An ISS and l-stability Approach to Forward Error Analysis of Iterative Numerical Algorithms

Ammar Hasan, Eric C. Kerrigan, George A. Constantinides

I. Introduction

The idea that iterative algorithms can be treated as dynamical systems and that they can be studied using control theory is not new. Hurt [1] showed that La Salle theorems for differential equations could be used to find the region of convergence for the Newton-Raphson and secant methods, both of which are iterative methods. Several connections between Lyapunov stability theory for ordinary differential equations and convergence of iterative methods were reviewed by Ortega [2]. However, little work can be found in the literature on the study of algorithms using control-theoretic ideas.

In recent years, some authors have shown interest in this research area [3]–[8]. However, most of the authors have focused on the design and analysis of algorithms in *exact* precision. It is well-known that control theory provides many tools to study the effects of disturbances. In this paper we will use control-theoretic ideas to study the effects of *finite* precision in algorithms. We will start with how algorithms in finite precision can be represented as dynamical systems and present a new systematic way of finding bounds on finite precision errors and the forward error for an algorithm; the forward error is the norm of the difference between the approximate solution obtained by an algorithm and the exact solution of the numerical problem. The proposed scheme will be applied on the successive iteration method and the classical iterative methods. The advantage of the presented scheme is demonstrated by comparing the obtained bounds with bounds given in the numerical analysis literature.

In [4], [5] authors have also studied the effects of finite precision errors using control theory. Kashima and Yamamoto [4] have shown that if the finite precision errors in iterations of Newton's method are bounded, then the forward error will also be bounded. However, they have not given a bound for either the finite precision errors or the forward error. Diene and Bhaya [5] have looked at the robustness of algorithms in finite precision. They have proposed the use of control Lyapunov functions to find a bound on the maximum finite precision error that would guarantee the algorithm remains stable. In contrast, in this paper we are concerned with bounds on the difference between the exact solution of a problem and the solution obtained by an algorithm in finite precision.

In Section II we give the primary notations and definitions. Section III discusses how algorithms can be represented as dynamical systems. Section IV presents two propositions for finding the error bounds for algorithms in finite precision. These propositions are applied in Sections V and VI to obtain bounds for the successive iterative method and the classical iterative methods, respectively; we also compare the obtained forward error bounds with bounds given in the numerical analysis literature. The concluding remarks are given in Section VII.

II. Notations and definitions

We denote matrices with capital letters; vectors, scalars and functions with small letters and sequences with small and bold letters. We denote the vector Euclidean norm and the matrix spectral norm (or the induced matrix 2-norm) with $\|\cdot\|_2$. We denote the vector infinity norm, sequence l_∞-norm and the induced matrix infinity norm with $\|\cdot\|_\infty$. We use $|\cdot|$ to denote the component-wise absolute value of a matrix or a vector.

We will often mention 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 $b \in \mathbb{R}^n$. Other notation will be defined as we require it.

A function $\zeta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a $\mathcal{K}$-function if it is continuous, strictly increasing and $\zeta(0) = 0$. A function $\zeta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a $\mathcal{KL}$-function if, for each fixed $t \geq 0$, the function $\zeta(\cdot, t)$ is a $\mathcal{K}$-function, and for each $s \geq 0$, the function $\zeta(s, \cdot)$ is decreasing and $\zeta(s, t) \to 0$ as $t \to \infty$.

III. Algorithms as dynamical systems

In this section we discuss how iterative algorithms can be represented as dynamical systems. Consider an iterative algorithm and a dynamical system with state vector x_k and dynamics $x_{k+1} = f(x_k)$. If, for some initial state $x_0 = \xi$, the state x_k of the dynamical system 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 state space form when the initial state is $x_0 = \xi$. By the 'k^{th} iterate' we mean the approximate solution obtained at iteration k of an algorithm.

