Iterative solution of nonlinear systems of equations
Lecture Notes in Math. 953, Springer 1982, pp. 88-105

SIMPLE BOUNDS FOR ZEROS OF SYSTEMS OF EQUATIONS

Arnold Neumaier

Institut für Angewandte Mathematik

Universität Freiburg

D-7800 Freiburg, W-Germany

Abstract. Let $F:D(\subseteq \mathbb{R}^n) \to \mathbb{R}^n$ be a continuous function, and suppose that for some $x_0 \in D$ and some nonsingular matrix A the vector $\delta_0 := A^{-1} F(x_0)$ is "small". Assuming the existence of a nonnegative vector $c \in \mathbb{R}^n$ such that

$$|F(x) - F(x_0) - A(x - x_0)| \leq \|\delta_0\| c$$

for all x in a suitable neighbourhood S of x_0, a simple condition is given which guarantees that S contains a zero $\hat{x}$ of F. The resulting bounds are shown to be quite accurate. They contain as a special case the bounds obtainable from a theorem of Kantorovic.

It is discussed how to compute the required vector c, and how to make efficient use of sparsity. The practical use of the bounds is demonstrated by an extensive example, the finite difference equations obtained by discretizing the minimal surface equation.

1. Introduction

Let D be a subset of $\mathbb{R}^n$, and let $F:D \to \mathbb{R}^n$ be a bounded continuous function. We assume the following:

(i) A good approximation $x_0 \in D$ for an isolated zero $\hat{x}$ of F(x) is known.

(ii) An approximation A for the Fréchet-derivative or a generalized slope operator near x_0 is known.

(iii) A is nonsingular, and the structure of A allows the practical solution of linear equations with coefficient matrix A by Gauss elimination.

With these assumptions, we ask for a bound on the magnitude of the error $\hat{x}-x_0$. The bound shall be such that we can take advantage of any special structure of A such as diagonal dominance or sparseness.

Assuming (i), (ii), and (iii), we have for x near x_0

$$F(x) - F(x_0) \approx A(x - x_0), \tag{1}$$

hence, for $x = \hat{x}$ with $F(\hat{x}) = 0$,

$$\hat{x} - x_0 \approx -A^{-1} F(x_0).$$

Thus, if we put

$$\delta_0 := A^{-1} F(x_0), \tag{2}$$

we have in any norm $\|\hat{x}-x_0\| \approx \|\delta_0\|$, and we can expect an error estimate

$$\|\hat{x} - x_0\| \leq \kappa \|\delta_0\|$$

with a reasonably small factor $\kappa > 1$. To guarantee $\hat{x} \in D$ we assume that

$$S := \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq \kappa \|\delta_0\|\} \subseteq D, \tag{3}$$

and to make (1) precise we require that

$$|F(x) - F(x_0) - A(x - x_0)| \leq \|\delta_0\| c \quad \text{for all } x \in S. \tag{4}$$

Hence c is a nonnegative vector; absolute value and order is understood to be componentwise.

It turns out that (2), (3), and (4), together with a simple inequality are sufficient to show that S must contain at least one zero of F(x). The precise statement and its proof is contained in Section 2. Section 3 discusses the practical computation of the quantities involved in the bound. In Section 4 we compare our bounds with those of the Kantorovic theorem, and in Section 5 we present some numerical examples.

The concepts used in this paper can be found, e.g. in the book by Ortega and Rheinboldt [12].

2. The main result

For the moment we shall forget about the heuristic introduction, and suppose that D is a subset of $\mathbb{R}^n$, $F:D \to \mathbb{R}^n$ is a continuous function, $x_0 \in D$ is a vector, A is a nonsingular $n\times n$-matrix, and

$$\delta_0 := A^{-1} F(x_0). \tag{2}$$

We suppose further that there is a constant $\kappa > 1$ such that

$$S := \{x \in \mathbb{R}^n \mid \|x - x_0\| \le \kappa \|\delta_0\|\} \subseteq D, \tag{3}$$

and a nonnegative vector $c \in \mathbb{R}^n$ such that, for some monotone norm $\|\cdot\|$

$$|F(x) - F(x_0) - A(x - x_0)| \le \|\delta_0\| c \quad \text{for all } x \in S. \tag{4}$$

Then the following theorem holds.

Theorem

If the vector

$$b := |A^{-1}|c \tag{5}$$

satisfies the condition

$$\|b\| \le \kappa - 1 \tag{6}$$

then $F(x)$ has at least one zero in S, and any such zero $\hat{x}$ satisfies

$$(2 - \kappa)\|\delta_0\| \le \|\hat{x} - x_0\| \le \kappa \|\delta_0\| . \tag{7}$$

Proof. We show that the map $\phi:D \to \mathbb{R}^n$ defined by

$$\phi(x) := x - A^{-1} F(x)$$

maps S into itself. Indeed, if $x \in S$ then

$$\begin{aligned} |\phi(x) - x_0| &= |x - x_0 - A^{-1} F(x)| \\ &= |- A^{-1} (F(x) - F(x_0) - A(x - x_0)) - A^{-1} F(x_0)| \\ &\le |A^{-1}|\cdot\|\delta_0\| c + |A^{-1} F(x_0)| \\ &= \|\delta_0\| b + |\delta_0| \end{aligned}$$

by (4), (5), and (2), hence

$$\|\phi(x) - x_0\| \le \|\delta_0\| \|b\| + \|\delta_0\| \le \kappa \|\delta_0\|$$

since $\|\cdot\|$ is monotone and by (6). Hence $\phi(x) \in S$. By Brouwer's fixed point theorem, S contains a fixed point $\hat{x}$ of ϕ, and clearly $\hat{x}$ satisfies $F(\hat{x}) = 0$.

