submitted for publication August 2, 2013; revised April 23, 2014

# VERIFIED BOUNDS FOR THE P-NORM CONDITION NUMBER

SIEGFRIED M. RUMP *

**Abstract.** Methods to compute verified error bounds for the $p$-norm condition number of a matrix are discussed for $p \in \{1, 2, \infty\}$ and the Frobenius norm. We consider the cases of a real or complex, point or interval input matrix. In the latter case the condition number of all matrices within the interval matrix are bounded. A special method for extremely ill-conditioned matrices is derived as well. Numerical results suggest that the quality of the bounds corresponds to the fact that the condition number of the condition number is the condition number.

**Key words.** Rigorous error estimates, condition number, p-norm, INTLAB.

**AMS subject classifications.** 65G20, 15A18

**1. Introduction and notation.** Let a real or complex square matrix $A$ be given, that is $A \in \mathbb{K}^{n\times n}$ where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. When using $\|\cdot\|_p$ we henceforth assume $p \in \{1, 2, \infty, F\}$, where $\|A\|_F$ denotes the Frobenius norm $\sqrt{\sum a_{ij}^2}$. We derive methods to compute verified error bounds for $\kappa_p(A) := \|A\|_p \|A^{-1}\|_p$. Note that finite bounds for the condition number imply $A$ to be non-singular. We also consider the case of an interval matrix $\mathbf{A} \in \mathbb{IK}^{n\times n}$, where all condition numbers of $A \in \mathbf{A}$ are bounded. Moreover, a special method for extremely ill-conditioned matrices is presented. We assume the reader to be familiar with basic concepts of interval arithmetic, cf. [5].

A straightforward INTLAB [7] statement to compute an inclusion `Cnd` $\in \mathbb{IR}$ of $\kappa_p(A)$ is

```
Cnd = norm(intval(A),p)*norm(inv(intval(A)),p)
```  (1.1)

which works both for point and interval, real and complex matrices `A`. It fails for too ill-conditioned matrices with `Cnd` = NaN. In the following we discuss alternative and specialized methods to compute an inclusion of the condition number. We frequently need

$$A, B \in \mathbb{K}^{n\times n} \text{ and } |A| \le B \quad \Rightarrow \quad \|A\|_p \le \|B\|_p \ , \tag{1.2}$$

where comparison is understood to be componentwise. This is clear for $p \in 1, \infty, F$. For $p = 2$ use $\|A\|_2^2 \le \| \, |A| \, \|_2^2 = \rho(|A|^T|A|) \le \rho(B^T B) = \|B\|_2^2$, where $\rho(\cdot)$ denotes the spectral radius and the latter inequality follows by Perron-Frobenius Theory [4].

**2. Main results.** Bounds for the $p$-norm of a matrix with $p \in \{1, \infty, F\}$ are immediate using interval arithmetic, and bounds for the spectral norm are obtaind by the methods in [8]. So we concentrate on bounds for the norm of the inverse of a matrix $A$.

Such bounds are computed on the basis of an *approximate* inverse $R$ of $A$. The quality of the bounds for $\|A^{-1}\|$ relates directly to the quality of the approximation $R \approx A^{-1}$. This quality depends, in turn, on the condition number of the matrix $A$, which is to be bounded and which is not known a priori.

*Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany, and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan (`rump@tu-harburg.de`).

What sounds like a vicious circle is resolved by computing a bound $\alpha$ for the norm of $E := I - RA$. If $\alpha < 1$, then lower and upper bounds for the norm of the true inverse $A^{-1}$ follow using the norm of $R$. The norm of $R$ is bounded using interval arithmetic as previously the norm of $A$.

Note that mathematically $R \in \mathbb{F}^{n\times n}$ is an arbitrary matrix. If $R$ is of poor quality, the condition $\alpha < 1$ is not satisfied. Practically $R$ is computed by some standard numerical algorithm, usually based on Gaussian elimination with partial pivoting. Numerically, we can expect $\alpha < 1$ if $\kappa(A) \lesssim \mathrm{eps}^{-1}$ for eps denoting the relative rounding error unit (about $1.1 \cdot 10^{-16}$ in IEEE 754 binary64). It is important to note that all our bounds are valid without any *a priori* assumption the condition number of $A$ or the the quality of $R$.

**2.1. Perturbations of the identity matrix.** Denoting the identity matrix of proper dimension by $I$, we first investigate $\|(I-E)^{-1}\|_p$ for admissible perturbations $E$.