Example 1: Consider the Jacobi algorithm [9, Sec. 10.1.1] for solving a linear system of equations $Ax^* = b$ in exact precision. We can represent the algorithm with the following dynamics and initial state:

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

A. Hasan and G. Constantinides are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, U.K. {ammar.hasan07,g.constantinides}@ic.ac.uk

E. Kerrigan is with the Department of Aeronautics and the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, U.K. e.kerrigan@ic.ac.uk

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

where M and N are matrices such that M is the diagonal of the matrix A and $A = M - N$. ◄

To represent an algorithm in finite precision, we also have to add finite precision errors in the state space model. This could be done by considering the finite precision errors as disturbance inputs w_k in the dynamical model. We illustrate this with the help of an example.

Example 2: Consider the Jacobi algorithm [9] for solving a linear system of equations $Ax^* = b$. The errors in the calculation of a Jacobi iteration in finite precision floating point arithmetic can be considered as additive errors [10, Sec. 17.2]. Therefore, the Jacobi algorithm 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 is equal to the finite precision floating point errors. ◄

We can also represent an algorithm in finite precision using an input-output dynamical model. We give an example to illustrate this.

Example 3: Consider the Jacobi algorithm [9] for solving a linear system of equations $Ax^* = b$. The algorithm in finite precision can be represented by the system in Figure 1, where P is the input-output map of dynamics $x_{k+1} = f(x_k, w_k) = M^{-1}Nx_k + M^{-1}b + w_k$ with initial condition $x_0 = 0$, w_k is the finite precision error at iteration k and E is the input-output map between the iterates of the algorithm and the finite precision errors. ◄

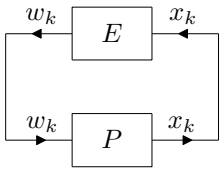


Fig. 1. Algorithm in finite precision as a dynamical system using I/O maps

IV. Bounds using control theory

In this section we give two propositions that together provide a systematic way of finding error bounds on solutions obtained from algorithms. The first proposition helps us to find bounds on the maximum finite precision error and the maximum norm of the iterates of an algorithm. The bound on maximum finite precision error is also required in the next proposition to obtain an error bound on the solution of an algorithm.

Proposition 1: Assume that the algorithm in finite precision can be considered as a dynamical system with input-output maps as in Figure 1. If the following hold:

1) System P is finite gain l_∞-stable [11, Sec. 6.7.1], i.e.
$$\|\mathbf{x}\|_\infty \leq \gamma_1 \|\mathbf{w}\|_\infty + \beta_1$$
for all input sequences $\mathbf{w}$, where $\mathbf{x} := (x_0, x_1, x_2, \ldots)$, $\mathbf{w} := (w_0, w_1, w_2, \ldots)$ and γ_1 and β_1 are positive scalars;
2) System E is finite gain l_∞-stable, i.e.
$$\|\mathbf{w}\|_\infty \leq \gamma_2 \|\mathbf{x}\|_\infty + \beta_2$$
for all sequences $\mathbf{x}$, where γ_2 and β_2 are positive scalars;
3) The small gain condition is satisfied, i.e. $\gamma_1\gamma_2 < 1$,

then

1) The output of system P, x_k, is bounded and we have the following relation:
$$\|\mathbf{x}\|_\infty \leq \frac{1}{(1-\gamma_1\gamma_2)}(\gamma_1\beta_2 + \beta_1);$$
2) The output of system E, w_k, which is equal to the finite precision errors is bounded and we have the following relation:
$$\|\mathbf{w}\|_\infty \leq \frac{1}{(1-\gamma_1\gamma_2)}(\gamma_2\beta_1 + \beta_2).$$

Proof: The relationships simply follow by applying the small gain theorem [11, Sec. 6.7.3]. ■

Given a bound on the maximum finite precision error, the next proposition helps us to find bounds on the forward error. Before stating the proposition we give a definition of input-to-state stability, which follows the definition given in [12].

Definition 1: A system $x_{k+1} = f(x_k, w_k)$, with equilibrium at the origin, is input-to-state stable (ISS) if there is a $\mathcal{KL}$-function ζ_1 and a $\mathcal{K}$-function ζ_2 such that, for each input sequence $\mathbf{w} := (w_0, w_1, w_2, \ldots)$ belonging to l_∞ and each initial state ξ, it holds that

$$\|x_k\| \leq \zeta_1(\|\xi\|, k) + \zeta_2(\|\mathbf{w}\|_\infty)$$

for each k, where $\|\mathbf{w}\|_\infty := \sup_k \|w_k\|$ and $\|\cdot\|$ can denote any vector norm. ◄

Let us define the solution error as $e_k := x_k - x^*$, where x_k is the approximate solution at iteration k and x^* is the exact solution of a numerical problem. Given a dynamical system representation of an algorithm in state space form, $x_{k+1} := f(x_k, w_k)$, we can find the dynamics for the solution error as $e_{k+1} = f_e(e_k, w_k) := f(e_k + x^*, w_k) + x^*$.

