Journal of Computational and Applied Mathematics 242 (2013) 185–212

Contents lists available at SciVerse ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Accurate solution of dense linear systems, Part II: Algorithms using directed rounding

Siegfried M. Rump *

Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany
Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ARTICLE INFO

Article history:
Received 12 August 2011
Received in revised form 22 August 2012

MSC:
65F05
65G20

Keywords:
Linear systems
Matlab
Data with tolerances
Inner bounds
(Extremely) ill-conditioned matrices
Rigorous error bounds

ABSTRACT

In Part I and this Part II of our paper we investigate how extra-precise evaluation of dot products can be used to solve ill-conditioned linear systems rigorously and accurately. In Part I only rounding to nearest is used. In this Part II we improve the results significantly by permitting directed rounding. Linear systems with tolerances in the data are treated, and a comfortable way is described to compute error bounds for extremely ill-conditioned linear systems with condition numbers up to about $\mathbf{u}^{-2}/n$, where $\mathbf{u}$ denotes the relative rounding error unit in a given working precision. We improve a method by Hansen/Bliek/Rohn/Ning/Kearfott/Neumaier. Of the known methods by Krawczyk, Rump, Hansen et al., Ogita and Nguyen we show that our presented Algorithm `LssErrBnd` seems the best compromise between accuracy and speed. Moreover, for input data with tolerances, a new method to compute componentwise inner bounds is presented. For not too wide input data they demonstrate that the computed inclusions are often almost optimal. All algorithms are given in executable Matlab code and are available from my homepage.

1. Introduction and notation

The paper divides into two parts: In Part I all algorithms use only the four basic floating-point operations in rounding to nearest, in the present Part II we use directed rounding and methods different from Part I to obtain superior results. All algorithms in both parts are presented in executable Matlab code.

The methods in Part I are based on norm estimates, verifying convergence of some residual matrix by approximating its Perron vector. In this Part II we verify the H-property of some matrix and demonstrate how this can be used to effectively compute verified error bounds of the solution of a linear system. Moreover, we show an efficient method to compute so-called “inner” inclusions of a linear system, the data of which is afflicted with tolerances.

Dividing the paper into two parts also serves didactical purposes. In Part I we demonstrate that using only rounding to nearest allows us to give simple algorithms to produce rigorous results. The algorithms presented in Part II can still be formulated in rounding to nearest, however, at the cost of easy readability.

All algorithms are given in executable Matlab code for which we reserve the “verbatim”-font. For instance, $\mathtt{C} = \mathtt{A} * \mathtt{B}$ means that `C` is the result of the floating-point multiplication $\mathtt{A} * \mathtt{B}$, where `A` and `B` are compatible quantities (scalar, vector, matrix). For analyzing the error we use ordinary mathematical notation, for example in $P = \mathtt{A} \cdot \mathtt{B}$ the verbatim-font is used for floating-point quantities so that P is the exact (real) product of `A` and `B`. For $\mathtt{A}, \mathtt{B} \in \mathbb{F}^{n\times n}$ this implies $|P - \mathtt{C}| \sim \mathbf{u}|\mathtt{A}| \cdot |\mathtt{B}|$.

* Correspondence to: Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany.
E-mail address: rump@tu-harburg.de.

0377-0427/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2012.09.024

Comparison between vectors and matrices is always to be understood entrywise, for example $x \le y$ for $x, y \in \mathbb{R}^n$ means $x_i \le y_i$ for $1 \le i \le n$. For $A \in \mathbb{R}^{n\times n}$, Ostrowski's comparison matrix $\langle A\rangle \in \mathbb{R}^{n\times n}$ is defined by

$$\langle A\rangle_{ii} := |A_{ii}| \quad \text{and} \quad \langle A\rangle_{ij} := -|A_{ij}| \text{ for } i \ne j. \tag{1.1}$$

If $\langle A\rangle$ is an M-matrix, then A is called an H-matrix. In that case A and $\langle A\rangle$ are non-singular, and $|A^{-1}| \le \langle A\rangle^{-1}$. Finally, $\varrho(A)$ denotes the spectral radius of A.

This Part II of the paper is organized as follows. In the next section we discuss several methods to obtain rigorous error bounds for linear systems based on H-matrices. In particular an efficient and best way is shown how to verify the H-property. In Section 3 we discuss how to implement these methods for rigorous error bounds using directed rounding.

