EFFICIENT NUMERICAL VALIDATION OF SOLUTIONS OF NONLINEAR SYSTEMS*

G. ALEFELD[†], A. GIENGER[†], AND F. POTRA[‡]

Abstract. A new stopping criterion for Newton's method is derived by combining the properties of the Krawczyk operator and a corollary of the Newton-Kantorovich theorem. When this criterion is satisfied the authors use the last three Newton iterates to compute an interval vector that is very likely to contain a solution of the given nonlinear system. The existence of such a solution is tested using Krawczyk's operator. Furthermore, each element from this interval vector considered as an approximation to the solution has a relative error that is of the order of the machine precision. Extensive numerical testing has shown that the proposed method has very good practical performance.

Key words. nonlinear systems, Newton-Kantorovich theorem, validation of solutions

AMS subject classifications. 65H10, 65G10

1. Introduction. Newton's method is the best-known algorithm for solving nonlinear systems of equations. The most famous theoretical result on the convergence of Newton's method is probably the theorem of Kantorovich (also known as the Newton-Kantorovich theorem), which gives sufficient conditions that guarantee the existence and uniqueness of a solution x^* as well as the convergence of the Newton iterates towards that solution. However, the sufficient conditions are phrased in terms of a global Lipschitz constant, so that they are very difficult to check in practice. Nevertheless, in the case of simple solutions, i.e., solutions at which the Jacobian is nonsingular, the Kantorovich hypothesis is always satisfied if the starting point is close to the solution. A common numerical practice is to stop the Newton iteration whenever the distance between two iterates is less than a given tolerance, i.e., when

$$||x^{k+1} - x^k|| \le \epsilon.$$

As shown in Gragg and Tapia [6], this practice is justified because if the Kantorovich hypothesis is satisfied at x^k , then

(2)
$$||x^k - x^*|| \le 2\epsilon, \qquad ||x^{k+1} - x^*|| \le \epsilon.$$

However, just the fact that (1) is satisfied does not guarantee the existence of a solution. Even if a solution for x^* exists, (1) alone does not provide an estimate for the distance between x^k or x^{k+1} and the solution.

Over the past thirty years a lot of effort has been spent in order to overcome these problems. By using interval arithmetic methods, different algorithms have been developed to generate sequences of *n*-dimensional intervals $[x]^k$, $k = 0, 1, \ldots$, that satisfy the relation

$$x^{\star} \in [x]^{k+1} \subseteq [x]^k \subseteq \cdots$$

Such an iterative algorithm can be safely stopped when the diameter of the interval $[x]^k$ becomes smaller than ϵ .

[‡] Department of Mathematics, University of Iowa, Iowa City, Iowa 52242.

^{*} Received by the editors October 26, 1992; accepted for publication (in revised form) February 9, 1993.

[†] Institut für Angewandte Mathematik, Universität Karlsruhe, Kaiserstraße 12, D-7500 Karlsruhe, Germany.

EFFICIENT NUMERICAL VALIDATION

Because of extensive use of interval arithmetic, these iterative procedures tend to be quite expensive computationally. Moreover they need a starting interval $[x]^0$ that is guaranteed to contain a simple solution x^* . A number of interval operators, like different variants of the Krawczyk operator K, have been developed to test whether or not an interval contains a root. (The precise definition of K is given in Chapter 3.) Namely, by using Brouwer's fixed point theorem, it can be shown that if

$$(3) K[x] \subseteq [x],$$

then $x^* \in K[x]$. The problem is then to find a suitable test interval [x], to compute the interval K[x] so that it includes all round-off errors, and to test (3). A recent survey on such interval operators has been given in [14].

Another approach is mentioned in Moore and Kioustelidis [8], where Miranda's theorem is used instead of Brouwer's fixed point theorem for proving the existence of a solution in a given test interval. As mentioned by the authors (see [8, p. 523]), the test interval has to be chosen neither too large nor too small. However, no numerical methods for obtaining such a test interval are given.