Proposition 2: Assume that the algorithm in finite precision can be considered as a dynamical system in state space form with dynamics $x_{k+1} := f(x_k, w_k)$. Also assume that x^* is an equilibrium point of the dynamical system when there is no disturbance, i.e. $w_k = 0$, and x^* is equal to the solution of the numerical problem.

If the dynamical system for the solution error, $e_{k+1} = f_e(e_k, w_k)$, which has an equilibrium at the origin for $w_k = 0$, is ISS, then the forward error, which is the norm of the solution error, is bounded by an expression of the following type:

$$\|e_k\| \leq \zeta_1(\|e_0\|, k) + \zeta_2(\|\mathbf{w}\|_\infty)$$

for each k, where ζ_1 is a $\mathcal{KL}$-function and ζ_2 is a $\mathcal{K}$-function.

Proof: The proof follows by the ISS property. ■

Corollary 1: We have the following bound:

$$\lim_{k\to\infty} \|e_k\| \leq \zeta_2(\|\mathbf{w}\|_\infty).$$

The process of bounding the forward error is called forward error analysis. Therefore, the above-mentioned propositions provide a systematic way for forward error analysis of iterative numerical algorithms.

In the above, we have used the tools of ISS and l_∞-stability, both of which are applicable to nonlinear systems. Therefore, the proposed forward error analysis scheme can be applied to both linear and nonlinear algorithms. Moreover, we have not made any assumption about the number representation, i.e. fixed point, floating point, etc. The model of the dynamical system and bounds on finite precision errors might change with the number representation. However, the propositions would still be applicable as long as their hypotheses are satisfied.

V. Bounds for the successive iteration method

In this section we will use the propositions given in Section IV to find bounds for the successive iteration method [13]:

$$x_0 = 0, \quad x_{k+1} = Hx_k + h, \tag{1}$$

where $H \in \mathbb{R}^{n\times n}$ and $h \in \mathbb{R}^n$. If the magnitude of the eigenvalues of the matrix H are all less than one, then the algorithm converges to the solution $x^* = (I-H)^{-1}h$, where I is the identity matrix.

A. Finite precision error analysis of a single iteration

If the algorithm is implemented in floating point, then the finite precision errors in the calculations in each iteration of the algorithm can be represented as an additive error and we have the expression

$$\mathrm{fl}(Hx_k + h) := (H \otimes x_k) \oplus h = Hx_k + h + w_k,$$

where operator $\mathrm{fl}(\cdot)$ represents calculation in floating point, $\odot$ represents arithmetic operation in floating point and w_k represents the finite precision errors in the whole calculation. For a tutorial on floating point number representation and finite precision errors we refer the reader to [10, Chapters 2 and 3].

According to [9, Sec. 2.4], we have the following bounds:

$$|\mathrm{fl}(Qz) - Qz| \le 1.01n\mu\,|Q|\,|z|\,;\ n\mu < 0.01, \tag{2a}$$

$$|\mathrm{fl}(z+y) - (z+y)| \le (|z| + |y|)\mu, \tag{2b}$$

where $Q \in \mathbb{R}^{n\times n}$, $z \in \mathbb{R}^n$, $y \in \mathbb{R}^n$, μ is the machine unit roundoff and the inequalities hold component wise. Using the expressions given above, we find the following bound on w_k:

$$\|w_k\|_2 \le c_1 \|H\|_2 \|x_k\|_2 + \|h\|_2 \mu, \tag{3}$$

where $c_1 := 1.01n^{\frac{3}{2}}\mu^2 + 1.01n^{\frac{3}{2}}\mu + \mu$.

