BIT manuscript No.
(will be inserted by the editor)

On a quality measure for interval inclusions

Siegfried M. Rump and Takeshi Ogita

Received: date / Accepted: date

Abstract Verification methods compute intervals which contain the solution of a given problem with mathematical rigour. In order to compare the quality of intervals some measure is desirable. We identify some anticipated properties and propose a method avoiding drawbacks of previous definitions.

Keywords Verification methods, interval, relative error, relative accuracy, INTLAB

Mathematics Subject Classification (2000) 65G40

1 Introduction

Verification methods prove that a given a numerical problem is solvable and produce mathematically rigorous error bounds for the solution of the problem. For an overview of verification methods cf. [5,8] and [in Japanese] [6].

When developing a new verification method, it is desirable to have some measure for the quality of an inclusion. We consider an inclusion interval X as error bounds for an unknown real quantity $\hat{x}$, i.e., $\hat{x} \in X$. Depending on the situation, we use synonymous notations for an inclusion interval, namely

$$\begin{aligned} X = [\underline{x}, \overline{x}] &:= \{x \in \mathbb{R} : \underline{x} \leqslant x \leqslant \overline{x}\} \\ &= \langle m, r \rangle := \{x \in \mathbb{R} : m - r \leqslant x \leqslant m + r\} \ . \end{aligned}$$

S. M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Am Schwarzenberg-Campus 3, Hamburg 21073, Germany, and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan E-mail: rump@tuhh.de

Takeshi Ogita
Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan (t.ogita@waseda.jp).

A colloquial notation is $\langle m, r\rangle = m \pm r$. Consider

$$X_1 := [-1, 2], \quad X_2 := [-1, 1], \quad \text{and} \quad X_3 := [1, 2] \ .$$

It seems that all three intervals do not give much information, only X_3 proves at least that $\hat{x}$ is positive. Now let A be a symmetric matrix with $\|A\|_2 = 10^{10}$ and let the X_ν be inclusions of an eigenvalue. Then all three inclusions X_ν reveal that the condition number $\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ of A is at least $5 \cdot 10^9$.

The quality of an interval inclusion depends on the context. Having said that, it may nevertheless be desirable to define a measure for the quality of an interval, knowing the pros and cons of such an attempt. There is some folklore about such measures, however, to that end we found only one paper in the literature, see below.

In this note we develop some criteria for such a measure. We start with some theoretical considerations in the next section, and conclude with some practical remarks.

2 Theoretical considerations

Let $\varrho : \mathbb{R}\times\mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ be a function for the quality $\varrho(m, r)$ of $\langle m, r\rangle$. The letter ϱ may remind of "relative error", however, we prefer the wording "quality" because mathematically ϱ may be interpreted as relative error, but only in a certain sense (see below). Note that $\varrho(m, r) = 0$ means best quality. We first list some desirable properties of such a function:

I)	non-negativity	$\varrho(m, r) \geqslant 0$
II)	zero value	$\varrho(m, r) = 0 \ \Leftrightarrow \ r = 0$
III)	scaling invariance	$\varrho(X) \ = \ \varrho(\alpha X) \quad \text{for } \ 0 \neq \alpha \in \mathbb{R}$
IV)	monotonicity for fixed m	$r' > r \ \Rightarrow \ \varrho(m, r') > \varrho(m, r)$
V)	monotonicity for fixed r	$\lvert m'\rvert > \lvert m\rvert \ \Rightarrow \ \varrho(m', r) < \varrho(m, r)$

The rationale is as follows. Properties I) and II) are clear. As for III), the quality of an inclusion interval X may well depend on the scaling for different settings, see the above example. However, without knowing any setting, invariance with respect to scaling seems the only option. For the monotonicity, an interval with constant midpoint but increasing radius gives less information, and with constant radius but increasing absolute value of the midpoint[1] the interval contains, in some sense, more information.