In the present paper we propose the use of the Kantorovich theorem in order to efficiently produce a good test interval that presumably contains a solution. Namely, we proceed with Newton's method performed in normal floating point arithmetic. For a given *eps* equal to the machine precision, we devise a stopping criterion and construct a test interval [x] such that (3) is very likely to be satisfied. Moreover, our method is designed in such a way that the condition

(4)
$$\frac{\|y - x^*\|_{\infty}}{\|x^*\|_{\infty}} \le eps$$

is also eventually satisfied. Here y denotes any point of the interval K[x].

Besides having an elegant theoretical justification, the resulting algorithm turns out to be very efficient in practice. It gives highly accurate results and at the same time provides a tool for establishing the existence of solutions of certain equations. For example, in [9] the problem of existence of a solution of the methanol-8 problem was mentioned as unsolved. By using the algorithm presented in the present paper we have managed to both establish the existence of a solution of the methanol-8 problem and to find a very good approximation for it.

2. Notation and preliminaries. Real numbers are denoted by a, b, \ldots . Real bounded and closed intervals are denoted by $[a] = [a_1, a_2], [b] = [b_1, b_2], \ldots$. The same notation is used for real vectors and interval vectors, e.g.,

$$a = (a_i), \quad a_i \in \mathbb{R},$$

 $[a] = ([a]_i), \quad \text{where } [a]_i \text{ are compact real intervals.}$

Real (n, n)-matrices are denoted by $A = (a_{ij}), \ldots$ and the corresponding interval matrices are denoted by $[A] = ([a]_{ij}), \ldots$ The operations in the set of intervals and in the set of interval matrices and/or interval vectors can be found in [2, Chap. 10].

The diameter of an interval $[a] = [a_1, a_2]$ is $d([a]) = a_2 - a_1$. The absolute value is defined as $|[a]| = \max\{|a_1|, |a_2|\}$. We mention just a few rules:

$$d([a] \pm [b]) = d([a]) + d([b]),$$

$$d([a][b]) \le d([a])|[b]| + |[a]|d([b]).$$

If $0 \in [a]$ and $0 \in [b]$, then

$$d([a][b]) \le d([a])d([b])$$
 (see [3]).

For interval vectors and interval matrices the diameter and the absolute value are defined componentwise.

3. The Krawczyk operator. Assume that $F: D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable mapping for which an interval arithmetic evaluation F'([x]) of the derivative exists for a certain set of interval vectors $[x] \subseteq D$. Let $x \in [x]$ be a real vector, and let C be a fixed real nonsingular (n, n)-matrix. The mapping

$$K([x], x, C) = x - CF(x) + (I - CF'([x]))([x] - x)$$

is called the Krawczyk operator. K([x], x, C) is again an interval vector.

The following result holds (see [2, Thm. 10, Chap. 13], or [7]).

THEOREM 3.1. If $K([x], x, C) \subseteq [x]$, then F has a zero x^* in K([x], x, C).

The proof of Theorem 3.1 is based on Brouwer's fixed point theorem. Theorem 3.1 is a very powerful result that allows the validation of solutions in a given interval vector.

In the discussion below we need the following result concerning the Krawczyk operator.

THEOREM 3.2. Assume that the mapping $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is differentiable and that the derivative has an interval arithmetic evaluation F'([x]) for all $[x] \in D$ such that

(5)
$$\|\mathrm{d}(F'([x]))\|_{\infty} \leq \hat{L}\|\mathrm{d}([x])\|_{\infty}, \qquad [x] \subseteq D,$$

for some $\hat{L} \geq 0$. If $C^{-1} \in F'([x])$, then the inequality

(6)
$$\|d(K([x], x, C))\|_{\infty} \le \gamma \|d([x])\|_{\infty}^{2}$$

holds with $\gamma = \|C\|_{\infty} \hat{L}$.

Proof. Following the rules for the interval arithmetic operations, for the diameter and for the absolute value, we obtain

$$d(K([x], x, C)) = d(x - CF(x) + (I - CF'([x]))([x] - x))$$

= d((I - CF'([x]))([x] - x))
$$\leq d(I - CF'([x]))d([x] - x)$$

= d(C(C⁻¹ - F'([x])))d([x])
= |C|d(F'([x]))d([x]).

Using (5) we obtain (6).

A fundamental problem in applying Theorem 3.1 is the question of finding an appropriate "test-interval" [x]. In practice one usually proceeds as follows.

