), $z\in\mathbb{R}^n$ and $y\in\mathbb{R}^n$. When this system is solved in finite precision floating point using the back substitution algorithm (Golub and Van Loan, 1996, Sec. 3.1.2), then the calculated approximate solution $\hat{z}$ is an exact solution of a system of linear equations $(M+\Delta M)\hat{z}=y$, where $\Delta M\in\mathbb{R}^{n\times n}$ and

$$|\Delta M|\le n\mu\,|M|\,, \qquad (17)$$

as shown on page 101 of Wilkinson (1963). Therefore, the following relation will hold in finite precision for iterations of classical iterative methods:

$$\begin{aligned} &(M+\Delta M_{k+1})x_{k+1}=\mathrm{fl}(Nx_k+b)\\ &\Leftrightarrow x_{k+1}=M^{-1}\left(\mathrm{fl}\left(Nx_k+b\right)-\Delta M_{k+1}x_{k+1}\right). \qquad (18)\end{aligned}$$

According to Section 2.4 in Golub and Van Loan (1996), we have the following bounds:

$$\begin{aligned} |\mathrm{fl}(Qz)-Qz| &\le 1.01n\mu\,|Q|\,|z|\,;\ n\mu<0.01, && (19a)\\ |\mathrm{fl}(z+y)-(z+y)| &\le (|z|+|y|)\mu, && (19b)\end{aligned}$$

where $Q\in\mathbb{R}^{n\times n}$, $z\in\mathbb{R}^n$, $y\in\mathbb{R}^n$ and the inequalities hold component wise.

Using (17)–(19) we can find the following bounds:

$$\begin{aligned} |w_k| \le\ & (I-n\mu\,|M|)^{-1}\times\\ & \left[\left(c_1\left|M^{-1}\right||N|+n\mu\,|M|\left|M^{-1}\right||N|\right)|x_k|+\right.\\ & \left.\left(\mu\left|M^{-1}\right|+n\mu\,|M|\left|M^{-1}\right|\right)|b|\right], \qquad (20)\end{aligned}$$

$$\begin{aligned} \|w_k\|_\infty \le\ & c_1\left\|M^{-1}\right\|_\infty\|N\|_\infty\|x_k\|_\infty+\mu\,\|b\|_\infty\\ & +n\mu\gamma_x\left\|M^{-1}\right\|_\infty\|M\|_\infty\|x^*\|_\infty\,, \qquad (21)\end{aligned}$$

where $c_1:=1.01n\mu^2+1.01n\mu+\mu$ and $\gamma_x:=\sup_k\frac{\|x_k\|_\infty}{\|x^*\|_\infty}$. Note that due to the condition in (19), these bounds only hold for the case $n\mu<0.01$.

5.2 Forward error analysis of classical iterative methods by Miller's technique

In this section we will show how Theorem 1 can be used to find forward error bounds for the classical iterative methods. Following the discussion in Section 3.2 we can write the dynamics of the error e_k for classical iterative methods in exact precision and finite precision as

$$e_{k+1}:=M^{-1}Ne_k \qquad (22)$$

and

$$e_{k+1}:=M^{-1}Ne_k+w_k, \qquad (23)$$

respectively.

Comparing (22)–(23) with (1)–(2), we note that matrix G in (1) is equivalent to $M^{-1}N$ and the additive perturbation is w_k. if the convergence conditions in Table 1 are satisfied, then the magnitude of all the eigenvalues of $M^{-1}N$ are less than one as required in Theorem 1. The perturbation w_k does not have a part that approaches zero as k approaches infinity, therefore we take $\alpha_k\equiv 0$ and $p(e_k)\equiv w_k$. Using (21) we get

$$\|w_k\|_2\le\epsilon_1\|e_k\|_2+\epsilon_2,$$

where

$$\begin{aligned} \epsilon_1 &:= c_1 n^{\frac{1}{2}}\left\|M^{-1}\right\|_\infty\|N\|_\infty\,, && (24)\\ \epsilon_2 &:= c_1 n^{\frac{1}{2}}\left\|M^{-1}\right\|_\infty\|N\|_\infty\|x^*\|_\infty+\\ &\quad n^{\frac{3}{2}}\mu\gamma_x\left\|M^{-1}\right\|_\infty\|M\|_\infty)\,\|x^*\|_\infty+n^{\frac{1}{2}}\mu\,\|b\|_\infty\,. && (25)\end{aligned}$$

From standard Lyapunov theory we know that for a stable discrete-time linear system a possible Lyapunov function is

$$V(e_k)=e_k^TPe_k,$$

where T in superscript denotes transposition and $P\in\mathbb{R}^{n\times n}$ is the solution of the discrete-time Lyapunov equation

$$\left(M^{-1}N\right)^TP\left(M^{-1}N\right)-P=-Q,$$

for some positive definite $Q\in\mathbb{R}^{n\times n}$ (Vidyasagar, 2002, Sec. 5.9). For this choice of Lyapunov function we have

$$\rho_1\le\lambda_{\max}(P),\quad \rho_2\le\frac{1}{\lambda_{\min}(P)},\quad \gamma\le 1-\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)},$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalue of a symmetric matrix, respectively.