Moreover, we may demand ϱ to be continuous in m and r except for $m = r = 0$ because for $r > 0$ it follows $\varrho(0, 0) < \varrho(0, r) = \varrho(0, 1)$. As for

[1] Note that III) implies $\varrho(m, r) = \varrho(-m, r)$.

differentiability note that $\varrho(m,r) = \varrho(-m,r)$ would imply $\frac{d\varrho}{dm}(0,r) = 0$ for all $r > 0$, but then V) and I) lead to a contradiction. Therefore we require

VI) continuity $\varrho(m,r)$ is everywhere continuous except for $m = r = 0$

VII) differentiability $\varrho(m,r)$ is everywhere differentiable except for $m = 0$

Having listed the desired properties, we look for possible candidates. An obvious choice is to use the midpoint m of $X = \langle m, r\rangle$ as an approximation and define $\varrho(X)$ to be the largest relative error of $x \in X$ with respect to m:

$$\varrho_1(m,r) := \max_{x\in X}\left|\frac{x-m}{m}\right| \quad \text{implying} \quad \varrho_1(X) = \left|\frac{\overline{x}-\underline{x}}{\underline{x}+\overline{x}}\right| \ . \tag{2.1}$$

All properties I) to VII) are satisfied, however, for a small or zero unknown real quantity $\hat{x}$ the midpoint may be zero causing an obvious problem. In this case $\varrho_1(0,r)$ is infinite no matter how small the radius r is.

A remedy is to use the maximum over the minimal relative error against some $\tilde{x} \in X$, i.e.,

$$\varrho_2(X) := \min_{\tilde{x}\in X}\max_{x\in X}\left|\frac{\tilde{x}-x}{\tilde{x}}\right| \ . \tag{2.2}$$

That is the definition in [4], the only reference we found. It is shown that

$$\varrho_2(m,r) = \begin{cases} \dfrac{r}{|m|} & \text{if } |m| - r \geqslant 0 \\ \dfrac{2r}{\max(|m-r|, m+r)} & \text{otherwise} \ . \end{cases}$$

The properties I) to VI) are satisfied for ϱ_2, however, differentiability VII) is not met:

$$\varrho_2(1, 1+e) = \begin{cases} 1+e & \text{if } e \leqslant 0 \\ \dfrac{1+e}{1+e/2} & \text{if } e \geqslant 0 \ . \end{cases}$$

As has been mentioned there is some folklore about quality measures, in particular

$$\varrho_3(X) := \frac{\overline{x}-\underline{x}}{|\underline{x}| + |\overline{x}|} \tag{2.3}$$

with $0/0 := 0$. That avoids the zero midpoint problem, but for all intervals X containing zero $\underline{x} \leqslant 0 \leqslant \overline{x}$ implies

$$0 \in X: \quad \varrho_3(X) = \frac{\overline{x} + |\underline{x}|}{|\underline{x}| + \overline{x}} = 1 \ .$$

The properties I) to VI are satisfied, but ϱ_3 is not differentiable for one endpoint zero:

$$\varrho_3([0,e]) = \begin{cases} 1 & \text{if } e > 0 \\ \dfrac{e}{|e|} & \text{if } e < 0 \ . \end{cases}$$

In order to find a function ϱ sharing all properties I) to VII) but avoiding the problems for zero midpoint we use, in view of $\varrho(m,r) = \varrho(-m,r)$, the ansatz

$$\varrho(m,r) = \frac{\alpha|m| + \beta r}{\gamma|m| + \delta r}$$

for constants $\alpha, \beta, \gamma, \delta$ to be determined. Property II) implies $\alpha = 0$ and $\gamma \neq 0$, so that using III) and some scaling we can restrict our attention to

$$\varrho(m,r) = \psi \, \frac{r}{\varphi|m| + r}$$

with a scaling factor ψ defining the maximum of ϱ. Rewriting $\varrho(m,r) = \psi \left(\varphi\frac{|m|}{r} + 1\right)^{-1}$ it is easy to verify that this definition satisfies all properties I) to VII) for any $\varphi > 0$. In order to find a suitable choice for φ we look at intervals with fixed left endpoint $\underline{x} = -1$ and right endpoints $-1 \leqslant \overline{x} \leqslant 1$, that is $X_r := \langle -1 + r, r\rangle$ for $0 \leqslant r \leqslant 1$. Then

$$\varrho(X_r) = \frac{\psi r}{\varphi(1-r) + r}.$$

A good choice may be $\varphi = 1$ in which case $\varrho(X_r)$ grows linearly with r. Hence,

$$\varrho(m,r) := \frac{\psi r}{|m| + r} \ .$$

Now it is a matter of taste to fix ψ. We may feel that $\varrho([0,1]) = 1$ should hold. That implies $\psi = 2$, so that we define

$$\varrho_4(m,r) := \frac{2r}{|m| + r} \tag{2.4}$$

implying $\varrho_4(m,r) \leqslant 2$ for all m, r. For $X = [\underline{x}, \overline{x}]$ it follows