B. Bound on maximum finite precision error over all iterates

In this section we state a theorem that gives bounds on the maximum finite precision error and the maximum iterates in the successive iteration method (1).

Theorem 1: For $\|H\|_2 < 1$ and $n\mu < 0.01$ we have the following bounds for the successive iteration method (1):

$$\|\mathbf{x}\|_\infty \le \frac{1}{1 - \frac{c_1\|H\|_2}{1-\|H\|_2}} \left(\frac{\|h\|_2 \mu + \|h\|_2}{1 - \|H\|_2} \right), \tag{4a}$$

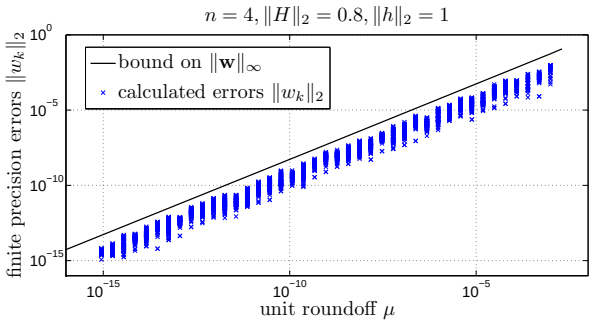


Fig. 2. Bound for maximum finite precision error over all iterations of the algorithm and the actual finite precision errors in single iterations of the algorithm for random problems

$$\|\mathbf{w}\|_\infty \le \frac{1}{1 - \frac{c_1\|H\|_2}{1-\|H\|_2}} \left(c_1 \frac{\|H\|_2 \|h\|_2}{1 - \|H\|_2} + \|h\|_2 \mu \right). \tag{4b}$$

Proof: See Appendix I. ∎

Figure 2 shows a plot of the obtained bound on maximum finite precision error in an iteration as a function of unit roundoff μ. The plotted bound is for problems with $\|H\|_2 \le 0.8$ and $\|h\|_2 \le 1$. The figure also shows calculated errors in iterations of the algorithm while solving random problems with $\|H\|_2 \le 0.8$ and $\|h\|_2 = 1$. The random problems were generated by generating a matrix H and a vector h for each random problem; each component of the matrix and the vector is an independent random number with standard Gaussian distribution, i.e. with zero mean and unit variance. In this paper, we have generated all random problems for the successive iteration method in this way. The calculated errors are obtained by simulation at various precisions using the Multiple Precision Toolbox [14]. In the sequel, all calculated errors are obtained in this way.

Due to space limitations we have only given a figure that shows the variation in bound with changes in unit roundoff μ and for specific values of $\|H\|_2$ and $\|h\|_2$. However, we will give a figure later that shows the variation in forward error bound with respect to changes in $\|H\|_2$. As for the variation of bound with changes in $\|h\|_2$, we observe from (4b) that we have a linear relation. Similar comments also apply to some of the other bounds we find in this paper.

C. Bound on forward error

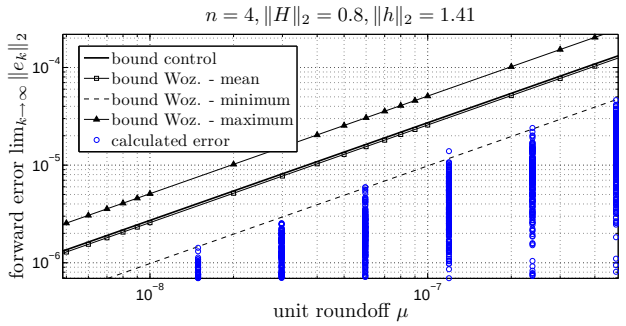
Theorem 2: For $\|H\|_2 < 1$ and $n\mu < 0.01$ we have the following forward error bound for the successive iteration method (1):

$$\|e_k\|_2 \le \left\|H^k\right\|_2 \|e_0\|_2 + \frac{1}{1-\|H\|_2} \frac{1}{1 - \frac{c_1\|H\|_2}{1-\|H\|_2}} \times \left(c_1 \frac{\|H\|_2 \|h\|_2}{1 - \|H\|_2} + \|h\|_2 \mu \right).$$