Now for any zero $\hat{x} \in S$ of $F(x)$, the right hand inequality of (7) is obvious. To prove the left hand inequality, we insert $x = \hat{x}$, $F(\hat{x}) = 0$ into (4) and obtain

$$|F(x_0) + A(\hat{x} - x_0)| \le \|\delta_0\| c,$$

hence

$$\begin{aligned} |\hat{x} - x_0| &= |\delta_0 - A^{-1}(F(x_0) + A(\hat{x} - x_0))| \\ &\ge |\delta_0| - |A^{-1}||F(x_0) + A(\hat{x} - x_0)| \\ &\ge |\delta_0| - |A^{-1}|\cdot\|\delta_0\| c \\ &= |\delta_0| - \|\delta_0\| b, \end{aligned}$$

and therefore

$$\|\hat{x} - x_0\| \ge \|\delta_0\| - \|\delta_0\| \|b\| \ge (2 - \kappa)\|\delta_0\| . \tag{8}$$

This completes the proof. □

To analyze the quality of the bound and the influence of the various quantities involved we adopt again the heuristic attitude of Section 1. Two important questions arise immediately:

(i) How restrictive is the condition (6) needed for the validity of (7)?

(ii) How good are the bounds (7)?

Obviously, (6) is satisfied if either c and hence b is sufficiently small, or if κ is sufficiently large.

Suppose first that κ is fixed. How small may we choose c? To answer this, we assume that $F(x)$ has a Fréchet-derivative $F'(x_0)$ in x_0, and $A = F'(x_0)$. Then $|F(x) - F(x_0) - A(x - x_0)| = o(\|x - x_0\|) = o(\|\delta_0\|)$ for $x \in S$, hence we may choose c arbitrarily small provided that $\|\delta_0\|$ is small enough. If the second derivative exists, too, then we even have $|F(x) - F(x_0) - A(x - x_0)| = O(\|\delta_0\|^2)$, hence $c = O(\|\delta_0\|)$. If A is close to $F'(x_0)$ then c has to be increased by a correction of order $\|A - F'(x_0)\|$. Thus we can satisfy (6) provided that δ_0 is sufficiently small (i.e. x_0 is sufficiently close to a zero $\hat{x}$) and A is sufficiently close to $F'(x_0)$.

Now we suppose that c is kept fixed. If $F(x)$ is differentiable then $|F(x) - F(x_0) - A(x - x_0)|$ will grow at least like a multiple of $\|x - x_0\|$, hence (4) implies that $\|x - x_0\|$ may not increase indefinitely. Therefore, the bounding factor κ in (3) must be kept sufficiently small.

Thus, if we increase κ, the vector c will increase as well. Unless F(x) is linear or almost linear, c will increase much faster than κ (for example $c = O(\kappa^2)$ if the second derivative exists). Therefore, κ has to be kept fairly small.
Of course there must be cases when no choice of c and κ will make (6) valid since F(x) may be nonzero in every set S of shape (3).

To estimate the accuracy of the bound (7) we compare the lower and upper bound. If $\kappa < 2$, then (7) shows that the upper bound overestimates the true error by a factor of at most

$$\mu = \frac{\kappa}{2-\kappa} ,$$

Thus for $\kappa = 3/2$ the true error is catched within a factor of $\mu = 3$.
If $\kappa > 2$, an overestimation factor of

$$\mu' = \frac{\kappa}{1-\|b\|}$$

can be derived from (8) provided that $\|b\| < 1$. But in this case (6) is satisfied for some $\kappa < 2$ as well. Hence from the point of view of accurate estimation, there is no incentive to use a factor $\kappa > 2$. On the other hand, there are cases when (6) can be satisfied with some $\kappa > 2$ but not with $\kappa \leq 2$ (see Section 4).

Remark
The argument leading to (8) can be adapted to give the componentwise error bound

$$|\bar{x} - x_1| \leq \|\delta_0\| b \quad \text{for } x_1 := x_0 - \delta_0 . \tag{9}$$

Thus, if the entries of b differ much in magnitude, it can be expected that some components of x_1 are much more accurate than others. (This may be an indication that the problem is badly scaled).

3. Practical aspects

In this section we discuss the choice of the parameters and the practical determination of the bound. For the sake of a simple presentation we assume that F(x) is (Fréchet-) differentiable.