LEMMA 2.1. *Let $E \in \mathbb{K}^{n\times n}$ and $p \in \{1, 2, \infty\}$ be given, and suppose $\|E\|_p \le \alpha < 1$. Then*

$$\frac{1}{1+\alpha} \le \|(I-E)^{-1}\|_p \le \frac{1}{1-\alpha} \ . \tag{2.1}$$

*Equality is attained in both bounds for $E = -\alpha I$ and $E = \alpha I$, respectively, for all $p$ and any $n$.*

PROOF. Follows by the Neumann series and $1 = \|(I-E)(I-E)^{-1}\|_p \le (1+\alpha)\|(I-E)^{-1}\|_p$, a commonly used "splitting of 1". □

Note the apparent unsymmetry of the bounds: For nonzero $\alpha$ the lower bound is closer to 1 than the upper bound. This will be utilized later. For the Frobenius norm things are a little more involved.

LEMMA 2.2. *Let $E \in \mathbb{K}^{n\times n}$ be given, and suppose $\|E\|_F \le \alpha < 1$. Then*

$$\frac{\sqrt{n}}{1+\|E\|_2} \le \|(I-E)^{-1}\|_F \le \frac{\sqrt{n}}{1-\|E\|_2} \ . \tag{2.2}$$

*Equality is attained in both bounds for $\beta = n^{-1/2}\alpha$ and $E = -\beta I$ and $E = \beta I$, respectively, for any $n$. Moreover,*

$$\sqrt{n} - \frac{\alpha}{1-\alpha} \le \|(I-E)^{-1}\|_F \le \sqrt{n} + \frac{\alpha}{1-\alpha} \ . \tag{2.3}$$

*The left and right bound in (2.3) can be better up to a factor 2 and $\sqrt{n}$ than the corresponding bounds in (2.2), respectively. Conversely, both bounds in (2.3) can be arbitarily weaker than the bounds in (2.2).*

PROOF. The standard perturbation bound for singular values (Theorem 3.3.16 in [3]) states that

$$A, E \in \mathbb{K}^{m\times n} \qquad \Rightarrow \qquad |\sigma_i(A+E) - \sigma_i(A)| \le \|E\|_2 \quad \text{for } 1 \le i \le \min(m,n) \ .$$

Setting $m = n$ and $A := I$ it follows that for all $1 \le i \le n$

$$1 - \|E\|_2 \ \le \ \sigma_i(I-E) \ \le \ 1 + \|E\|_2 \ .$$

Then $\|E\|_2 \le \|E\|_F$ and $\|E\|_F \le \alpha < 1$ shows that $I - E$ has full rank and

$$\frac{1}{1+\|E\|_2} \ \le \ \sigma_i\big((I-E)^{-1}\big) \ \le \ \frac{1}{1-\|E\|_2} \ . \tag{2.4}$$

For any matrix $M \in \mathbb{K}^{n\times n}$

$$\|M\|_F^2 = \mathrm{trace}(M^T M) = \sum_{i=1}^{n} \sigma_i(M^T M) = \sum_{i=1}^{n} \sigma_i^2(M)$$

because the eigenvalues and singular values of $M^T M$ coincide, so that (2.4) implies the bounds in (2.2). Furthermore,

$$\|(I-E)^{-1}\|_F = \|I + \sum_{\nu=1}^{\infty} E^\nu\|_F \quad \text{and} \quad \|\sum_{\nu=1}^{\infty} E^\nu\|_F \le \frac{\alpha}{1-\alpha}$$

imply the bounds in (2.3). Using $\|E\|_2 \le \|E\|_F$ implies

$$\frac{\text{left (2.3)}}{\text{left (2.2)}} \le \frac{\sqrt{n} - \alpha/(1-\alpha)}{\sqrt{n}/(1+\alpha)} \le 1 + \alpha < 2$$

and

$$\frac{\text{right (2.2)}}{\text{right (2.3)}} \le \frac{\sqrt{n}}{\sqrt{n}(1-\alpha)+\alpha} = \frac{1}{1-\alpha(1-1/\sqrt{n})} \le \sqrt{n}\ .$$

For $\alpha$ close to 1 the bounds in (2.3) tend to $\pm\infty$ whereas, with a gap between $\|E\|_2$ and $\|E\|_F$, the bounds in (2.2) remain finite. A simple example is again $E$ being a multiple of the identity matrix. This completes the proof. □

**2.2. General point matrices.** Bounds for the condition number of a point matrix $A \in \mathbb{K}^{n\times n}$ follow by direct application of the previous lemma.