$$\varrho_4(X) = \min\left(\left|\frac{\overline{x} - \underline{x}}{\underline{x}}\right|, \left|\frac{\overline{x} - \underline{x}}{\overline{x}}\right|\right)$$

with the convention $\frac{0}{0} = 0$, the minimal relative error of the endpoints against each other. In verification methods $\mathrm{mag}(X) := \max\{|x| : x \in X\}$ is called the magnitude of an interval. Hence $\varrho_4(X) = \mathrm{diam}(X)/\mathrm{mag}(X)$. An advantage over ϱ_3 is that no case distinction is necessary in the computation. An almost identical formulation

$$\varrho_4'(X) = \frac{\overline{x} - \underline{x}}{\max(|\underline{x}|, |\overline{x}|, \eta)}$$

was suggested by Demmel [1]. It is equal to ϱ_4 except that it is tailored to binary64 of the IEEE754 [3] arithmetic standard by using the gradual underflow unit, i.e., the smallest positive floating-point number $\eta = 2^{-1074}$. If the endpoints $\underline{x}, \overline{x}$ are binary64 floating-point numbers, then $\varrho_4(X) = \varrho_4'(X)$.

In Figure 2.1 the four definitions ϱ_ν are compared for fixed midpoint $m = 1$ and for fixed left endpoint $\underline{x} = -1$.

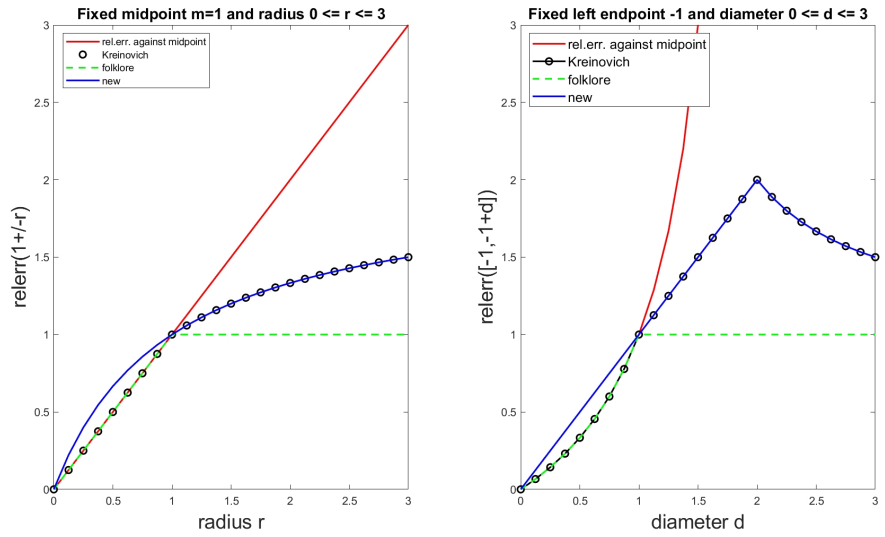

Fig. 2.1: The functions ϱ_ν for fixed midpoint $m = 1$ (left) and fixed left endpoint -1 (right)

The first function ϱ_1 [relative error against midpoint, red] shows a linear behaviour for fixed midpoint and growing radius, and tends to infinity if the midpoint approaches zero. As discussed the second function ϱ_2 [Kreinovich's definition, black with circles] it is not differentiable at $m = r$. The "folklore" function ϱ_3 [green] is not differentiable for zero endpoint and flat equal to the maximal value 1 for intervals containing zero, no discrimination in terms of small or large radius. Moreover, it is not concave. Finally, the new definition ϱ_4 [blue] is, as ϱ_1, linear for fixed midpoint and growing radius, and everywhere differentiable except for $m = 0$.

The first three definitions coincide in the left picture for $X = \langle 1, r\rangle$ with $r \in [0, 1]$, and in the right picture for $X = [-1, -1 + d]$ with $d \in [0, 1]$. In both pictures Kreinovich's definition ϱ_2 and the proposed ϱ_4 coincide for $r \geqslant 1$ and $d \geqslant 1$, respectively. So the proposed measure ϱ_4 differs from the other definitions for $r \in [0, 1]$ and $d \in [0, 1]$ in the left and right picture, respectively. This ensures differentiability everywhere except zero midpoint.