3.1. Choice of x_0 and A. For good bounds, x_0 should be an approximation of a zero of F(x). From (6) we see that this essentially means that δ_0 is kept small. Thus it is sensible to determine x_0 from a quasi-Newton iteration.

$$z_0 \text{ suitable, } z_{i+1} := z_i - \sigma_i A_i^{-1} F(z_i),\ A_i \approx F'(z_i). \tag{10}$$

If $A_m^{-1} F(z_m)$ is sufficiently small we put

$$x_0 := z_m,\ A := A_m,\ \delta_0 := A_m^{-1} F(z_m). \tag{11}$$

But it should be kept in mind that the bound (7) does not depend on this choice of x_0 and A. For example, x_0 might be the solution of a previously solved near by system $\tilde{F}(x) = 0$, or (in particular if $F'(x_0)$ is an M-matrix), A might be the lower triangle of $F'(x_0)$.

3.2. Choice of κ. We already commented on the need of keeping κ small, preferably $\kappa < 2$. If x_0 is indeed close to a zero of F(x), and $A = F'(x_0)$ then $\kappa = 3/2$ works quite well since $c = O(\|\delta_0\|)$. But if $F'(x_0)$ is ill-conditioned, or if A is only a crude (e.g. triangular) approximation of $F'(x_0)$ then κ should be taken larger, e.g. $\kappa = 10$. If this does not work, a better approximation for $F'(x_0)$ or a better approximation x_0 of the zero is required.

3.3. Determination of c. There are various ways of finding a vector c satisfying (4).
Perhaps the most straightforward approach is the following. Suppose that the expression

$$F(x_0 + w) - F(x_0) - Aw \tag{12}$$

which comes from (4) by replacing x by x_0+w can be rewritten into an analytically equivalent expression each term of which is of order O(w) or smaller. The resulting expression can be bounded by eliminating w with $\|w\| \leq \kappa \|\delta_0\|$, using inequalities like the triangle inequality. If this is possible, it gives very good values for c; an example is given in Section 5.
Alternatively, the expression can be evaluated in interval arithmetic (see e.g. Moore [8]), using in place of w an interval vector containing all w with $\|w\| \leq \kappa \|\delta_0\|$. Interval arithmetic has the advantage that rounding errors are automatically taken into account (cf. 3.5), but at present most interval arithmetic implementations

are rather slow (cf., however, Moore [9]).

Another approach makes use of a technique of Alefeld [1], and can be applied if F(x) is a so-called polynomial operator. Alefeld shows how to find (with interval arithmetic) two matrices $\underline{A}$ and $\overline{A}$ such that for every $x \in S$,

$$F(x) - F(x_0) = A_x(x - x_0) \tag{13}$$

for a suitable matrix A_x with

$$\underline{A} \le A_x \le \overline{A}; \tag{14}$$

then (4) amounts to finding a vector c such that

$$|(A_x - A)(x - x_0)| \le \|\delta_0\| c \quad \text{for all } x \in S. \tag{15}$$

If a scaled l_∞ - norm is used,

$$\|x\| = \|x\|_u := \max_{1 \le i \le n} \frac{|x_i|}{|u_i|} = \min \{\alpha \ge 0 \mid |x| \le \alpha u\}, \tag{16$_\infty$}$$

where $u = (u_1, \ldots, u_n) > 0$, then $x \in S$ implies $|x - x_0| \le \kappa \|\delta_0\| u$, whence $|(A_x - A)(x - x_0)| \le \kappa \|\delta_0\| \max(|\overline{A} - A|u, |\underline{A} - A|u)$; therefore (15) holds with

$$c := \kappa \cdot \max(|\overline{A} - A|u, |\underline{A} - A|u). \tag{17$_\infty$}$$

A similar, slightly more complicated argument shows that for the Euclidean norm

$$\|x\| = \|x\|_2 = \sqrt{x^T x} \tag{16$_2$}$$

we can satisfy (15) with

$$c := \kappa \cdot (v_1, \ldots, v_n)^T, \tag{17$_2$}$$

where v_i is the maximum of the Euclidean norms of the i-th rows of $\overline{A}$-A and $\underline{A}$-A.

Slightly less accurate estimates are obtainable from lower and upper bounds of the derivative,

$$\underline{A} \le F'(x) \le \overline{A} \quad \text{for all } x \in S; \tag{18}$$

then (13) and (14) hold with

$$A_x = \int_0^1 F'(x_0 + t(x - x_0))\, dt,$$

whence again (17) can be used to compute c.

A further possibility to get an expression for c is to use bounds for the norm of the second derivative F"(x); we do not recommend the use of F"(x) and hence give no details (but cf. Section 4).

3.4. Determination of b. We assume that a triangular decomposition

$$A = LR \tag{19}$$

of A is already available (from the computation of δ_0). The straightforward way to compute $b = |A^{-1}|c$ is to form A^{-1}, take absolute values, and multiply with c. The formation of A^{-1} at least triples the work needed in (19), but if A is sparse the work is much more (e.g. for tridiagonal A: $O(n)$ for (19), $O(n^2)$ for A^{-1}). Thus we look for ways to avoid the formation of A^{-1}.