By some iteration method a floating point approximation \tilde{x} to a solution x^* of F(x) = 0 is computed. Then one defines $x := \tilde{x}$ and an interval vector [x] is constructed with \tilde{x} as the center and with diameter 2ϵ where ϵ is "small." Finally, C is chosen as a floating point approximation to $F'(\tilde{x})^{-1}$. If the Krawczyk operator maps [x] into itself, then [x] contains a solution. Otherwise a certain strategy, called ϵ -inflation, is used to construct a new test interval. See [13], for example.

In this paper we introduce a strategy for computing a test interval that is based on some elementary conclusions following from the Newton-Kantorovich theorem.

254

4. The Newton-Kantorovich theorem and the construction of test intervals. Consider Newton's method

(7)
$$x^{k+1} = x^k - F'(x^k)^{-1}F(x^k), \qquad k = 0, 1, 2, \dots,$$

applied to a mapping $F: D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. The Newton-Kantorovich theorem gives sufficient conditions for the convergence of Newton's method starting at x^0 . Furthermore, it contains an error estimation. A simple discussion of this estimation in conjunction with Theorem 3.2 will lead us to a test interval that can be computed by using only iterates of Newton's method.

THEOREM 4.1 (see [10, Thm. 12.6.2]). Assume that $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is differentiable in the ball $\{x | | |x - x^0|| \leq r\}$ and that

(8)
$$||F'(x) - F'(y)|| \le L||x - y||$$

for all x, y from this ball. Suppose that $F'(x^0)^{-1}$ exists and that $||F'(x^0)^{-1}|| \leq B_0$. Let

$$||x^1 - x^0|| = ||F'(x^0)^{-1}F(x^0)|| \le \eta_0,$$

and assume that

$$h_0 = B_0 \eta_0 L \le \frac{1}{2}, \qquad r_0 = \frac{1 - \sqrt{1 - 2h_0}}{h_0} \eta_0 \le r.$$

Then the Newton iterates (7) are well defined, remain in the ball $\{x | ||x - x^0|| \le r_0\}$, where

$$r_0 = \frac{1 - \sqrt{1 - 2h_0}}{h_0} \eta_0,$$

and converge to a solution x^* of F(x) = 0 that is unique in $D \cap \{x | ||x - x^0|| < r_1\}$, where

$$r_1 = \frac{1 + \sqrt{1 - 2h_0}}{h_0} \eta_0$$

provided $r \geq r_1$. Moreover, the error estimate

(9)
$$||x^{\star} - x^{k}|| \le \frac{1}{2^{k-1}} (2h_0)^{2^{k}-1} \eta_0, \qquad k \ge 0$$

holds.

We mention that this theorem has been used in [11] to prove the existence of solutions by explicitly computing L (this can be done by interval arithmetic evaluation of the second partial derivatives) and the bounds B_0 and η_0 . A comparison between such an approach and the test (3) is given in [12].

Since $h_0 \leq \frac{1}{2}$, the error estimate (9) (for k = 0, 1 and the ∞ -norm) leads to

$$\|x^{\star} - x^{0}\|_{\infty} \le 2\eta_{0} = 2\|x^{1} - x^{0}\|_{\infty},$$

 $\|x^{\star} - x^{1}\|_{\infty} \le 2h_{0}\eta_{0} \le \eta_{0} = \|x^{1} - x^{0}\|_{\infty}.$

By replacing x^0 with x^k one can show in a similar manner that if the hypothesis of the Newton-Kantorovich theorem is satisfied, then (1) implies (2).



FIG. 1. Error estimate (9) for k = 1 and the ∞ -norm.

This suggests a simple construction of an interval vector containing the solution. The situation is illustrated in Fig. 1. If x^0 is close enough to the solution x^* , then x^1 is much closer to x^* than x^0 since Newton's method is quadratically convergent. The same holds if we choose any vector $(\neq x^*)$ from the ball $\{x | ||x - x^1||_{\infty} \leq \eta_0\}$ as the starting vector for Newton's method. Because of (6) and since $x^* \in K([x], x, C)$, it is reasonable to assume that $K([x], x^1, F'(x^0)^{-1}) \subseteq [x]$ for

(10)
$$[x] = \left\{ x | ||x - x^1||_{\infty} \le \eta_0 \right\}.$$

The important point is that this test interval [x] can be computed without knowing B_0 and L. Of course all the arguments above are based on the assumption that the hypothesis of the Newton-Kantorovich theorem is satisfied, which may not be the case if x^0 is far away from x^* .