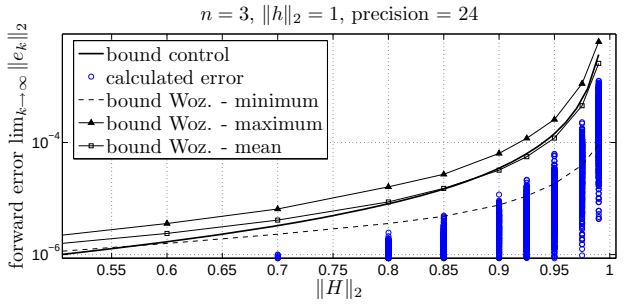
Proof: See Appendix II. ∎

Corollary 2: For $k \to \infty$ we have the following bound:

$$\lim_{k\to\infty} \|e_k\|_2 \le \frac{1}{1-\|H\|_2} \frac{1}{1 - \frac{c_1\|H\|_2}{1-\|H\|_2}} \times \left(c_1 \frac{\|H\|_2 \|h\|_2}{1 - \|H\|_2} + \|h\|_2 \mu \right). \tag{5}$$



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



(b) Variation in bound w.r.t. changes in $\|H\|_2$

Fig. 3. Comparison of forward error bounds by control techniques and Wozniakowski [13] for the successive iteration method

D. Comparison with bound by Wozniakowski

In this section we compare the error bound obtained in this paper with a bound given in the numerical analysis literature [13]. The expression for the bound for the case $\|H\|_2 < 1$ is

$$\lim_{k\to\infty} \|e_k\|_2 \leq \frac{c_2\mu(\|H\|_2 + \|I - H\|_2)}{1 - \|H\|_2} \|x^*\|_2 + \mathcal{O}(\mu^2), \tag{6}$$

where c_2 is a constant and $\mathcal{O}(\mu^2)$ denotes the quadratic and higher order terms in μ. A possible value of c_2 is $1.01n^{\frac{3}{2}} + 1$ (Appendix III).

In contrast to the expression for the bound obtained using control tools the bound (6) also involves the problem solution x^*. One usually does not know the solution *a priori*, thus the expression for the bound using control tools is arguably more practical.

Figure 3 shows the comparison of the bound (6) (ignoring $\mathcal{O}(\mu^2)$ terms) and the bound obtained using control tools (5). The expressions for the error bound by Wozniakowski [13] depends on the solution of the problem, therefore we have calculated the error bounds using the exact solution of random problems. In Figure 3, we have plotted the average, maximum and minimum bounds for these problems. The figure also shows the errors in the calculated solution of these random problems to get some idea of the conservativeness of these bounds.

In Figure 3 note that the maximum error bound obtained from the expression given in [13] for the random problems is higher than the bound using control tools. Therefore, the bound calculated using control tools is tighter.

VI. Bounds for the classical iterative methods

In this section we will find bounds for the classical iterative methods [9, Sec. 10.1] for solving a system of linear equations $Ax^* = b$. The classical iterative methods have the following iterations:

$$x_0 = 0, \quad x_{k+1} = M^{-1}(Nx_k + b), \tag{7}$$

where M and N are $n \times n$ matrices such that $A = M - N$. Table I lists some of the classical iterative methods [9, 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, magnitudes of all the eigenvalues of the matrix $M^{-1}N$ are less than one and the algorithm converges to the solution $x^* = A^{-1}b$.

TABLE I

Classical iterative methods

M	Iterative 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

A. Finite precision error analysis for a single iteration

If the algorithm is implemented in floating point, then the finite precision errors in the calculations in each iteration of the algorithm can be represented as an additive error and we have the expression

$$\text{fl}(M^{-1}(Nx_k + b)) = M^{-1}(Nx_k + b) + w_k,$$

where w_k represents the finite precision errors.