The simplest case is when A is an M-matrix; then A^{-1} is nonnegative, and b can be computed as the solution of the doubly triangular system

$$LRb = c. \tag{20}$$

In general, if A^{-1} has not a constant sign, b cannot be formed without explicit knowledge of A^{-1}. Fortunately we only need an upper bound for the norm of b, hence the computation of an upper bound $\overline{b}$ of b is sufficient. It is shown in Neumaier [16] that such a bound is given by the solution of the system

$$\langle L\rangle\langle R\rangle \overline{b} = c, \tag{21}$$

where the Ostrowski operator $\langle\cdot\rangle$ replaces a matrix $M = (m_{ik})$ by

$$\langle M\rangle := (\bar{m}_{ik}) \text{ with } \bar{m}_{ik} = \begin{cases} |m_{ik}| & \text{if } i = k \\ -|m_{ik}| & \text{otherwise.} \end{cases} \tag{22}$$

If we write

$$\beta := \|b\|, \quad \overline{\beta} := \|\overline{b}\|$$

then $\beta \le \overline{\beta}$, hence the condition

$$\overline{\beta} \le \kappa - 1 \tag{23}$$

implies the hypothesis of our theorem.

If A is an M-matrix then the results of [6] imply that $\langle L\rangle = L$ and $\langle R\rangle = R$ whence $\overline{\beta} = \beta$. For other matrices, it is of critical importance to known how much $\overline{\beta}$ overestimates β. Examples show that $\overline{\beta}/\beta$ can grow exponentially with the dimension n, but for diagonally dominant matrices the situation is much better, see [10]. As an illustration, the following table gives the range of $\log_2(\overline{\beta}/\beta)$ for certain dense 20x20 matrices (band matrices behave slightly better). We chose constant entries for c, and randomly generated matrices B with unit diagonal and off-diagonal entries uniformly distributed in [-max,max]. For each value of max $\in$ {5.0, 1.0, 0.25, 0.1}, 40 such matrices were generated, and $\overline{\beta}/\beta$ was calculated for A=B, A=|B|, and A = $\langle B\rangle$, respectively.

max	5.0	1.0	0.25	0.1
A = B	15...21	12...18	2...4	$\leq$ 1.5
A = |B|	16...21	12...19	2...4	$\leq$ 1.5
A = <B>	15...23	9...16	7...12	$\leq$ 0.5

Table 1. Range of $\log_2 (\overline{\beta}/\beta)$

An estimate for β which usually is smaller than β can be obtained by using $\beta = \|b\| \leq \|A^{-1}\| \, \|c\|$ and replacing $\|A^{-1}\|$ by a lower bound for $\|A^{-1}\|$. Several such lower bounds have been proposed, all based on $\|A^{-1}\| \geq \|x\| / \|Ax\|$, with x suitably chosen to maximize the lower bound (see [3], [4], [11]). While such an estimate $\underline{\beta}$ cannot be used for bounding the error $\|\tilde{x}-x_0\|$, it is of value in assessing whether a computed $\overline{\beta}$ is a realistic bound for β.

3.5. Control of rounding errors. Due to rounding errors, the quantities δ_0, c, b, and β will usually be calculated only approximately. Then the bounds (7) and (9) are valid only approximately, and if $\|b\|$ is close to $\kappa-1$, the test (6) on which (7) depends may be spoilt by rounding errors. Nevertheless, if A is not ill-conditioned, $\|b\|$ and $\|\delta_0\|$ will have the right order of magnitude, whence

$$\|b\| << \kappa - 1 \tag{6a}$$

is still an indication that (7) holds.
If reliable bounds are required we may use rounded interval arithmetic (Moore [8]) which automatically takes account of the rounding errors made by working with the intervals of uncertainly of each variable. The solution of linear interval equations is easiest if A is diagonally dominant or at least an H- matrix, for then the interval version of Gauss elimination can be carried out without problems (cf. Alefeld [1]), moreover the system (21) can be programmed in the equivalent but simpler form

$$[-\overline{b},\overline{b}] := \text{solution of } LRz = [-c,c]. \tag{21a}$$

4. The Kantorovic bound

In this section we compare our bounds with the a priori bound which is part of the famous theorem of Kantorovic. This theorem gives conditions which guarantee that Newton's iteration converges to a zero of a function F(x). As a byproduct, the following bounds for a zero are obtained (cf. e.g. Ortega and Rheinboldt [12]).

Proposition (Kantorovic)
Let D be a subset of $\mathbb{R}^n$, and suppose that the function $F:D \to \mathbb{R}^n$ has in D a continuous Fréchet derivative F'(x). Assume that for some $x_0 \in D$, $F'(x_0)$ is nonsingular, and

$$\|F'(x_0)^{-1} F(x_0)\| \leq \alpha \neq 0, \tag{24}$$

$$\|F'(x_0)^{-1}\| \leq \beta ,$$

$$\|F'(x) - F'(y)\| \leq \gamma \|x - y\| \quad \text{for } x, y \in S_0,$$

where S_0 is a suitable convex subset of D. If

$$h := 2\alpha\beta\gamma \leq 1, \tag{25}$$

and

$$S := \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq r := \frac{2\alpha}{1+\sqrt{1-h}}\} \subseteq S_0 \tag{26}$$

then F(x) has a unique zero in S. □

Remarks. 1. There is also an affine invariantive form of the proposition, see e.g. Miel [7]. But in practical applications this simply amounts to applying the preceding to BF(x) in place of F(x), with a suitable preconditioning matrix B.