LEMMA 2.3. *Let $A, R \in \mathbb{K}^{n\times n}$ and $p \in \{1, 2, \infty, F\}$ be given, abbreviate $E := I - RA$ and assume $\|E\|_p \le \alpha$ for some $\alpha \in \mathbb{R}$. Suppose $\alpha < 1$ in case $p \in \{1, 2, \infty\}$, and $\|E\|_2 < 1$ in case $p = F$. Then $A$ and $R$ are non-singular and*

$$\frac{\|A\|_p \|R\|_p}{1+\alpha} \le \kappa_p(A) \le \frac{\|A\|_p \|R\|_p}{1-\alpha} \qquad \text{for } p \in \{1, 2, \infty\}\ , \tag{2.5}$$

*and*

$$\|A\|_F \|R\|_F \big(1-\alpha\beta\big) \le \kappa_F(A) \le \|A\|_F \|R\|_F \big(1+\alpha\beta\big) \tag{2.6}$$

*for $\beta := \sqrt{n}/(1-\|E\|_2)$. If $\alpha < 1$ in case $p = F$, then (2.6) is also true for $\beta = \sqrt{n} + \alpha/(1-\alpha)$.*

PROOF. The assumptions imply $I - E$ and therefore $A$ and $R$ to be invertible. Using

$$A^{-1} = (I-E)^{-1}R = R + (I-E)^{-1}ER$$

togther with Lemmata 2.1 and 2.2 yields the right bound in (2.5) and both bounds in (2.6), whereas the left bound in (2.5) follows by $\|R\|_p = \|(I-E)(I-E)^{-1}R\|_p = \|(I-E)A^{-1}\|_p \le (1+\alpha)\|A^{-1}\|_p$. □

Note that for both lower bounds only upper bounds of $\|(I-E)^{-1}\|$ have been used. For the lower bound in (2.5) we could use the splitting of 1 since $\|I\|_p = 1$ for $p \in \{1, 2, \infty\}$; this is not possible for the lower bound in (2.6). One might be inclined to expand $(I-E)^{-1} = I + (I-E)^{-1}E$ to improve the bounds in (2.5) as well. However, this yields

$$\|R\|\,\big(1 - \frac{\alpha}{1-\alpha}\big) \le \|A^{-1}\| \le \|R\|\,\big(1 + \frac{\alpha}{1-\alpha}\big) \tag{2.7}$$

implying the same upper, but a weaker lower bound. This is due to the mentioned unsymmetry of the bounds in Lemma 2.1.

**2.3. General interval matrices.** Let $\mathbf{A} \in \mathbb{IK}^{n\times n}$ be given, so that

$$\mathbf{A} = \langle M, \Delta\rangle := \{\widetilde{A} \in \mathbb{K}^{n\times n} : |\widetilde{A} - M| \le \Delta\} \qquad \text{for } M \in \mathbb{K}^{n\times n} \text{ and } 0 \le \Delta \in \mathbb{R}^{n\times n} \ . \tag{2.8}$$

Let $\widetilde{A} \in \mathbf{A}$ be fixed but arbitrary, so that $\widetilde{A} = M + \widetilde{\Delta}$ with $|\widetilde{\Delta}| \le \Delta$. Denoting $\widetilde{E} := I - R\widetilde{A}$ and using (1.2) it follows

$$\|\widetilde{E}\| \le \|I - RM\| + \|R\|\|\widetilde{\Delta}\| \le \|I - RM\| + \|R\|\|\Delta\| \ .$$

Furthermore, again using (1.2),

$$\|M\|_p - \|\Delta\|_p \le \|M\|_p - \|\widetilde{\Delta}\|_p \le \|\widetilde{A}\|_p \le \|M\|_p + \|\widetilde{\Delta}\|_p \le \|M\|_p + \|\Delta\|_p \ .$$

Proceeding as before by using $\widetilde{A}^{-1} = (I - \widetilde{E})^{-1}R = R + (I - \widetilde{E})^{-1}\widetilde{E}R$ we obtain