Up to this point, standard Matlab suffices. From Section 4 on, the algorithms become too involved and we use INTLAB [1], the Matlab toolbox for reliable computing. The toolbox INTLAB is entirely written in Matlab and thus portable in many environments. The only features of INTLAB we need are the basic interval operations, so other libraries such as Intlib [2], Profil/Bias [3,4] or b4m [5] may be used as well.

In Section 4 linear systems, the data of which are afflicted with tolerances, are treated. Using interval operations the code becomes easier to read without sacrificing performance and/or accuracy. Now the floor is prepared to discuss alternative approaches to compute rigorous error bounds in Section 5. In particular a method originated by Hansen is discussed and improved. In Section 6 we show how to obtain inclusions for extremely ill-conditioned matrices, i.e. with condition number up to $\mathbf{u}^{-2}$, and finally we show a new method to calculate inner inclusions, even if only one entry of the matrix and/or the right hand side is afflicted with a tolerance. This allows us to judge the quality of outer inclusions. Detailed computational results and a conclusion finish the paper.

2. Rigorous error bounds for linear systems

Let a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ be given. As in Part I let $R \in \mathbb{R}^{n\times n}$ be an approximate inverse of A, for example computed by the Matlab command `inv`. Note that there are no *a priori* assumptions on A and R, in particular no accuracy requirement on R. For the following explanation assume the matrices A and R to be non-singular; in the following theorems this will be verified *a posteriori* by the methods.

Define $C := RA$. For an approximation $\widetilde{x}$ to the solution $A^{-1}b$ we show several ways how to estimate

$$\Delta := \widetilde{x} - A^{-1}b = C^{-1}R(A\widetilde{x} - b) = C^{-1}c \quad \text{with } c := R(b - A\widetilde{x}). \tag{2.1}$$

In Part I of this paper we used normwise error estimates. Define $T := \operatorname{diag}(t)$ for a positive vector $t \in \mathbb{R}^n$. Defining $F := I - C$ and exploring

$$C^{-1}c = c + C^{-1}Fc = c + T(I - T^{-1}FT)^{-1}T^{-1}Fc \tag{2.2}$$

and using $|Ex| \le \|x\|_\infty \cdot |E|e$ for $E \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ and $e := (1, \ldots, 1) \in \mathbb{R}^n$ we obtain

$$\begin{aligned} |\widetilde{x} - A^{-1}b| &\le |c| + \|(I - T^{-1}FT)^{-1}T^{-1}Fc\|_\infty \cdot Te \\ &\le |c| + \frac{\|T^{-1}Fc\|_\infty}{1 - \|T^{-1}FT\|_\infty} \cdot t, \end{aligned} \tag{2.3}$$

provided $\|T^{-1}FT\|_\infty < 1$ (see Theorem 4.14 in Part I). Note that this is true for any $0 < t \in \mathbb{R}^n$. Since $\|T^{-1}FT\|_\infty = \|T^{-1}|F|T\|_\infty = \|T^{-1}|F|Te\|_\infty$, the obvious choice for Te is the Perron vector t of $|F|$. Then $|F|t = \varrho t$ implies $T^{-1}|F|Te = \varrho e$, minimizing $\|T^{-1}FT\|_\infty$. For very ill-conditioned matrices, the choice $t = e$ may fail due to $\|F\|_\infty \ge 1$, see Fig. 4.1 in Part I; otherwise, however, due to rounding errors in finite precision, sometimes the choice $t = e$ is superior to the Perron vector.

The previous result is based on the Neumann expansion $C^{-1} = (I - F)^{-1} = I + C^{-1}F$. If A is not too ill-conditioned, then $C = RA$ is not too far from the identity matrix and likely to be an H-matrix, which means that $\langle C\rangle$ is an M-matrix. The matrix C is an H-matrix, also called generalized diagonally dominant, if and only if there exists some positive $v \in \mathbb{R}^n$ such that $u := \langle C\rangle v > 0$. For the moment assume such a vector v to be given.

Then $C = RA$ is an H-matrix, so that A and R are non-singular. Denote by

$$\langle C\rangle := D - E \quad \text{with } D > 0, E \ge 0 \tag{2.4}$$

the splitting of $\langle C\rangle$ into diagonal and off-diagonal part. Note that $u = \langle C\rangle v > 0$ implies $D > 0$. Set

$$G := I - \langle C\rangle D^{-1} = ED^{-1} \ge 0. \tag{2.5}$$

Following [6] define

$$w_k := \max_i \frac{G_{ik}}{u_i} \quad \text{for } 1 \le k \le n, \tag{2.6}$$