2. Usually, the constant γ is determined as a bound of the norm of the second derivative F"(x) in S_0.

3. γ usually depends on the choice of S_0. Since in (26), $\alpha < r \leq 2\alpha$; it is sensible to choose

$$S_0 := \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq \kappa\alpha\}. \tag{27}$$

with a parameter κ satisfying

$$1 < \kappa \leq 2; \tag{28}$$

of course, we must assume that $S_0 \subseteq D$ (note that Rall [14], Theorem 24.1. recommends to use (27) with $\kappa = 2$). With this choice of S_0,

(25) and (26) are equivalent with

$$h := 2\alpha\beta\gamma \le 4(\kappa - 1)/\kappa^2 . \tag{29}$$

4. The relation

$$\|F(x)-F(x_0)-F'(x_0)(x-x_0)\| \le \tfrac{1}{2}\gamma \|x-x_0\|^2 \quad \text{for } x,y \in S_0 \tag{30}$$

is a consequence of (24); see [12].

Now assume that $\|\ \| = \|\cdot\|_u$ is a scaled l_∞-norm. Then, with S_0 as in (27), we find from (30) that

$$|F(x) - F(x_0) - F'(x_0)(x - x_0)|_u \le \tfrac{1}{2}\gamma\kappa^2\alpha^2 u. \tag{31}$$

Hence we are in the situation of Section 2 with

$$A = F'(x_0),\ \|\delta_0\| = \alpha,\quad c = \tfrac{1}{2}\alpha\gamma\kappa^2 u,$$

and we find $\|b\| = \| |A^{-1}|c| \|_u = \frac{1}{2}\alpha\gamma\kappa^2 \| |A^{-1}|u| \|_u = \frac{1}{2}\alpha\gamma\kappa^2 \|A^{-1}\|_u \le \frac{1}{2}\alpha\gamma\kappa^2\beta = \frac{1}{4}h\kappa^2$.

We see that (6) and (29) are equivalent inequalities. Hence, for scaled l_∞-norms, the a priori Kantorovic bound is equivalent to our theorem, when c is computed from the relation (30).

Now we show by an example (in dimension n=1) that our theorem may produce a bound when Kantorovic's bound does not work. Consider

$$F(x) = x^3 + 12x + 12, \qquad D = \mathbb{R}$$
$$x_0 = 0,\ A = F'(x_0) = 12.$$

Then Kantorovic's constants are

$$\alpha = 1,\ \beta = \frac{1}{12},\ \gamma = \sup_{x,y \in S_0} 3|x + y| \ge 6 \sup_{x \in S_0}|x|.$$

Condition (25) implies that $\gamma \le 6$, hence $\sup_{x \in S_0} |x| \le 1 = \alpha$, whence

$$S_0 \subseteq \{x \in \mathbb{R} \mid \|x - x_0\| \le \alpha\},$$

and (26) cannot be satisfied. Therefore, no bound is obtained. On the other hand, our theorem, applied with $\kappa = 3/2$ gives

$$\delta_0 = A^{-1} F(x_0) = 1,$$
$$S = \{x \in \mathbb{R} \mid |x| \le 3/2\},$$
$$|F(x) - F(x_0) - A(x - x_0)| = |x^3| \le 27/8,$$

whence

$$c = 27/8,\ b = |A^{-1}|c = 9/32 \le 1/2 = \kappa - 1.$$

Therefore, F(x) has a zero ξ with $|\xi| \le 3/2$. Note that (9) gives the sharper estimate $|\xi+1| \le 9/32$, whence $-1.3 < \xi < -0.7$. In fact, we have $\xi \approx -0.932441$.

From the relationship to Kantorovic's bound it again appears that it is sensible to choose $\kappa \le 2$. But there are cases when a value of $\kappa > 2$ is useful. For example, choose $t > 0$ and put

$$F(x) = \begin{cases} e^x - 1 & \text{if } x \le t, \\ e^t - 1 + e^t(x - t) & \text{if } x > t. \end{cases} \tag{32}$$

Then F(x) has a continuous derivative

$$F'(x) = e^{\min(x,t)}.$$

The unique zero of F(x) is $\xi = 0$. Hence if we take

$$x_0 := t,\quad A := F'(x_0) = e^t,$$

so that

$$\delta_0 = 1 - e^{-t}\ (>0),$$

the set (3) contains the zero only if $\kappa\|\delta_0\| > t$, or

$$\kappa \ge t/(1 - e^{-t}). \tag{33}$$

Now suppose we choose κ according to (33), so that

$$r := \kappa\delta_0 \ge t.$$

Since

$$F(x_0 + w) - F(x_0) - Aw = \begin{cases} e^t(e^w - w - 1) & \text{if } w < 0, \\ 0 & \text{if } w \ge 0, \end{cases}$$

is nonnegative and decreasing, it takes its maximal absolute value for the smallest admissible w. Therefore, (4),(5) hold with

$$c = e^t(e^{-r} + r - 1)/\delta_0,$$
$$b = (e^{-r} + r - 1)/\delta_0 = \kappa - (1 - e^{-r})/(1 - e^{-t}) \le \kappa - 1.$$

By our theorem, the bound

$$|\xi - x_0| \le \kappa\delta_0 \tag{34}$$

is valid for every κ satisfying (33).