LEMMA 2.4. *Let* $\mathbf{A}$ *as in (2.8),* $R \in \mathbb{K}^{n\times n}$ *and* $p \in \{1, 2, \infty, F\}$ *be given, and assume* $\|I - RM\|_p + \|R\|_p\|\Delta\|_p < \gamma_p$ *for some* $\gamma_p \in \mathbb{R}$. *Suppose* $\gamma_p < 1$ *in case* $p \in \{1, 2, \infty\}$, *and* $\gamma_2 < 1$ *in case* $p = F$. *Then for every* $\widetilde{A} \in \mathbf{A}$

$$\frac{(\|M\|_p - \|\Delta\|_p)\|R\|_p}{1 + \gamma_p} \le \kappa_p(\widetilde{A}) \le \frac{(\|M\|_p + \|\Delta\|_p)\|R\|_p}{1 - \gamma_p} \qquad \text{for } \ p \in \{1, 2, \infty\} \ , \tag{2.9}$$

*and*

$$\big(\|M\|_F - \|\Delta\|_F\big)\|R\|_F\big(1 - \gamma_F\beta\big) \le \kappa_F(\widetilde{A}) \le \big(\|M\|_F + \|\Delta\|_F\big)\|R\|_F\big(1 + \gamma_F\beta\big) \tag{2.10}$$

*for* $\beta := \sqrt{n}/(1 - \gamma_2)$. *If* $\gamma_F < 1$, *then (2.10) is also true for* $\beta = \sqrt{n} + \gamma_F/(1 - \gamma_F)$.

Again one might try to improve the bounds using $\widetilde{A}^{-1} = (I + M^{-1}\widetilde{\Delta})^{-1}M^{-1}$ and the implicit estimate for $\|M^{-1}\|$ in (2.5). But again a computation yields that the upper bounds coincide, whereas the lower bound is weaker than in (2.9).

**2.4. Extremely ill-conditoned matrices.** We finally address matrices with $\kappa(A) > \text{eps}^{-1}$ for eps denoting the relative rounding error unit. In IEEE 754 binary64 (double precision) this means $\kappa(A) > 10^{16}$. In that case there is likely no preconditioner $R \in \mathbb{K}^{n\times n}$ at all such that $I - RM$ is convergent.

Here a method as in [9, Section 6], which is based on unpublished results of the mid 1980's (see the citations over there), can be used. If $\kappa(A) > \text{eps}^{-1}$, then an approximate inverse has, in general, no correct digit, but it contains enough information to ensure $\kappa(RA) \approx \text{eps} \cdot \kappa(A)$. Denote $B := RA$, and let $S$ be an approximate inverse of $B$. Abbreviating $F := I - SB$ and assuming $\|F\|_p \le \alpha < 1$ we obtain as before $A^{-1} = B^{-1}R = (I - F)^{-1}SR$, and Lemma 2.3 implies

$$\frac{\|A\|_p\|SR\|_p}{1 + \alpha} \le \kappa_p(A) \le \frac{\|A\|_p\|SR\|_p}{1 - \alpha} \tag{2.11}$$

for $p \in \{1, 2, \infty\}$, and similarly for the Frobenius norm using Lemma 2.2.

It is seen in the final section that this approach works up to condition numbers near $\text{eps}^{-2}$, which is up to $10^{32}$ in IEEE 754 binary64. Note that this is far beyond the scope of common numerical algorithms. It is mandatory in that case to use extra-precise accumulation of dot products for computing the error bounds for $B = RA$ (and only there). A number of such algorithms using only floating-point operations in the current working precision are available, see [8] and the literature cited over there.

We close this section with some hints on the implementation for $p = 2$, the spectral condition number. The spectral norm of a matrix can be bounded using the methods in [8], however, this requires $\mathcal{O}(n^3)$ operations. In contrast, bounds for the spectral norm of a non-negative matrix $C$ are computed in $\mathcal{O}(n^2)$ operations using $\|C\|_2^2 = \rho(C^TC)$, $C^TC \geq 0$ and Perron-Frobenius Theory. The inequalities

$$\min_i \frac{(C^TCx)_i}{x_i} \leq \rho(C^TC) \leq \max_i \frac{(C^TCx)_i}{x_i} \tag{2.12}$$

are valid for any positive vector $x$, see [1]. Here $(\cdot)_i$ denotes the $i$-th entry of a vector. Since the vector $x$ is arbitrary, a power set iteration replacing $x$ by $C^T(Cx)$ may be perfomed. Each iteration requires $\mathcal{O}(n^2)$ floating-point operations. It is known [10] that this produces a nested sequence of lower and upper bounds, guaranteed to improve in each step.

