Overestimations in Bounding Solutions of Perturbed Linear Equations*

Jiří Rohn

Faculty of Mathematics and Physics
Charles University
and
Institute of Computer Science
Academy of Sciences
Prague, Czech Republic

Submitted by

ABSTRACT

It is proved that some classical bounds on solutions of perturbed systems of linear equations may yield arbitrarily large overestimations for arbitrarily narrow perturbations. The proofs are constructive.

---

1. INTRODUCTION

For a system of linear equations

$$Ax = b \tag{1}$$

with an $n \times n$ nonsingular matrix A, consider a family of perturbed systems

$$A'x' = b' \tag{2}$$

with data satisfying

$$|A' - A| \leq \Delta \tag{3}$$

*This work was supported by the Czech Republic Grant Agency under grant GAČR 201/95/1484

655 Avenue of the Americas, New York, NY 10010 0024- 3795/94/$6.00

and

$$|b' - b| \le \delta, \tag{4}$$

where $\Delta \ge 0$ and $\delta \ge 0$ are $n \times n$ perturbation matrix and perturbation n-vector, respectively. Here, the absolute value of a matrix $A = (a_{ij})$ is defined by $|A| = (|a_{ij}|)$ and the inequalities are understood componentwise; the same notation applies to vectors as well. The classical numerical argument using Neumann series shows that if the condition

$$\varrho(|A^{-1}|\Delta) < 1 \tag{5}$$

is met (where ϱ stands for the spectral radius), then each A' satisfying (3) is nonsingular and the solution of each system (2) with data (3), (4) satisfies

$$|x' - x| \le d, \tag{6}$$

where d is an n-vector defined by

$$d = (I - |A^{-1}|\Delta)^{-1}|A^{-1}|(\Delta|x| + \delta) \tag{7}$$

and I is the unit matrix (see Skeel [5] or Rump [4]). To keep the paper self-contained, we give here another simple proof of this result: for the solutions x, x' of (1), (2) under (3), (4) we have

$$\begin{aligned} |x' - x| &= |A^{-1}A(x' - x)| \\ &\le |A^{-1}| \cdot |(A - A')(x' - x) + (A - A')x + b' - b| \\ &\le |A^{-1}|(\Delta|x' - x| + \Delta|x| + \delta). \end{aligned}$$

Here, as before, the inequalities hold componentwise. Hence

$$(I - |A^{-1}|\Delta)|x' - x| \le |A^{-1}|(\Delta|x| + \delta),$$

and premultiplying this inequality by $(I - |A^{-1}|\Delta)^{-1}$, which is nonnegative in view of (5), we obtain (6), where d is given by (7).

The quality of the estimation (6) has been paid little attention in the literature. Obviously, the bound d is exact if $\Delta = 0$. In fact, in this case, for each $i \in \{1, \ldots, n\}$, if we take $b'_j = b_j + \delta_j$ if $(A^{-1})_{ij} \ge 0$ and $b'_j = b_j - \delta_j$ otherwise, then b' satisfies (4) and for the solution x' of $Ax' = b'$ we have

$$|x'_i - x_i| = \sum_j |A^{-1}|_{ij}\delta_j = d_i,$$

hence the bound is achieved. However, this argument fails in the case $\Delta \ne 0$. In this paper we show that for each $n \ge 4$ and for arbitrary positive real numbers ε, ζ and α we may construct $n \times n$ matrices A, $\Delta \geq 0$ and n-vectors b, $\delta \geq 0$ such that

$$\|\Delta\|_{1,\infty} := \max_{i,j} |\Delta_{ij}| = \varepsilon,$$

$$\|\delta\|_\infty := \max_i |\delta_i| = \zeta$$

hold and the solution x' of each system (2) with data (3), (4) satisfies

$$|x'_1 - x_1| + \alpha \leq d_1,$$

where d is given by (7) (section 2, Theorem 1). Hence, the formula (6) may yield an arbitrarily large overestimation α for arbitrarily narrow perturbations ε, ζ.