Remarks. 1. If $t \ge 2$ then (33) implies $\kappa > 2$, whence no value of $\kappa \le 2$ can be used to bound the zero.

2. If x_0 is "close" to the zero ξ then t is small, and we may take κ small as well; note that $t/(1-e^{-t}) = 1 + \frac{1}{2}t + O(t^2)$.

3. The example is unusual in that every large κ gives a bound. This is due to the fact that the function F(x) is asymptotically linear for large x.

4. If κ is chosen optimally (equality in (33)) then the bound (34) gives the error exactly (equality in the bound). Although this does not generalize to arbitrary functions F(x), it indicates the quality of the estimate.

5. Numerical example

As an illustration, we consider the nonlinear bounding value problem ("minimal surface equation")

$$\begin{aligned} (1+u_y^2)\,u_{xx} - 2u_x u_y u_{xy} + (1+u_x^2)\,u_{yy} &= 0 \quad \text{in } D, \\ u(x,y) &= g(x,y) \quad \text{on the boundary } \partial D. \end{aligned} \tag{35}$$

We look at the problem when D is the square,

$$D = \{(x,y) \mid 0 \le x \le 1,\ 0 \le y \le 1\},$$

replace the square by a grid of $(N+1)^2$ points,

$$\left(\tfrac{i}{N}\,,\ \tfrac{k}{N}\right),\quad i,k \in \{0,1,\dots,N\},$$

and approximate the function u and its derivative in these points as follows:

$$u\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx v_{ik},$$

$$u_x\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx \tfrac{N}{2}\,\Delta_x v_{ik},\quad u_y\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx \tfrac{N}{2}\,\Delta_y v_{ik},$$

$$u_{xx}\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx N^2\,\Delta_{xx} v_{ik},\quad u_{yy}\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx N^2\,\Delta_{yy} v_{ik},$$

$$u_{xy}\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \approx \tfrac{N^2}{4}\,\Delta_{xy} v_{ik},$$

where

$$\Delta_x v_{ik} = v_{i+1,k} - v_{i-1,k},\quad \Delta_y v_{ik} = v_{i,k+1} - v_{i,k-1},$$

$$\Delta_{xx} v_{ik} = v_{i+1,k} - 2v_{ik} + v_{i-1,k},\quad \Delta_{yy} v_{ik} = v_{i,k+1} - 2v_{i,k} + v_{i,k-1},$$

$$\Delta_{xy} v_{ik} = v_{i+1,k+1} - v_{i+1,k-1} - v_{i-1,k+1} + v_{i-1,k-1}.$$

The boundary conditions give

$$v_{ik} = g\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right) \text{ if } i \in \{0,N\} \text{ or } k \in \{0,N\},$$

and the remaining v_{ik} are arranged as an $(N-1)^2$-dimensional block vector

$$v = (v_1,\dots,v_{N-1})^T;\quad \text{with } v_i = (v_{i1},\dots,v_{i,N-1})^T.$$

At a fixed interior point $\left(\tfrac{i}{N}\,,\tfrac{k}{N}\right)$, the differential equation is replaced by the discrete equation

$$0 = F_{ik}(v) = (\alpha+\delta_y^2)\,\delta_{xx} - \tfrac{1}{2}\,\delta_x\delta_y\delta_{xy} + (\alpha+\delta_x^2)\,\delta_{yy}, \tag{36_{ik}}$$

where

$$\alpha = 4/N^2,$$

$$\delta_x = \Delta_x v_{ik},\ \delta_y = \Delta_y v_{ik},\ \text{etc.}$$

For a small vector w with

$$\|w\|_\infty \le \omega, \tag{37}$$

$$w_{ik} = 0 \quad \text{if } i \in \{0,N\} \text{ or } k \in \{0,N\}, \tag{38}$$

the quantities

$$\epsilon_x = \Delta_x v_{ik},\ \epsilon_y = \Delta_y v_{ik},\ \text{etc.}$$

are bounded by

$$|\epsilon_x| \le 2\omega,\ |\epsilon_y| \le 2\omega,$$

$$|\epsilon_{xx}| \le 4\omega,\ |\epsilon_{xy}| \le 4\omega,\ |\epsilon_{yy}| \le 4\omega.$$

Now

$$\begin{aligned} F_{ik}(v+w) = {} & (\alpha+(\delta_y+\epsilon_y)^2)\,(\delta_{xx}+\epsilon_{xx}) - \tfrac{1}{2}\,(\delta_x+\epsilon_x)\,(\delta_y+\epsilon_y)\,(\delta_{xy}+\epsilon_{xy}) \\ & + (\alpha+(\delta_x+\epsilon_x)^2)\,(\delta_{yy}+\epsilon_{yy}). \end{aligned}$$