The components of the matrix $E = I - RA$ or $I - RM$ are, for not too ill-conditioned matrices, likely to be small, and usually it does no harm to bound the norm by $\| \, |E| \, \|$. To this end (2.12) can be used, and also for bounding $\| \, |\Delta| \, \|$. This method should not be used to bound $\|A\|$ or $\|R\|$.

Note that all approaches use an approximate inverse, so they may be time and memory consuming for sparse matrices. However, error bounds for the condition number of a matrix imply bounds for the solution of linear systems. For general, sparse matrices this is one of the grand challenges in verification methods [6]. The presented algorithms will be included in Version 7.2 of INTLAB [7].

**3. Computational results.** Following we show some computational results. Matrices of specified condition number are generated by predefining a geometric sequence of singular values as in the Matlab function `randsvd`; for extremely ill-conditioned matrices we use the INTLAB-function `randmat`, see [9].

| | condition number | | | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $10^2$ | $10^5$ | $10^{10}$ | $10^{13}$ | $10^{14}$ | $10^{15}$ | fail |
| 10 | $3.1 \cdot 10^{-14}$ | $1.6 \cdot 10^{-11}$ | $1.5 \cdot 10^{-6}$ | $1.5 \cdot 10^{-3}$ | $1.4 \cdot 10^{-2}$ | $1.3 \cdot 10^{-1}$ | 0 |
| 20 | $5.9 \cdot 10^{-14}$ | $2.7 \cdot 10^{-11}$ | $2.1 \cdot 10^{-6}$ | $2.0 \cdot 10^{-3}$ | $1.9 \cdot 10^{-2}$ | $1.9 \cdot 10^{-1}$ | 0 |
| 50 | $1.9 \cdot 10^{-13}$ | $6.6 \cdot 10^{-11}$ | $4.3 \cdot 10^{-6}$ | $3.5 \cdot 10^{-3}$ | $3.9 \cdot 10^{-2}$ | $4.2 \cdot 10^{-1}$ | 0 |
| 100 | $1.0 \cdot 10^{-5}$ | $1.0 \cdot 10^{-5}$ | $1.9 \cdot 10^{-5}$ | $7.8 \cdot 10^{-3}$ | $7.1 \cdot 10^{-2}$ | 1.2 | 5 |
| 200 | $1.0 \cdot 10^{-5}$ | $1.0 \cdot 10^{-5}$ | $3.1 \cdot 10^{-5}$ | $1.6 \cdot 10^{-2}$ | $1.6 \cdot 10^{-1}$ | 30 | 95 |
| 500 | $1.0 \cdot 10^{-5}$ | $1.0 \cdot 10^{-5}$ | $6.5 \cdot 10^{-5}$ | $4.1 \cdot 10^{-2}$ | $4.6 \cdot 10^{-1}$ | - | 100 |
| 1000 | $1.0 \cdot 10^{-5}$ | $1.0 \cdot 10^{-5}$ | $9.7 \cdot 10^{-5}$ | $6.5 \cdot 10^{-2}$ | $9.4 \cdot 10^{-1}$ | - | 100 |

TABLE 3.1

*Test for $p = 2$, median over 100 samples, number of failures for condition number $10^{15}$ in the last column.*

For a given matrix $A$ and computed bound $\mathbf{c} \in \mathbb{IR}$ for $\kappa(A)$ the following tables show the relative error $\phi(\mathbf{c}, c) := \mathrm{rad}(\mathbf{c})/c$, where $\mathrm{rad}(\mathbf{c}) := (\overline{c} - \underline{c})/2$ denotes the radius of $\mathbf{c} = [\underline{c}, \overline{c}]$. Here $c =$ `cond(A)` is the approximate condition number computed by ordinary Matlab for anticipated condition numbers up to $10^{15}$, and using the symbolic toolbox beyond. The quality of the approximation seems still acceptable for $\kappa(A) \lesssim 10^{15}$.