If

$$\gamma_1:=(\gamma+\rho_1\rho_2\epsilon_1)=1-\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}+\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}\epsilon_1<1,$$

then using Theorem 1 we have

$$\lim_{k\to\infty} \|e_k\| \leq \frac{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}{\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)} - \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}\epsilon_1}\epsilon_2, \qquad (26)$$

where ϵ_1 and ϵ_2 are given in (24) and (25), respectively.

5.3 Forward error analysis of classical iterative methods by Kofman's technique

We note that (23) has the same form as (10) with $G := M^{-1}N$ and $u_k := w_k$. If the convergence conditions in Table 1 are satisfied, then the magnitude of all the eigenvalues of $M^{-1}N$ are less than one as required in Theorem 2. Using (20) we get

$$|w_k| \leq D\,|e_k| + d,$$

where

$$D := (I - n\mu\,|M|)^{-1}\,(c_1 + n\mu\,|M|)\left(\left|M^{-1}\right|\,|N|\right), \qquad (27)$$

$$d := (I - n\mu\,|M|)^{-1}\left[\left(\mu\left|M^{-1}\right| + n\mu\,|M|\left|M^{-1}\right|\right)|b| + \left(c_1\left|M^{-1}\right|\,|N| + n\mu\,|M|\left|M^{-1}\right|\,|N|\right)|x^*|\right]. \qquad (28)$$

The inverse of $(I - n\mu\,|M|)$ in the above equations will exist if the unit roundoff μ is sufficiently small; see Appendix A. A map $\delta : \mathbb{R}^n_{\geq 0} \to\in \mathbb{R}^n_{\geq 0}$ which satisfies property (11) can be chosen as $\delta(z) := Dz + d$, where $z \in \mathbb{R}^n_{\geq 0}$.

The map T defined in (12) is given as

$$\begin{aligned} T(y) &:= |\Lambda|\,y + \left|W^{-1}\right|\delta(|W|\,y) \\ &= |\Lambda|\,y + \left|W^{-1}\right|(D\,|W|\,y + d), \qquad (29) \end{aligned}$$

where Λ is the Jordan canonical form of $M^{-1}N$, i.e. $\Lambda := W^{-1}M^{-1}NW$.

A point $\zeta \in \mathbb{R}^n$ satisfying $\zeta = T(\zeta)$ can be computed as

$$\zeta := (I - \Lambda - \left|W^{-1}\right| D\,|W|)^{-1}\left|W^{-1}\right| d \qquad (30)$$

if the inverse of matrix $(I - \Lambda - \left|W^{-1}\right| D\,|W|)$ exists, which will be the case for a sufficiently small μ; see Appendix A.

Using Theorem 2 an expression for the bound on the forward error is as follows:

$$\lim_{k\to\infty} |e_k| \leq |W|\,(I - \Lambda - \left|W^{-1}\right| D\,|W|)^{-1}\left|W^{-1}\right| d, (31)$$

where D and d are given by (27) and (28), respectively.

5.4 Forward error analysis of classical iterative methods in numerical analysis literature

A forward error bound for classical iterative methods is given in Section 17.2 of Higham (2002). The expression for the bound is

$$\lim_{k\to\infty} \|e_k\|_\infty \leq c_2\mu(1 + \gamma_x)(\|M\|_\infty + \|N\|_\infty) \times \|x^*\|_\infty \frac{\left\|M^{-1}\right\|_\infty}{1 - \|M^{-1}N\|_\infty} \qquad (32)$$

if $\left\|M^{-1}N\right\|_\infty < 1$, where c_2 is a constant depending on n. In Higham (2002), the constant c_2 is defined such that the following holds:

$$|\xi_k| \leq c_2\mu(|M|\,|x_{k+1}| + |N|\,|x_k| + |b|),$$

for all k, where

$$\xi_k := \Delta M_{k+1}x_{k+1} + (Nx_k + b) - \mathrm{fl}(Nx_k + b),$$

see (17.2) in Higham (2002). According to an analysis similar to the one in Section 5.1, it can be shown that a possible value of c_2 is $(1.01n + 1)$ if $n\mu < 0.01$.