We extract from this the part linear in the epsilons and obtain a formula for the derivative:

$$\begin{aligned} [F'(v)w]_{ik} = {} & (\alpha+\delta_y^2)\,\epsilon_{xx} - \tfrac{1}{2}\,\delta_x\delta_y\epsilon_{xy} + (\alpha+\delta_x^2)\,\epsilon_{yy} \\ & + (2\delta_x\delta_{yy} - \tfrac{1}{2}\,\delta_y\delta_{xy})\,\epsilon_x + (2\delta_y\delta_{xx} - \tfrac{1}{2}\,\delta_x\delta_{xy})\,\epsilon_y. \end{aligned} \tag{39_{ik}}$$

If we take differences we find

$$\begin{aligned} F_{ik}(v+w) - F_{ik}(v) - [F'(v)w]_{ik} = {} & \\ = {} & \epsilon_y^2(\delta_{xx}+\epsilon_{xx}) - \tfrac{1}{2}\,\epsilon_x\epsilon_y(\delta_{xy}+\epsilon_{xy}) + \epsilon_x^2(\delta_{yy}+\epsilon_{yy}) \\ & + 2\delta_y\epsilon_y\epsilon_{xx} - \tfrac{1}{2}\,(\delta_x\epsilon_y + \epsilon_x\delta_y)\,\epsilon_{xy} + 2\delta_x\epsilon_x\epsilon_{yy}, \end{aligned}$$

hence

$$|F(v+w) - F(v) - F'(v)w|_{ik} \le 4\omega^2(e_{ik} + 10\omega), \tag{40_{ik}}$$

where

$$e_{ik} = |\delta_{xx}| + \tfrac{1}{2}|\delta_{xy}| + |\delta_{yy}| + 5|\delta_x| + 5|\delta_y|. \tag{41_{ik}}$$

Now, if v is an approximation to a zero of the system $F(v)=0$ given by (36_{ik}) for $i,k \in \{1,\dots,N-1\}$ then we may apply our theorem with $A=F'(v)$; an explicit (block tridiagonal) matrix expression is available from (39_{ik}) by expansion to linear combinations of the $v_{i+j,k+t}$, $j,t \in \{0,1,-1\}$. Once $\delta_0=A^{-1}F(v)$ and

$$\omega = \kappa \|\delta_0\|$$

are computed, c is available from (40) as

$$c_{ik} = 4\kappa\omega(e_{ik} + 10\omega); \qquad (42_{ik})$$

if the e_{ik} are computed simultaneously with $F(v)$ and $F'(v)$, little extra effort is necessary.

As a numerical example we consider the problem (35) on the unit square D, with

$$g(x,y) = \sqrt{xy(2-xy)}.$$

We take g(x,y) as initial approximation for u(x,y). For N = 8, we obtain the following initial block vector $v = (v_{ik})$:

.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
.0000	.1761	.2480	.3026	.3480	.3875	.4227	.4547	.4841
.0000	.2480	.3480	.4227	.4841	.5367	.5830	.6242	.6614
.0000	.3026	.4227	.5113	.5830	.6433	.6953	.7407	.7806
.0000	.3480	.4841	.5830	.6614	.7262	.7806	.8268	.8660
.0000	.3875	.5367	.6433	.7262	.7929	.8472	.8914	.9270
.0000	.4227	.5830	.6953	.7806	.8472	.8992	.9391	.9682
.0000	.4547	.6242	.7407	.8268	.8914	.9391	.9721	.9922
.0000	.4841	.6614	.7806	.8660	.9270	.9682	.9922	1.0000

We use $\kappa = 1.5$ and find after i Newton steps

$\bar{\beta} = 630.5$, $\omega = 0.3270$, for $i = 0$,

$\bar{\beta} = 0.4788$, $\omega = 0.00065$, for $i = 6$,

$\bar{\beta} = 0.1316$, $\omega = 0.00018$, for $i = 7$.

Thus the error estimate is unreliable for $i = 0$ ($\bar{\beta} \gg \kappa-1 = 0.5$), probably reliable for $i = 6$ ($\bar{\beta} \approx \kappa-1$), and reliable for $i = 7$ ($\bar{\beta} \ll \kappa-1$). Hence the discretized solution agrees with the approximation for $i = 7$,

.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
.0000	.0219	.0446	.0690	.0972	.1333	.1860	.2780	.4841
.0000	.0445	.0903	.1391	.1941	.2615	.3516	.4812	.6614
.0000	.0690	.1391	.2121	.2914	.3821	.4908	.6235	.7806
.0000	.0972	.1942	.2914	.3910	.4958	.6088	.7317	.8660
.0000	.1333	.2615	.3821	.4958	.6047	.7110	.8169	.9270
.0000	.1861	.3517	.4909	.6088	.7110	.8017	.8852	.9682
.0000	.2782	.4812	.6235	.7317	.8169	.8852	.9413	.9922
.0000	.4841	.6614	.7806	.8660	.9270	.9682	.9922	1.0000

Within two units of the last decimal. Of course, the discretization error is not included in our bound.

Our second example is the catenoid

$$u(x,y) = \cosh^{-1}\sqrt{(x+1)^2+(y-0.5)^2}, \qquad (44)$$

which is a solution of the minimal surface equation. We take it as initial approximation for the discretized equation with N = 8 on the unit square. The computed error bound is now a bound for the discretization error of this particular problem. We obtain (for $\kappa = 1.5$)

$\bar{\beta} = 3.2265$, $\omega = 0.0062$,

hence no reliable bound. But after two Newton steps, we have

$\bar{\beta} = 0.0987$, $\omega = 0.0002$;

inspection shows that the two approximations differ by at most 0.0045, whence the first estimate is also valid.