We try to overcome this difficulty by first performing a certain number of Newton steps until we are close enough to a solution x^* of F(x) = 0. Then we compute the interval (10) and using the Krawczyk operator we test whether this interval contains a solution. The question of when to terminate the Newton iteration is answered by the following considerations.

Our general assumption is that the Newton iterates are convergent to x^* . For ease of notation we set

$$[y] := K([x], x^{k+1}, F'(x^k)^{-1}),$$

256

where

(11)

$$[x] = \left\{ x \in \mathbf{R}^n | \| x^{k+1} - x \|_{\infty} \le \eta_k \right\},\\eta_k = \| x^{k+1} - x^k \|_{\infty}$$

for some fixed k.

Our goal is to terminate Newton's method as soon as

(12)
$$\frac{\|\mathbf{d}([y])\|_{\infty}}{\|x^{k+1}\|_{\infty}} \le eps$$

holds, where *eps* is the machine precision of the floating point system. If $x^* \in [x]$, then $x^* \in [y]$ so that for any $y \in [y]$ we have

(13)
$$\frac{\|x^* - y\|_{\infty}}{\|x^*\|_{\infty}} \le \frac{\|d([y])\|_{\infty}}{\|x^*\|_{\infty}}.$$

Since $||x^*||_{\infty}$ differs only slightly from $||x^{k+1}||_{\infty}$ if x^{k+1} is near x^* , (12) guarantees that the relative error with which any $y \in [y]$ approximates x^* is close to machine precision.

Now let $\tilde{L} = \max{\{\hat{L}, L\}}$, where \hat{L} and L are defined by (5) and (8), respectively. Since by Theorem 3.2 we have

$$\|d([y])\|_{\infty} \leq \|C\|_{\infty} L \|d([x])\|_{\infty}^{2}$$

and since $||d([x])||_{\infty} = 2\eta_k$, (12) holds if

(14)
$$4\frac{\|C\|_{\infty}\tilde{L}\eta_k^2}{\|x^{k+1}\|_{\infty}} \le eps$$

is true.

From Newton's method we have

$$x^{k+1} - x^k = C\left\{F(x^k) - F(x^{k-1}) - F'(x^{k-1})(x^k - x^{k-1})\right\}$$

and by 3.2.12 in [10] it follows that

(15)
$$\eta_k \leq \frac{1}{2} \|C\|_{\infty} \tilde{L} \eta_{k-1}^2.$$

Replacing the inequality sign by the equality in this relation and eliminating $||C||_{\infty}L$ from (14) we get the following stopping criterion for Newton's method:

(16)
$$\frac{8\eta_k^3}{\|x^{k+1}\|_{\infty}\eta_{k-1}^2} \le eps.$$

Of course, it is not a mathematical proof that if (16) is satisfied, then the interval [y] constructed as above will contain x^* and that the vectors in [y] will approximate x^* with a relative error close to eps. However, it seems reasonable that (16) should work well in practice.

Inequality (16) is not explicitly formulated in terms of \tilde{L} and $||C||_{\infty}$, but contains only eps, x^{k-1}, x^k , and x^{k+1} . Hence (16) can be checked at each step of Newton's method as soon as three successive iterates have been computed. If (16) is fulfilled we take the interval vector [x] defined in (12) and compute

$$\begin{split} &K([x], x^{k+1}, F'(x^k)^{-1}) \\ &= x^{k+1} - F'(x^k)^{-1} F(x^{k+1}) - (I - F'(x^k)^{-1} F'([x]))([x] - x^{k+1}). \end{split}$$

Now if

$$[y] = K([x], x^{k+1}, F'(x^k)^{-1}) \subseteq [x],$$

we know by Theorem 3.1 that F(x) = 0 has a solution $x^* \in [y] \subseteq [x]$.