5.5 Comparison

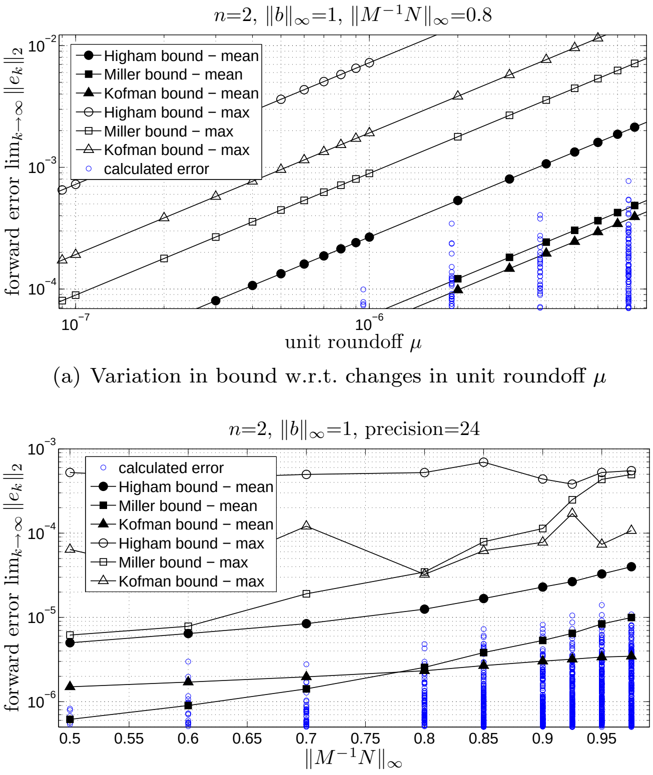
In this section we compare the bound obtained using Miller's technique (26), the bound obtained using Kofman's technique (31) and the bound given by Higham (32). In contrast to Higham's bound (32), the other two bounds do not require that the ∞-norm of matrix $M^{-1}N$ should be less than one. Regarding the matrix $M^{-1}N$, the only requirement for the control-theoretic bounds is that the magnitude of all the eigenvalues are less than one, which is guaranteed if the convergence conditions in Table 1 are satisfied.

The bound (31) requires that the inverse of matrix $(I - n\mu\,|M|)$ in (27)–(28) and matrix $(I - \Lambda - \left|W^{-1}\right| D\,|W|)$ in (30) exist. This could be a restrictive condition for low machine precisions or large values of unit roundoff.

Figure 1 shows a numerical comparison of the bounds. To calculate the bound obtained using Miller's technique (26) we need a matrix Q. Although any positive definite matrix can be chosen as Q, we would like to select a value that gives a lower bound. To calculate the bound we select Q by random search over one hundred random values. The random values of Q are generated by setting $Q = R^TR$, where R is a non-singular matrix with each element as an independent random number with a standard Gaussian distribution, i.e. with zero mean and unit variance.

The expressions for all of the bounds also depend on the solution of the problem. Therefore, we have calculated the error bounds using the exact solution of random problems and plotted the maximum and mean of the bounds obtained for these problems. The random problems were generated such that each element of vector b and each off-diagonal element of matrix A is an independent random number with standard Gaussian distribution. Moreover, the diagonal elements of A are the absolute sum of the off-diagonal elements in the corresponding row plus a small positive constant, so that the matrix A is strictly diagonally dominant, which ensures convergence for the Gauss-Seidel algorithm. In the figure we have also shown the errors in the calculated solution of these random problems to get some idea of the conservativeness of these bounds. These random problems were solved using the Gauss-Seidel algorithm and the calculated errors were obtained by simulation at various precisions using the Multiple Precision Toolbox by Barrowes (2010).

In the figure we observe that the bounds using control-theoretic ideas are tighter than Higham's bound. Moreover, Figure 1(b) shows that for higher values of $\|M^{-1}N\|_\infty$ the bound by Kofman's technique is the lowest, whereas for lower values of $\|M^{-1}N\|_\infty$ the bound obtained by Miller's technique is better. Although Higham's expression gives the most conservative bounds, we note that the simplicity of the expression (32) allows us to easily comment on the



(a) Variation in bound w.r.t. changes in unit roundoff μ

(b) Variation in bound w.r.t. changes in $\left\|M^{-1}N\right\|_\infty$

Fig. 1. Comparison of forward error bounds for classical iterative methods

dependence of the bound on problem size n and machine precision μ. The discussion suggests that each bound has its own advantages and it is useful to consider more than one forward error analysis approach.