Tables 3.1 and 3.2 show the results for the spectral norm and the 1-norm, respectively. Bounds for the condition number are calculated for 100 samples, and the median of $\phi(\mathbf{c}, c)$ of the bounds is displayed. The lower bound of $\mathbf{c}$ is set to zero if it becomes negative.

| | condition number | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $10^2$ | $10^5$ | $10^{10}$ | $10^{13}$ | $10^{14}$ | fail | $10^{15}$ | fail |
| 10 | $3.8 \cdot 10^{-14}$ | $2.4 \cdot 10^{-11}$ | $2.3 \cdot 10^{-6}$ | $2.2 \cdot 10^{-3}$ | $2.1 \cdot 10^{-2}$ | 0 | $2.1 \cdot 10^{-1}$ | 0 |
| 20 | $7.4 \cdot 10^{-14}$ | $4.5 \cdot 10^{-11}$ | $3.6 \cdot 10^{-6}$ | $3.5 \cdot 10^{-3}$ | $3.3 \cdot 10^{-2}$ | 0 | $3.7 \cdot 10^{-1}$ | 0 |
| 50 | $2.2 \cdot 10^{-13}$ | $1.2 \cdot 10^{-10}$ | $8.3 \cdot 10^{-6}$ | $7.0 \cdot 10^{-3}$ | $7.0 \cdot 10^{-2}$ | 0 | 1.1 | 11 |
| 100 | $5.4 \cdot 10^{-13}$ | $2.6 \cdot 10^{-10}$ | $1.6 \cdot 10^{-5}$ | $1.4 \cdot 10^{-2}$ | $1.4 \cdot 10^{-1}$ | 0 | 3.6 | 87 |
| 200 | $1.4 \cdot 10^{-12}$ | $6.2 \cdot 10^{-10}$ | $3.7 \cdot 10^{-5}$ | $3.1 \cdot 10^{-2}$ | $3.3 \cdot 10^{-1}$ | 0 | – | 100 |
| 500 | $9.2 \cdot 10^{-12}$ | $3.9 \cdot 10^{-9}$ | $2.3 \cdot 10^{-4}$ | $2.0 \cdot 10^{-1}$ | – | 100 | – | 100 |
| 1000 | $2.1 \cdot 10^{-11}$ | $7.7 \cdot 10^{-9}$ | $3.7 \cdot 10^{-4}$ | $3.4 \cdot 10^{-1}$ | – | 100 | – | 100 |

TABLE 3.2

*Test for $p = 1$, median over 100 samples, number of failures for condition numbers $10^{14}$ and $10^{15}$.*

For the spectral norm the method works in all examples up to condition number $10^{14}$, for $p = 1$ always up to condition number $10^{13}$ in our examples. If there are failures for a certain condition number, the number of failures is displayed in the following column. For the spectral norm and dimensions up to 50, bounds for the norm are computed on the base of a full eigendecomosition, otherwise the accuracy is restricted to about $10^{-6}$ to save computing time. This is because an inclusion of the condition accurate to a few digits seems appropriate and sufficient in almost all applications.

Despite this the computational results concur with the fact that the condition number of the condition number is the condition number [2]. The results for the Frobenius norm are a little bit better in accuracy than those for the 1-norm, and (of course) equal to those for the spectral norm in terms of failures.

The final Tables 3.3 and 3.4 show the results for the spectral norm and the 1-norm, respectively, for extremely ill-conditioned matrices. As has been mentioned, (only) the product $RA$ is computed with extra-precise dot product accumulation. It means that each entry $(RA)_{ij}$ is computed with algorithm `Dot2` in [8]. This algorithm uses only floating-point operations in working precision and produces a result *as if* computed in twice the working precision. Otherwise the bounds in (2.11) are straightforwardly computed.

The results are as expected in terms of accuracy and failure. As before, the results for the Frobenius norm are similar in accuracy to those of the 1-norm, and equal to those the spectral norm in terms of failure.

| | condition number | | | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $10^{15}$ | $10^{20}$ | $10^{25}$ | $10^{28}$ | fail | $10^{29}$ | fail |
| 10 | $7.2 \cdot 10^{-15}$ | $1.7 \cdot 10^{-13}$ | $9.3 \cdot 10^{-9}$ | $6.7 \cdot 10^{-6}$ | 0 | $3.3 \cdot 10^{-6}$ | 0 |
| 20 | $1.3 \cdot 10^{-14}$ | $2.1 \cdot 10^{-13}$ | $3.7 \cdot 10^{-8}$ | $1.6 \cdot 10^{-5}$ | 1 | $3.0 \cdot 10^{-5}$ | 0 |
| 50 | $3.5 \cdot 10^{-14}$ | $1.3 \cdot 10^{-12}$ | $3.0 \cdot 10^{-8}$ | $1.6 \cdot 10^{-4}$ | 1 | $2.6 \cdot 10^{-1}$ | 91 |
| 100 | $1.0 \cdot 10^{-5}$ | $1.5 \cdot 10^{-5}$ | $2.0 \cdot 10^{-4}$ | $8.5 \cdot 10^{-2}$ | 26 | $8.8 \cdot 10^{-1}$ | 99 |