Finally, we check an example on the square $[0,20] \times [0,20]$ given by Douglas [5], who used essentially the same discretization and calculated the following approximation (in 1928, by hand) using coordinate relaxation:

.0	9.7	13.2	15.6	17.3	18.5	19.4	19.8	20.0
.0	5.4	9.4	12.4	14.6	16.3	17.8	18.9	19.8
.0	3.5	6.8	9.6	12.0	14.1	16.0	17.8	19.4
.0	2.5	4.9	7.4	9.7	11.9	14.1	16.3	18.5
.0	1.8	3.7	5.5	7.5	9.7	12.0	14.6	17.3
.0	1.3	2.6	4.0	5.5	7.4	9.6	12.4	15.6
.0	.8	1.7	2.6	3.7	4.9	6.8	9.4	13.2
.0	.4	.8	1.3	1.8	2.5	3.5	5.4	9.7
.0	.0	.0	.0	.0	.0	.0	.0	.0

For $\kappa = 1.5$ we obtain

$\bar{\beta} = 38.3463$, $\omega = 0.8544$.

The correctly rounded solution

.0	9.7	13.2	15.6	17.3	18.5	19.4	19.8	20.0
.0	5.6	9.6	12.5	14.6	16.3	17.7	18.8	19.8
.0	3.7	7.0	9.8	12.2	14.2	16.0	17.7	19.4
.0	2.7	5.2	7.6	9.9	12.1	14.2	16.3	18.5
.0	1.9	3.9	5.8	7.8	9.9	12.2	14.6	17.3
.0	1.4	2.8	4.2	5.8	7.6	9.8	12.5	15.6
.0	.9	1.8	2.8	3.9	5.2	7.0	9.6	13.2
.0	.4	.9	1.4	1.9	2.7	3.7	5.6	9.7
.0	.0	.0	.0	.0	.0	.0	.0	.0

Acknowledgment I want to thank Prof. Collatz for his critical remarks, and Prof. Törnig for suggesting the minimal surface equation as an interesting application.

References

1. G. Alefeld, Über die Durchführbarkeit des Gaußschen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten, Computing , Suppl. 1 (1977), 15-19.
2. G. Alefeld, Bounding the slope of polynomial operators and some applications, Computing 26 (1981), 227-237.
3. A.L. Cline, C.B. Moler, G.W. Stewart, and J.H. Wilkinson, An estimate for the condition number of a matrix, SIAM J.Numer.Anal. 16 (1979), 368-375.
4. J.J. Dongarra, J.R. Bunch, C.B. Moler, and G.W. Stewart, LINPACK Users Guide. SIAM, Philadelphia 1979.
5. J. Douglas, A method of numerical solution of the problem of Plateau, Annals of Math. (2) 29 (1928), 180-188.
6. M. Fiedler and V. Ptak, On matrices with nonpositive off-diagonal elements and positive principal minors, Czech. Math.J. 12 (1962), 382-400.
7. G. Miel, An updated version of the Kantorovich theorem for Newton's method, Computing, to appear.
8. R.E. Moore, Methods and Applications of Interval Analysis, SIAM Publications, Philadelphia 1979.
9. R.E. Moore, New results on nonlinear systems. In: Interval Mathematics 1980 (Ed. K. Nickel), Acad. Press, New York - London 1980, pp. 165-180.
10. A. Neumaier, Hybrid norms, the Ostrowski operator, and bounds for solutions of linear equations, to appear.
11. D.P. O'Leary, Estimating matrix condition numbers, SIAM J. Sci.Stat.Comput. 1 (1980), 205-209.
12. J.M. Ortega and W.C. Rheinboldt, Iterative solution of nonlinear equations in several variables. Acad. Press, New York - London 1970.
13. A.M. Ostrowski, Über die Determinanten mit überwiegender Hauptdiagonale, Comm.Math.Helv. 10 (1937), 69-96.
14. L.B. Rall, Computational solution of nonlinear operator equations. Wiley, New York - London 1969.

has error parameters

$$\overline{\beta} = 2.6193, \qquad \omega = 0.0742.$$

In both cases we have $\overline{\beta} > \kappa - 1$ although the true maximal error is $< \omega$. This is partially explained by the fact that the Newton iteration converges very slowly: For the first 25 iterations starting from Douglas' approximation, the convergence is only linear, with factor 0.7. The first iterate with $\overline{\beta} < \kappa - 1$ is the 16th, with

$$\overline{\beta} = 0.2808, \qquad \omega = 0.0082$$

(this allows checking the validity of the other two bounds). Thus we have here a very nonlinear problem.

All examples were calculated on a UNIVAC 1110. The time for one Newton iteration was 0.13 sec (without error information) resp. 0.16 sec (with error information)

Use was made of (21) to compute $\overline{\beta}$, although the matrices A were neither M-matrices nor diagonally dominant. Perhaps this accounts for the fact that in the examples ω was a bound even when $\overline{\beta}$ was too large.