The definition $\varrho_4(X) = \frac{\operatorname{diam}(X)}{\operatorname{mag}(X)}$ with the interpretation $\frac{0}{0} = 0$ can be used for complex intervals as well. It replaced the function `relerr` in the latest Version 13 of INTLAB [7], the Matlab/Octave toolbox for reliable computing. Executable Matlab/INTLAB code is as follows:

```
function res = relerr(X)
   diamX = diam(X);
   res = diamX;
   index = find(res);          % careful with sparse input
```

```
if any(index(:))              % diam(X)./mag(X)
   magX = mag(X);
   res(index) = res(index)./magX(index);
end
res(isinf(diamX)) = 1;
```

The code is working for scalar, vector and matrix input `X`, full or sparse, real or complex. The "if"-statement takes care of $\frac{0}{0}$, and of sparse input avoiding full output.

3 Practical considerations

Our definition $\varrho_4(X)$ seems a good theoretical measure for the relative error of an interval X. However, from a practical and numerical point of view, there is a drawback. Mathematically a small $\varrho_4(Y)$ means a small *forward error*, i.e., a small relative error with respect to the true result. But numerically we can only hope for a small *backward error*, introduced and popularized by Wilkinson [11,12], see also [2]. The backward error of an approximation $\tilde{x}$ is small if $\tilde{x}$ is the true solution of the original problem after a small perturbation of the input data. Without further measurements such as a residual iteration that is about the best what we can expect.

Now consider, similar to our introductory problem, an approximation $\tilde{x} = 1.23456 \cdot 10^{-10}$ of a singular value of a matrix A with $\|A\|_2 = 1$ to the true singular value $\hat{x} = 1.23457 \cdot 10^{-10}$. Then $\varrho_4(\tilde{x} \underline{\cup} \hat{x}) = 8.1 \cdot 10^{-6}$. If computed in binary64 equivalent to some 16 decimals precision, the accuracy of $\tilde{x}$ might be considered as not bad, but far from best possible. With the additional information of the context $\|A\|_2 = 1$, however, we know that this is close to the best possible approximation we can hope for.

Therefore, from a practical and numerical point of view it seems reasonable to pass information about the context. We therefore propose a *relative accuracy* defined by

$$\alpha(X, \tau) := \frac{\mathrm{diam}(X)}{\max(\mathrm{mag}(X), \tau)} \ , \tag{3.1}$$

where τ is the context information. That implies $\alpha(X, \|A\|_2) = 10^{-15}$, a value we may expect from a practical, numerical point of view. In Version 13 of INTLAB the function `relacc` computes the relative accuracy. A typical call is

```
alpha = relacc(X,'thresh',tau);
```

The following Figure 3.1 illustrates this definition and compares it to the relative error ϱ_4. We compute approximations s_k of the singular values of a square matrix with 1000 rows and condition number 10^{12}. The well accepted rule thumb says that the approximations s_k of the smallest singular values may be correct to some 4 decimals. The dotted green line[2] in Figure 3.1 displays

[2] Relative errors zero are set to 10^{-25} to avoid gaps.

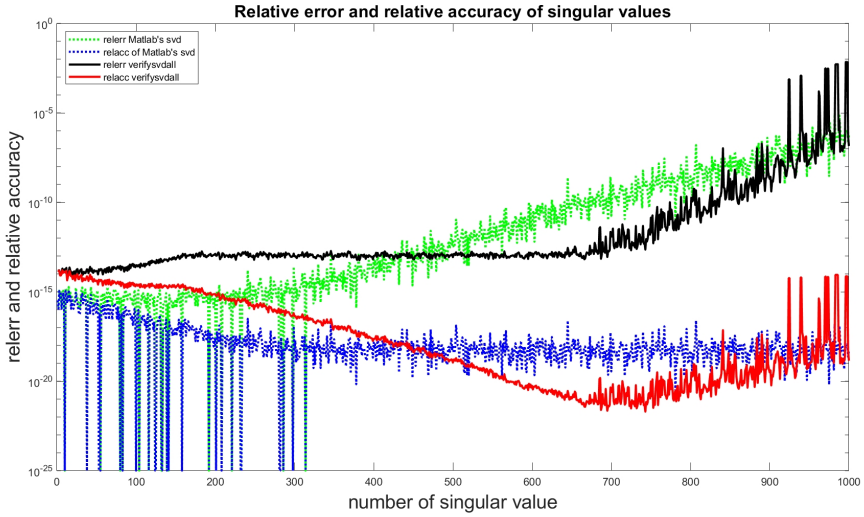

Fig. 3.1: Relative error and relative accuracy of singular value inclusions.

the values $\varrho_4(s_k \underline{\cup} \sigma_k)$, where σ_k are the true singular values of A. As expected the relative error increases from 10^{-14} for the largest to about 10^{-6} for the smallest singular values. The dotted blue line displays the relative accuracy $\alpha(X, \|A\|_2)$ and reflects what we would expect from a numerical point of view.