TABLE 3.3

*Test for $p = 2$, median over 100 samples, number of failures for condition numbers $10^{28}$ and $10^{29}$.*

We finally mention that using (1.1) produces, in general, more accurate inclusions but need more computing time. Table 3.5 shows the median of the computing times of 100 examples each. We choose some random matrices because there is not much difference in time for well- or ill-conditioned matrices. As can be seen the new approach is faster than (1.1). The timing for interval matrices is very similar and therefore omitted.

| | condition number | | | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $10^{15}$ | $10^{20}$ | $10^{25}$ | $10^{28}$ | fail | $10^{29}$ | fail |
| 10 | $3.1 \cdot 10^{-15}$ | $1.9 \cdot 10^{-13}$ | $7.1 \cdot 10^{-9}$ | $5.1 \cdot 10^{-6}$ | 0 | $3.6 \cdot 10^{-6}$ | 0 |
| 20 | $5.4 \cdot 10^{-14}$ | $2.3 \cdot 10^{-13}$ | $3.2 \cdot 10^{-8}$ | $1.1 \cdot 10^{-5}$ | 1 | $1.5 \cdot 10^{-5}$ | 1 |
| 50 | $1.2 \cdot 10^{-14}$ | $1.3 \cdot 10^{-12}$ | $3.9 \cdot 10^{-8}$ | $1.8 \cdot 10^{-4}$ | 0 | $3.0 \cdot 10^{-1}$ | 92 |
| 100 | $2.4 \cdot 10^{-14}$ | $1.4 \cdot 10^{-9}$ | $1.6 \cdot 10^{-4}$ | $7.1 \cdot 10^{-2}$ | 15 | $8.2 \cdot 10^{-1}$ | 97 |

TABLE 3.4

*Test for $p = 1$, median over 100 samples, number of failures for condition numbers $10^{28}$ and $10^{29}$.*

Matrices with condition number beyond $\text{eps}^{-1}$ are outside the scope of (1.1).

| | $n = 100$ | | $n = 200$ | | $n = 500$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|---|---|
| | new | (1.1) | new | (1.1) | new | (1.1) | new | (1.1) |
| $\|\cdot\|_1$ | 0.0034 | 0.0089 | 0.0092 | 0.032 | 0.070 | 0.35 | 0.35 | 2.01 |
| $\|\cdot\|_2$ | 0.0141 | 0.0196 | 0.0276 | 0.054 | 0.158 | 0.45 | 0.93 | 2.57 |
| $\|\cdot\|_\infty$ | 0.0038 | 0.0092 | 0.0090 | 0.035 | 0.070 | 0.35 | 0.35 | 2.00 |
| $\|\cdot\|_F$ | 0.0051 | 0.0098 | 0.0089 | 0.033 | 0.065 | 0.34 | 0.34 | 1.97 |

TABLE 3.5

*Computing time in seconds new vs. (1.1)*

**4. Acknowledgment.** My dearest thanks to an anonymous referee for fruitful and valuable comments.

REFERENCES

[1] L. Collatz. Einschließungssatz für die charakteristischen Zahlen von Matrizen. *Math. Z.*, 48:221–226, 1942.
[2] J. B. Demmel. The Geometry of Ill-Conditioning. *Journal of Complexity*, 3:201–229, 1987.
[3] R.A. Horn and Ch. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
[4] R.A. Horn and Ch. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012.
[5] A. Neumaier. *Interval Methods for Systems of Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.
[6] A. Neumaier. Grand challenges and scientific standards in interval analysis. *Reliable Computing*, 8(4):313–320, 2002.
[7] S.M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999.
[8] S.M. Rump. Error estimation of floating-point summation and dot product. *BIT Numerical Mathematics*, 52(1):201–220, 2012.
[9] S.M. Rump. Accurate solution of dense linear systems, Part II: Algorithms using directed rounding. *Journal of Computational and Applied Mathematics (JCAM)*, 242:185–212, 2013.
[10] R.S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1962.