If x_{k+1} in (7) is calculated in finite precision floating point, then we can think of the finite precision errors being produced by 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 $Mz = \zeta$, 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 $\zeta \in \mathbb{R}^n$. It is known that when this system is solved in finite precision floating point using the back substitution algorithm [9, Sec. 3.1.2], then the calculated approximate solution $\hat{z}$ is an exact solution of a system of linear equations $(M + \Delta M)\hat{z} = \zeta$, where $\Delta M \in \mathbb{R}^{n\times n}$ and

$$|\Delta M| \leq n\mu\,|M|\,, \tag{8}$$

as shown in [15, p. 101]. Therefore, for iterations of classical iterative methods, the following relation will hold in finite precision:

$$\begin{aligned}(M + \Delta M_{k+1})x_{k+1} &= \text{fl}(Nx_k + b)\\ \Rightarrow x_{k+1} &= M^{-1}\left(\text{fl}\left(Nx_k + b\right) - \Delta M_{k+1}x_{k+1}\right).\end{aligned} \tag{9}$$

Using (8) and (2), we can find the following bound:

$$\begin{aligned}\|w_k\|_\infty &\leq c_3 \left\|M^{-1}\right\|_\infty \|N\|_\infty \|x_k\|_\infty + \mu \|b\|_\infty \\ &\quad + n\mu\gamma_x \left\|M^{-1}\right\|_\infty \|M\|_\infty \|x^*\|_\infty , \end{aligned} \tag{10}$$

where $c_3 := 1.01n\mu^2 + 1.01n\mu + \mu$ and $\gamma_x := \sup_k \frac{\|x_k\|_\infty}{\|x^*\|_\infty}$. Note that due to the condition in (2), this bound only holds for the case $n\mu < 0.01$.

B. Bound on maximum finite precision error over all iterations

In this section we state a theorem that gives bounds on the maximum finite precision error over all iterations for classical iterative methods.

Theorem 3: For $\left\|M^{-1}N\right\|_\infty < 1$ and $n\mu < 0.01$ we have the following bound for the classical iterative methods:

$$\|\mathbf{w}\|_\infty \leq \frac{1}{1-\gamma_1\gamma_2}\left(\gamma_2\beta_1 + \beta_2\right), \tag{11}$$

where

$$\begin{aligned}\gamma_1 &:= (1 - \left\|M^{-1}N\right\|_\infty)^{-1}, \\ \gamma_2 &:= c_3 \left\|M^{-1}\right\|_\infty \|N\|_\infty , \\ \beta_1 &:= (1 - \left\|M^{-1}N\right\|_\infty)^{-1} \left\|M^{-1}b\right\|_\infty , \\ \beta_2 &:= \mu \|b\|_\infty + \mu n\gamma_x \left\|M^{-1}\right\|_\infty \|M\|_\infty \|x^*\|_\infty .\end{aligned}$$

Proof: The proof is similar to that of Theorem 1. ■

C. Bound on forward error

Theorem 4: For $\left\|M^{-1}N\right\|_\infty < 1$ and $n\mu < 0.01$ we have the following forward error bound for the successive iteration method (1):

$$\|e_k\|_\infty \leq \left\|(M^{-1}N)^k\right\|_\infty \|e_0\|_\infty + \frac{\|\mathbf{w}\|_\infty}{1 - \|M^{-1}N\|_\infty}$$

and the bound on $\|\mathbf{w}\|_\infty$ is given by (11).

Proof: The proof is similar to that of Theorem 2. ■

Corollary 3: We have the following bound:

$$\lim_{k\to\infty} \|e_k\|_2 \leq \frac{1}{1 - \|M^{-1}N\|_\infty} \|\mathbf{w}\|_\infty . \tag{12}$$

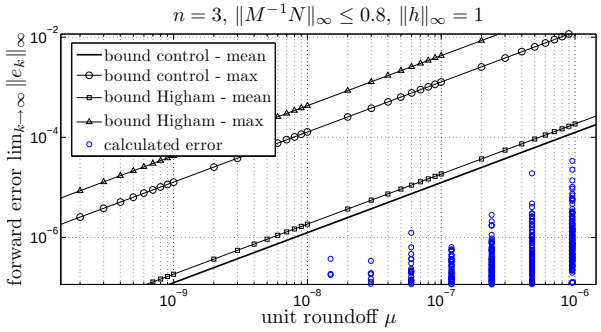
D. Comparison with bound by Higham

In this section we compare the error bound obtained in this paper for the classical iterative methods with the bound given by Higham in [10, Sec. 17.2]. The bound given by Higham is as follows:

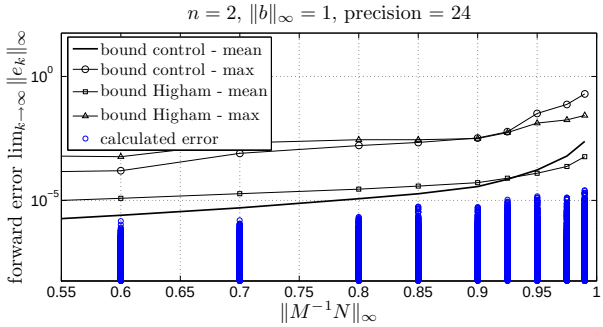
$$\begin{aligned}\lim_{k\to\infty} \|e_k\|_\infty &\leq c_4\mu(1+\gamma_x)(\|M\|_\infty + \|N\|_\infty) \times \\ &\quad \|x^*\|_\infty \frac{\left\|M^{-1}\right\|_\infty}{1 - \|M^{-1}N\|_\infty}, \end{aligned} \tag{13}$$

where $c_4 = (1.01n + 1)$ (see Appendix IV) is a constant.

Figure 4 shows the comparison of the bound (13) and the bound using control tools (12). The expressions for the bounds 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 the mean of bounds obtained for these problems. The random problems



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



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

Fig. 4. Comparison of forward error bounds by control techniques and Higham [10] for classical iterative methods

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, so that the matrix A is diagonally dominant. We generate matrix A in this way because diagonal dominance ensures convergence in 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. In the figures we observe that the bound using control tools is tighter than the bound calculated using (13), unless the value of $\left\|M^{-1}N\right\|_\infty$ is close to unity.

VII. Conclusions

In this paper we have used results from ISS and l_∞-stability theory to find a systematic way of finding bounds on the forward error and the maximum finite precision error. The proposed approach is applied on the successive iteration method and the classical iterative methods. The obtained forward error bound is numerically observed to be tighter than the bounds given in the numerical analysis literature. Moreover, in contrast to bounds in the numerical analysis literature for the successive iteration method, the bound obtained using a control approach does not involve the solution of the problem. Although the proposed scheme provides a new systematic way for forward error analysis, we note that the conditions of ISS and l_∞-stability might be restrictive and not all algorithms will satisfy them.

APPENDIX I
PROOF OF THEOREM 1

The successive iteration algorithm (1) in finite precision floating point can be represented as the following dynamical system:

$$x_{k+1} = Hx_k + h + w_k, \tag{14}$$

with initial condition $x_0 = 0$. This dynamical system can also be represented in the form given in Figure 1 with dynamics of system P given by

$$x_{k+1} = Hx_k + h + w_k, \quad x_0 = 0.$$

For system P, simple analysis leads to the expression

$$\|\mathbf{x}\|_\infty \leq q \|\mathbf{w}\|_\infty + q \|h\|_2 .$$

where $q := \left(\sum_{j=0}^{\infty} \left\|H^j\right\|_2\right)$. The series in the expression for q converges if the spectral radius of H is less than one [10, Prob. 17.1], which is the case here. Therefore, system P is finite gain l_∞-stable with gain $\gamma_1 = q$ and bias $\beta_1 = q\|h\|$. For system E, using (3) we can find the expression

$$\|\mathbf{w}\|_\infty \leq c_1 \|H\|_2 \|\mathbf{x}\|_\infty + \|h\|_2 \mu,$$

which implies that system E is finite gain l_∞-stable with gain $\gamma_2 = c_1 \|H\|_2$ and bias $\beta_2 = \|h\|_2 \mu$.

If the small gain condition, $\gamma_1\gamma_2 = c_1 q \|H\|_2 < 1$, is satisfied, then all the hypotheses of Proposition 1 are true and we have the following bounds:

$$\|\mathbf{x}\|_\infty \leq (1 - c_1 q \|H\|_2)^{-1} q(\mu + 1) \|h\|_2 ,$$

$$\|\mathbf{w}\|_\infty \leq (1 - c_1 q \|H\|_2)^{-1} (c_1 q \|H\|_2 + \mu) \|h\|_2 .$$

If $\|H\|_2 < 1$, then we have

$$q = \left(\sum_{j=0}^{\infty} \left\|H^j\right\|_2\right) \leq \left(\sum_{j=0}^{\infty} \|H\|_2^j\right) = \tfrac{1}{1-\|H\|_2},$$

which along with the above-mentioned bounds completes the proof.