In numerical linear algebra, normwise estimations are preferred to the componentwise ones. For each absolute norm $\|\cdot\|$ (i.e., satisfying $\||x|\| = \|x\|$ for each x; such a norm has the property $|x| \leq |y| \Rightarrow \|x\| \leq \|y\|$, see Higham [2]), the componentwise estimation (6) yields the normwise estimation

$$\|x' - x\| \leq \|d\|. \tag{8}$$

In Theorem 2 of section 3 we prove an analogous result for normwise overestimations: for each $n \geq 4$ and arbitrary positive real numbers ε, ζ and α satisfying an additional assumption

$$\frac{1}{2}\zeta \leq \alpha$$

we may construct $n \times n$ matrices A, Δ and n-vectors b, δ satisfying $\|\Delta\|_{1,\infty} = \varepsilon$, $\|\delta\|_\infty = \zeta$ (in fact, the same data as in the proof of Theorem 1) such that

$$\|x' - x\|_1 + \alpha \leq \|d\|_1,$$

$$\|x' - x\|_\infty + \alpha \leq \|d\|_\infty$$

and

$$\|x' - x\|_2^2 + \alpha^2 \leq \|d\|_2^2$$

hold for the solution x' of each system (2) with data satisfying (3), (4) (where, as usual, $\|x\|_1 = \sum_i |x_i|$, $\|x\|_\infty = \max_i |x_i|$ and $\|x\|_2 = \sqrt{x^T x}$). Hence again, an arbitrarily large normwise overestimation may occur for arbitrarily narrow perturbations.

These results show that formulae (6), (8) should be used with some care.

2. COMPONENTWISE OVERESTIMATIONS

For an integer $n \geq 2$, denote by I the $(n-1) \times (n-1)$ unit matrix and let
$$E = ee^T,$$
where $e = (1, \ldots, 1)^T \in R^{n-1}$; hence, E is the $(n-1) \times (n-1)$ matrix of all ones. For given positive real numbers ε, ζ and α, define $n \times n$ matrices A, Δ and n-vectors b, δ by

$$A = \begin{pmatrix} \frac{\varepsilon\zeta}{\alpha} & 0^T \\ 0 & \frac{1}{n}(I+E) \end{pmatrix}, \tag{9}$$

$$\Delta = \begin{pmatrix} 0 & \varepsilon e^T \\ 0 & 0 \end{pmatrix}, \tag{10}$$

$$b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{11}$$

$$\delta = \begin{pmatrix} 0 \\ \zeta e \end{pmatrix}. \tag{12}$$

This definition implies that A, Δ, b and δ are all nonnegative and that

$$\|\Delta\|_{1,\infty} = \varepsilon, \tag{13}$$

$$\|\delta\|_\infty = \zeta \tag{14}$$

hold. Moreover, we have
$$\varrho(|A^{-1}|\Delta) = 0. \tag{15}$$

In fact, from $E^2 = (n-1)E$ it follows

$$\frac{1}{n}(I+E)(nI-E) = I,$$

hence
$$\left(\frac{1}{n}(I+E)\right)^{-1} = nI - E, \tag{16}$$

which implies
$$A^{-1} = \begin{pmatrix} \frac{\alpha}{\varepsilon\zeta} & 0^T \\ 0 & nI-E \end{pmatrix},$$

$$|A^{-1}| = \begin{pmatrix} \frac{\alpha}{\varepsilon\zeta} & 0^T \\ 0 & (n-2)I+E \end{pmatrix} \tag{17}$$

and

$$|A^{-1}|\Delta = \begin{pmatrix} 0 & \frac{\alpha}{\zeta}e^T \\ 0 & 0 \end{pmatrix}, \tag{18}$$

hence (15) holds. The following theorem is our main result for componentwise overestimations:

THEOREM 1. *Let $n \geq 4$, let ε, ζ and α be arbitrary positive real numbers and let A, Δ, b, δ be given by (9)–(12). Then (13)–(15) hold and for the solution x' of each system (2) with data satisfying (3), (4) we have*

$$|x'_1 - x_1| + \alpha \leq d_1, \tag{19}$$

where x is the solution of (1) and d is given by (7).

Proof. Let $|A' - A| \leq \Delta$, $|b' - b| \leq \delta$. Then the system $A'x' = b'$ can be equivalently written in the form

$$\frac{\varepsilon\zeta}{\alpha}x'_1 + a^T\tilde{x} = 0, \tag{20}$$

$$-\zeta e \leq \frac{1}{n}(I + E)\tilde{x} \leq \zeta e, \tag{21}$$

where $\tilde{x} = (x'_2, \ldots, x'_n)^T \in R^{n-1}$ and $a^T = (A'_{12}, \ldots, A'_{1n})$ satisfies $|a| \leq \varepsilon e$. Hence for the quantity

$$\overline{x}_1 := \max\{|x'_1|;\ x' \text{ solves (2) under (3), (4)}\}$$

we have from (20), (21) that

$$\overline{x}_1 = \frac{\alpha}{\varepsilon\zeta}\max\{\varepsilon e^T|\tilde{x}|;\ -\zeta e \leq \frac{1}{n}(I + E)\tilde{x} \leq \zeta e\}.$$