Additionally we use INTLAB's routine `verifysingvalall` to compute inclusions X_k of all singular values of A. The solid black line shows the relative error $\varrho_4(X)$ of the inclusions, while the solid line displays the relative accuracy $\alpha(X, \|A\|_2)$. From the black line we might conclude that the inclusions are of reasonable, but not too good quality for the smallest singular values, whereas the red line shows that the inclusions are of almost best quality for an inclusion method without extra iterative refinement. For other problems the context may be passed similarly.

We want to stress that neither the function `relerr` nor `relacc` is a panacea. As noted at the beginning of this note the judgement of the quality of an inclusion depends on the context. As an example let matrices R, A be given. Then $\|I - RA\| < 1$ for any matrix norm proves that both R and A are nonsingular. Typically, a good choice for R is an approximate inverse of A. Denote by $\mathbf{X}$ the stacked columns of an inclusion of the residual $I - RA$. As an example, we display the first and last two elements in Table 3.1 .

It is well known that one step of iterative refinement in working precision implies backward stability of the result of Gaussian elimination [9,10]. A forward stable result, i.e., an approximation with close to maximum accuracy can be achieved with residuals computed in twice the working precision.

The computed $\mathbf{X}$ may be applied in some iterative refinement. The intervals have relatively wide diameters but are small in magnitude. If that is true for

Table 3.1: Inclusion vector $\mathbf{X}$ with relative error and relative accuracy

$\mathbf{X}$	`relerr(X)`	`relacc(X,'thresh',norm(A))`
$[-1.45 \cdot 10^{-11}, 2.18 \cdot 10^{-11}]$	1.7	$3.6 \cdot 10^{-11}$
$[-9.09 \cdot 10^{-13}, 2.73 \cdot 10^{-12}]$	1.3	$3.6 \cdot 10^{-12}$
...	...	...
$[2.93 \cdot 10^{-11}, 8.00 \cdot 10^{-11}]$	0.6	$5.1 \cdot 10^{-11}$
$[0, 3.64 \cdot 10^{-12}]$	1.0	$3.7 \cdot 10^{-12}$

all entries, the wide diameters show that a residual of that quality is not suited for iterative refinement, so that `relerr` provides that information. However, the small magnitude shows that the residuals are good enough to prove that A is nonsingular, so that `relacc` provides that information.

4 Acknowledgement

The authors wish to thank the two anonymous referees for the thorough reading and fruitful comments.

5 Conflict of interest

Not applicable.

References

1. J. B. Demmel: private communication, 2012.
2. N. J. Higham: *Accuracy and Stability of Numerical Algorithms*, SIAM Publications, Philadelphia, 2nd edition, 2002.
3. IEEE standard for floating-point arithmetic, *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pages 1–84, 2019.
4. V. Kreinovich: How to define relative approximation error of an interval estimate: a proposal, *Applied Mathematical Sciences*, 7(5):211–216, 2013.
5. A. Neumaier: *Interval methods for systems of equations*. Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1990.
6. S. Oishi, K. Ichihara, M. Kashiwagi, T. Kimura, X. Liu, H. Masai, Y. Morikura, T. Ogita, K. Ozaki, S. M. Rump, K. Sekine, A. Takayasu, N. Yamanaka: *Principle of Verified Numerical Computations*, Corona Publisher, Tokyo, Japan, 2018. [in Japanese].
7. S. M. Rump: INTLAB – INTerval LABoratory, In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Springer Netherlands, Dordrecht, 1999.
8. S. M. Rump: Verification methods: Rigorous results using floating-point arithmetic, *Acta Numerica*, 19:287–449, 2010.
9. R. Skeel. Scaling for Numerical Stability in Gaussian Elimination. *Journal of the ACM*, 26(3):494–526, 1979.
10. R. Skeel. Iterative Refinement Implies Numerical Stability for Gaussian Elimination. *Math. Comp.*, 35(151):817–832, 1980.
11. J. H. Wilkinson: Error analysis of floating-point computation, *Numer. Math.*, 2:319–340, 1960.
12. J. H. Wilkinson: *Rounding Errors in Algebraic Processes*, Prentice-Hall Inc., 1963.