The test based on the stopping criterion (16) works extremely well in practice, as will be seen in the examples in the next section. Occasionally, however, it happens that the radius η_k of the ball is already too small for ensuring $K([x], x^{k+1}, F'(x^k)^{-1}) \subseteq [x]$. In this case we replace [x] from (12) by

(17)
$$[x] = \left\{ x \in \mathbb{R}^n | \| x^{k+1} - x \|_{\infty} \le \sqrt{\eta_k \eta_{k-1}} \right\}.$$

This modification was necessary only in a few examples and worked very well.

5. Test examples and numerical results. The ideas of this paper have been extensively tested on a large number of numerical examples. These tests can be found in [5]. Because of lack of space, we explicitly list here only two examples. The first example is quite standard and easy to handle, while the other one is more difficult.

Example 1. The usual discretization of the nonlinear boundary problem (see [1])

$$3y''y + (y')^2 = 0,$$

 $y(0) = 0, y(1) = 20,$

with the exact solution $y = 20t^{3/4}$ leads to the nonlinear system

$$F_1(x) = 3x_1(x_2 - 2x_1) + \frac{1}{4}x_2^2,$$

$$F_i(x) = 3x_i(x_{i+1} - 2x_i + x_{i-1}) + \frac{1}{4}(x_{i+1} - x_{i-1})^2, \qquad i = 2(1)n - 1,$$

$$F_n(x) = 3x_n(20 - 2x_n + x_{n-1}) + \frac{1}{4}(20 - x_{n-1})^2.$$

We take $x_i = 10$, i = 1(1)n as the starting value for Newton's method.

Example 2. The second problem we list here is the so-called "distillation column test problem," which is described in full detail in [9]. We consider the so-called methanol-8 problem, which leads to a nonlinear system with 31 unknowns and equations. In the discussion of the numerical experience with this system, Moré reports in [9]: "I still do not know if there exists a solution to the methanol-8 problem." We treated this problem and we were successful in proving the existence of a solution. The numerical values are available upon request. (We were also able to validate the solutions of the hydrocarbon-6 and hydrocarbon-20 problems.)

In Tables 1-2 below we have used the following abbreviations: n: number of unknowns and equations; k + 1: number of Newton steps until the termination criteria (16) is fulfilled; η_k : infinity norm of the difference of the two last iterates; γ_k : geometric mean of η_k and η_{k-1} , $\gamma_k = (\eta_k \eta_{k-1})^{1/2}$; Val: indicates the radius r of the ball with center x^{k+1} for which the validation test was successful; ρ_k : infinity norm $\|d([y])\|_{\infty}$ of the diameter of [y], rel: approximation of the relative error of $y \in [y]$, $rel = \|d([y])\|_{\infty}/\|x^{k+1}\|_{\infty}$.

EFFICIENT NUMERICAL VALIDATION

n	k+1	η_k	η_{k-1}	γ_k	Val	ρ_k	rel
10	8	$6.89 \cdot 10^{-13}$	$3.74 \cdot 10^{-6}$	$1.61 \cdot 10^{-9}$	η_k	$1.07 \cdot 10^{-14}$	$5.73 \cdot 10^{-16}$
20	8	$8.91 \cdot 10^{-11}$	$4.31 \cdot 10^{-5}$	$6.20 \cdot 10^{-8}$	η_k	$2.49 \cdot 10^{-14}$	$1.29 \cdot 10^{-15}$
50	9	$6.32 \cdot 10^{-13}$	$8.11 \cdot 10^{-7}$	$7.16 \cdot 10^{-10}$	η_k	$1.42 \cdot 10^{-14}$	$7.21 \cdot 10^{-16}$
100	10	$1.78 \cdot 10^{-15}$	$1.75 \cdot 10^{-8}$	$5.57 \cdot 10^{-12}$	γ_k	$1.42 \cdot 10^{-14}$	$7.16 \cdot 10^{-16}$

TABLE 1 Numerical results for Example 1.

	TABLE 2							
Numerical	results	for	Example	2.				

n	k+1	η_k	η_{k-1}	γ_k	Val	ρ_k	rel
31	5	$1.30 \cdot 10^{-10}$	$1.62 \cdot 10^{-4}$	$1.46 \cdot 10^{-7}$	η_k	$2.00 \cdot 10^{-10}$	$2.05 \cdot 10^{-13}$

The numerical examples have been computed on an HP-9000 workstation using the programming language PASCAL-XSC. The machine eps on this computer is approximately 10^{-16} .