Put

$$\hat{x} = \frac{1}{\zeta n}(I + E)\tilde{x},$$

then we have $\tilde{x} = \zeta(nI - E)\hat{x}$ due to (16), hence

$$\overline{x}_1 = \alpha\max\{\|(nI - E)\hat{x}\|_1;\ -e \leq \hat{x} \leq e\}. \tag{22}$$

In view of convexity of the norm the maximum in (22) is achieved at some of the vertices of the hyperrectangle $\{\hat{x};\ -e \leq \hat{x} \leq e\}$, which are exactly the points satisfying $|\hat{x}| = e$ (i.e., the ± 1-vectors). Hence (22) implies

$$\overline{x}_1 = \alpha\max\{\|(nI - E)\hat{x}\|_1;\ |\hat{x}| = e\}. \tag{23}$$

Now, since each ± 1-vector $\hat{x} \in R^{n-1}$ satisfies

$$|e^T \hat{x}| \leq e^T e = n - 1,$$

for each $i \in \{1, \ldots, n-1\}$ we have

$$\hat{x}_i((nI - E)\hat{x})_i = n - \hat{x}_i(e^T \hat{x}) \geq 1 > 0,$$

hence

$$\begin{aligned} \|(nI - E)\hat{x}\|_1 &= \sum_i |(nI - E)\hat{x}|_i = \sum_i \hat{x}_i((nI - E)\hat{x})_i \\ &= \hat{x}^T (nI - E)\hat{x} = n(n-1) - (e^T \hat{x})^2 \end{aligned}$$

and from (23) we get

$$\overline{x}_1 = \alpha n(n-1) - \alpha \min\{(e^T \hat{x})^2; \, |\hat{x}| = e\},$$

hence

$$\overline{x}_1 = \alpha n(n-1) \tag{24}$$

if n is odd and

$$\overline{x}_1 = \alpha(n(n-1) - 1) \tag{25}$$

if n is even, in both cases

$$\overline{x}_1 \leq \alpha n(n-1). \tag{26}$$

Let us now compute d_1. Since

$$\begin{pmatrix} 1 & -\frac{\alpha}{\zeta} e^T \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \frac{\alpha}{\zeta} e^T \\ 0 & I \end{pmatrix}$$

and since $x = 0$ due to $b = 0$, from (7) using (18), (17) we obtain

$$\begin{aligned} d &= \begin{pmatrix} 1 & \frac{\alpha}{\zeta} e^T \\ 0 & I \end{pmatrix} \begin{pmatrix} \frac{\alpha}{\varepsilon\zeta} & 0^T \\ 0 & (n-2)I + E \end{pmatrix} \begin{pmatrix} 0 \\ \zeta e \end{pmatrix} \\ &= \begin{pmatrix} \alpha(2n-3)(n-1) \\ \zeta(2n-3)e \end{pmatrix}, \end{aligned} \tag{27}$$

hence

$$d_1 = \alpha(2n-3)(n-1). \tag{28}$$

Since

$$n(n-1)+1 \leq (2n-3)(n-1) \tag{29}$$

holds for each $n \geq 4$ (as it can be easily verified), from (26), (28) and (29) we finally obtain

$$\overline{x}_1 + \alpha \leq d_1. \tag{30}$$

Hence for the solution x' of each system (2) with data satisfying (3), (4) we have

$$|x'_1 - x_1| + \alpha = |x'_1| + \alpha \leq \overline{x}_1 + \alpha \leq d_1,$$

which is (19) and the proof is complete. ■

3. NORMWISE OVERESTIMATIONS

In this section we show that the componentwise overestimation result of Theorem 1 can be given a normwise overestimation form provided any of the three most frequently used vector norms $\|\cdot\|_1$, $\|\cdot\|_\infty$ or $\|\cdot\|_2$ is used.