6. CONCLUSION

In this paper we have shown that some problems in the literature on quantization effects in control systems are similar to the problems in numerical analysis. We have used results from the study of quantization effects to propose two new systematic ways for forward error analysis of iterative algorithms. The proposed schemes are applied on the classical iterative methods and the obtained bounds are shown to be tighter than the bounds in the numerical analysis literature. We have also discussed the limitations and advantages of each forward error analysis scheme and argued that it might be useful to consider more than one forward error analysis scheme for an algorithm.

REFERENCES

Barrowes, B. (2010). Multiple Precision Toolbox. `http://www. sondette.com/math/mp_toolbox.html`.

Bhaya, A. and Kaszkurewicz, E. (2007). A control-theoretic approach to the design of zero finding numerical methods. *IEEE Transactions on Automatic Control*, 52(6), 1014–1026.

Chehab, J.P. and Laminie, J. (2005). Differential equations and solution of linear systems. *Numerical Algorithms*, 40(2), 103–124.

Chu, M.T. (2008). Linear algebra algorithms as dynamical systems. *Acta Numerica*, 17, 1–87.

Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations.* The Johns Hopkins University Press.

Haimovich, H. (2006). *Quantisation Issues in Feedback Control.* Ph.D. thesis, The University of Newcastle.

Hasan, A., Kerrigan, E.C., and Constantinides, G.A. (2010). An ISS and l-stability approach to forward error analysis of iterative numerical algorithms. Submitted to the 49th IEEE Conference on Decision and Control.

Higham, N.J. (2002). *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics.

Hurt, J. (1967). Some stability theorems for ordinary difference equations. *SIAM Journal on Numerical Analysis*, 4(4), 582–596.

Kashima, K. and Yamamoto, Y. (2007). System theory for numerical analysis. *Automatica*, 43(7), 1156–1164.

Kofman, E., Haimovich, H., and Seron, M.M. (2007). A systematic method to obtain ultimate bounds for perturbed systems. *International Journal of Control*, 80(2), 167–178.

Michel, A.N., Hou, L., and Liu, D. (2008). *Stability of dynamical systems: continuous, discontinuous, and discrete systems.* Birkhäuser.

Miller, R.K., Mousa, M.S., and Michel, A.N. (1988). Quantization and overflow effects in digital implementations of linear dynamic controllers. *IEEE Transactions on Automatic Control*, 33(7), 698–704.

Ortega, J.M. (1973). Stability of difference equations and convergence of iterative processes. *SIAM Journal on Numerical Analysis*, 10(2), 268–282.

Vidyasagar, M. (2002). *Non-linear Systems Analysis.* Society for Industrial and Applied Mathematics.

Wilkinson, J.H. (1963). *Rounding Errors in Algebraic Processes.* Prentice-Hall, Inc.

Zhou, K., Doyle, J.C., and Glover, K. (1995). *Robust and optimal control.* Prentice Hall.

Appendix A. EXISTENCE OF INVERSE

The inverse of matrix $(I - n\mu|M|)$ will exist if all the singular values of the matrix are non-zero. For square matrices A and B we have

$$\underline{\sigma}(A+B) \geq \underline{\sigma}(A) - \bar{\sigma}(B),$$

where $\underline{\sigma}(\cdot)$ and $\bar{\sigma}(\cdot)$ denote the smallest and the largest singular value of a matrix, respectively (Zhou et al., 1995, Lemma 2.12). For matrix $(I - n\mu|M|)$ we have

$$\begin{aligned}\underline{\sigma}\left(I + (-n\mu|M|)\right) &\geq \underline{\sigma}(I) - \bar{\sigma}(-n\mu|M|),\\ &= 1 - n\mu\bar{\sigma}(|M|),\end{aligned}$$

where we have used the fact that all the singular values of the identity matrix I are equal to one. Therefore, the smallest singular value of the matrix $(I - n\mu|M|)$ will be greater than zero (or the matrix will be invertible) if

$$n\mu\bar{\sigma}(|M|) < 1.$$

There will always exist a $\mu > 0$ sufficiently small such that this condition is true.

Similar arguments can be used to show that matrix $(I - \Lambda - \left|W^{-1}\right| D|W|)$ is also invertible if μ is sufficiently small.