We close this section with a few remarks concerning the practical programming of the ideas in this paper.

(a) Newton's method and the test for the stopping criteria are performed in floating point arithmetic;

(b) Assume that C denotes a floating point approximation of $F'(x^k)^{-1}$ (where $F'(x^k)$ is the floating point derivative for the floating point approximation of x^k). Such an approximation can be computed very cheaply, since from the last Newton step the floating point triangular decomposition of $F'(x^k)$ is known. Then testing $K([x], x^{k+1}, C) \subseteq [x]$ we have to take into account the rounding errors in computing

$$K([x], x^{k+1}, C) = x^{k+1} - CF(x^{k+1}) - (I - CF'([x]))([x] - x^{k+1}).$$

This is done in the following way. The floating point vector x^{k+1} is considered as an interval (with diameter zero) and $F(x^{k+1})$ is computed by following the rules of interval arithmetic and by rounding outwards in each step. Therefore we get an inclusion of $F(x^{k+1})$. We proceed similarly in computing F'([x]). Afterwards we compute the right-hand side of the Krawczyk operator by again following the rules of interval arithmetic and rounding outwards where necessary. In this manner we get an interval vector $\tilde{K}([x], x^{k+1}, C)$ with $K([x], x^{k+1}, C) \subseteq \tilde{K}([x], x^{k+1}, C)$. Therefore, if $\tilde{K}([x], x^{k+1}, C) \subseteq [x]$, then $K([x], x^{k+1}, C)$ and a solution of F(x) = 0 exists in $\tilde{K}([x], x^{k+1}, C)$.

More details can be found in [5].

REFERENCES

- J. P. ABBOTT AND R. P. BRENT, Fast local convergence with single and multistep methods for nonlinear equations, Austral. Math. Soc., 19 (1975), pp. 173-199.
- [2] G. ALEFELD AND J. HERZBERGER, Introduction to Interval Computations, Academic Press, New York and London, 1983.
- [3] H. CORNELIUS AND R. LOHNER, Computing the range of values of real functions with accuracy higher than second order, Computing, 33 (1984), pp. 331-347.
- [4] M. Y. COSNARD AND J. J. MORÉ, Numerical solution of nonlinear equations, ACM Trans. Math. Software, 5 (1979), pp. 64-85.
- [5] A. GIENGER, Zur Lösungsverifikation bei nichtlinearen Gleichungssystemen, Diplomarbeit, Universität Karlsruhe, Germany, 1992.

G. ALEFELD, A. GIENGER, AND F. POTRA

- [6] W. B. GRAGG AND R. A. TAPIA, Optimal error bounds for the Newton Kantorovich theorem, SIAM J. Numer. Anal., 11 (1974), pp. 10–13.
- [7] R. E. MOORE, A test for existence of solutions to nonlinear systems, SIAM J. Numer. Anal., 14(1977), pp. 611-615.
- [8] R. E. MOORE AND J. B. KIOUSTELIDIS, A simple test for accuracy of approximate solution to nonlinear (or linear) systems, SIAM J. Numer. Anal., 17 (1980), pp. 521-529.
- [9] J. J. MORÉ, A collection of nonlinear model problems, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., Lectures in Applied Mathematics, Volume 26, American Mathematical Society, Providence, RI, 1990, pp. 723-762.
- [10] J. M. ORTEGA AND W. C. RHEINBOLDT, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York and London, 1970.
- [11] L. B. RALL, Computational Solution of Nonlinear Operator Equations, Wiley, New York, 1969.
- [12] —, A comparison of the existence theorems of Kantorovich and Moore, SIAM J. Numer. Anal., 17 (1980), pp. 148–161.
- [13] S. M. RUMP, Solving nonlinear systems with least significant bit accuracy, Computing, 29 (1982), pp. 183-200.
- [14] M.A. WOLFE, On certain computable tests and componentwise error bounds, a talk given at Schloss Dagstuhl, Conference on Symbolic, Algebraic and Validated Numerical Computation, August 1992.