APPENDIX II
PROOF OF THEOREM 2

As seen in (14), the successive iteration algorithm in finite precision can be represented as

$$x_{k+1} = Hx_k + h + w_k.$$

For $\|H\|_2 < 1$, the equilibrium point of the unforced system is $x^* := (I - H)^{-1}h = v^*$. To shift the equilibrium we subtract x^* from both sides of the equation to get

$$e_{k+1} = He_k + w_k,$$

where $e_k := x_k - x^*$ is the solution error.

Simple analysis leads to the following expression:

$$\|e_k\|_2 \leq \left\|H^k\right\|_2 \|e_0\|_2 + q \|\mathbf{w}\|_\infty .$$

Since the spectral radius of H is less than one, $\left\|H^k\right\|_2 \to 0$ as $k \to \infty$ and the series in the equation converges [10, Prob. 17.1], therefore the dynamical system is ISS.

If $\|H\|_2 < 1$, from the above expression we get

$$\|e_k\|_2 \leq \left\|H^k\right\|_2 \|e_0\|_2 + (1 - \|H\|_2)^{-1} \|\mathbf{w}\|_\infty .$$

Using (4b) in the above equation completes the proof.

APPENDIX III
CONSTANT IN BOUND BY WOZNIAKOWSKI

It is assumed in [13] that for a matrix $M \in \mathbb{R}^{n\times n}$ and vectors z, $y \in \mathbb{R}^n$ we have

$$\mathrm{fl}(Mz + y) = (M + \delta M)z + (I + \delta I)y,$$

where $\|\delta M\| \leq \mu c_5 \|M\|$, $\|\delta I\| \leq \mu c_6$; c_5 and c_6 are constants only depending on n. From (2) and some simple manipulation we obtain possible values for these constants; $c_5 = 1.01n^{3/2} + 1$ and $c_6 = 1$.

The constant c_2 in (6) is defined in [13] as $c_2 := \max(c_5, c_6)$. Therefore, a possible value for the constant is $c_2 = 1.01n^{3/2} + 1$.

APPENDIX IV
CONSTANT IN BOUND BY HIGHAM

In [10], the constant c_4 in (13) is defined such that the following holds:

$$|\xi_k| \leq c_4\mu(|M|\,|x_{k+1}| + |N|\,|x_k| + |b|),$$

for all k, where

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

see [10, Eq. 17.2]. According to analysis similar to the one in Section VI-A, a possible value of c_4 is $(1.01n + 1)$.

REFERENCES

[1] James Hurt. Some stability theorems for ordinary difference equations. *SIAM Journal on Numerical Analysis*, 4(4):582–596, 1967.

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

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

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

[5] Oumar Diene and Amit Bhaya. A study of the robustness of iterative methods for linear systems. In *8th International Conference on Numerical Analysis and Applied Mathematics*, Rethymno, Greece, 2009.

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

[7] M. T. Chu. Linear algebra algorithms as dynamical systems. *Acta Numerica*, 17:1–87, 2008.

[8] Uwe Helmke, Jens Jordan, and Alexander Lanzon. A control theory approach to linear equation solvers. In *Proc. 17th International Symposium on Mathematical Theory of Networks and Systems*, Kyoto, Japan, 2006.

[9] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.

[10] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2002.

[11] M. Vidyasagar. *Non-linear Systems Analysis*. Society for Industrial and Applied Mathematics, 2002.

[12] Z. P. Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6):857–869, 2001.

[13] H. Wozniakowski. Round-off error analysis of iterations for large linear systems. *Numerische Mathematik*, 30(3):301–314, 1978.

[14] B. Barrowes. Multiple Precision Toolbox. `http://www.sondette.com/math/mp_toolbox.html`, February 2010.

[15] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall, Inc., 1963.