THEOREM 2. *Let $n \geq 4$, let ε, ζ and α be arbitrary positive real numbers satisfying*

$$\frac{1}{2}\zeta \leq \alpha, \tag{31}$$

and let A, Δ, b, δ be given by (9)– (12). Then (13)–(15) hold and for the solution x' of each system (2) with data satisfying (3), (4) we have

$$\|x' - x\|_1 + \alpha \leq \|d\|_1, \tag{32}$$

$$\|x' - x\|_\infty + \alpha \leq \|d\|_\infty \tag{33}$$

and

$$\|x' - x\|_2^2 + \alpha^2 \leq \|d\|_2^2, \tag{34}$$

where x is the solution of (1) and d is given by (7).

Proof. Define $\overline{x} = (\overline{x}_j)$ by

$$\overline{x}_j := \max\{|x'_j|;\, x' \text{ solves (2) under (3), (4)}\}$$

$(j = 1, \ldots, n)$. Formulae for $\overline{x}_1$ were given in (24), (25). For $j \geq 2$ we obtain from (21)

$$\begin{aligned} \overline{x}_j &= \max\{\tilde{x}_j;\, -\zeta e \leq \frac{1}{n}(I+E)\tilde{x} \leq \zeta e\} \\ &= \max\{((nI - E)\hat{x})_j;\, -\zeta e \leq \hat{x} \leq \zeta e\} = (2n-3)\zeta. \end{aligned} \tag{35}$$

Since

$$\frac{2n-3}{n^2-n-1} \leq \frac{1}{2}$$

holds for $n \geq 4$, we have

$$\overline{x}_j = (2n-3)\zeta \leq \frac{1}{2}(n^2-n-1)\zeta \leq \alpha(n^2-n-1) \leq \overline{x}_1$$

for each $j \geq 2$ due to (31) and (24), (25), which gives

$$\overline{x}_1 = \max_j \overline{x}_j. \tag{36}$$

Next, (27) and (31) imply

$$d_j = (2n-3)\zeta \leq (2n-3)2\alpha \leq (2n-3)(n-1)\alpha = d_1$$

for $j \geq 2$, hence also

$$d_1 = \max_j d_j. \tag{37}$$

Taking into account the inequality

$$\overline{x}_1 + \alpha \leq d_1 \tag{38}$$

established in the previous proof (eq. (30)) and the fact that

$$\overline{x}_j = d_j \tag{39}$$

holds for $j \geq 2$ ((35), (27)), from (36)–(39) we obtain that

$$\|\overline{x}\|_p + \alpha \leq \|d\|_p$$

is valid for $p = 1$ or $p = \infty$. Hence for the solution x' of each system (2) with data satisfying (3), (4) we have

$$\|x' - x\|_p + \alpha = \|x'\|_p + \alpha \leq \|\overline{x}\|_p + \alpha \leq \|d\|_p$$

for $p \in \{1, \infty\}$, which proves (32) and (33). Next, (38) and (39) imply

$$\|\overline{x}\|_2^2 + \alpha^2 \leq \|d\|_2^2$$

and again

$$\|x' - x\|_2^2 + \alpha^2 = \|x'\|_2^2 + \alpha^2 \leq \|\overline{x}\|_2^2 + \alpha^2 \leq \|d\|_2^2,$$

which is (34). ■

4. CONCLUDING REMARKS

We have proved that the classical formulae (6), (8) may yield arbitrarily large overestimations for arbitrarily narrow perturbations. This, of course, is a worst-case result relying heavily on the special form of the data (9)–(12). In particular, perturbations affect zero coefficients only, a situation which is very unlikely to happen in practical applications. Nevertheless, the results show that the formulae (6), (8) should be used with some care.

5. ACKNOWLEDGEMENT

In the first version of this paper, a weaker version of Theorem 1 (as the sole result) was proved in a nonconstructive way under the assumption of validity of the famous conjecture “P≠NP”, where P and NP are the complexity classes (see Garey and Johnson [1]). The proof, rather complicated, was based on the recently established fact that computing the subordinate matrix norm $\|A\|_{\infty,1}$ is NP-hard [3]. The anonymous referee of this paper suggested that the author try to find an unconditional result not relying on the conjecture “P≠NP”. Hence the paper in its present form owes much to the referee whose contribution is gladly acknowledged.

REFERENCES

1 M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

2 N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

3 J. Rohn, NP-hardness results for some linear and quadratic problems, Technical Report No. 619, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague 1995, 11 p.

4 S. M. Rump, Bounds for the componentwise distance to the nearest singular matrix, to appear in *SIAM J. Matr. Anal. Appl.*

5 R. Skeel, Iterative refinement implies numerical stability for Gaussian elimination, *Math. of Comp.* 35:817–832 (1980).