Interval Computations
No 3, 1993

Computing Zeros of Functions Using Generalized Interval Arithmetic

Eldon R. Hansen

We consider the use of a generalized interval arithmetic in algorithms for solving nonlinear equations or systems of nonlinear equations. The algorithms can involve either derivatives or slopes. The convergence rate is improved for either form. The improvement is greater if slopes, rather than derivatives, are used. However, the slope method is applicable to only rational functions. For multidimensional problems we introduce the generalized interval arithmetic into the Hansen-Sengupta method. Again, the convergence rate is improved.

Вычисление нулей функций при помощи обобщенной интервальной арифметики

Э. Р. Хансен

Рассматривается использование обобщенной интервальной арифметики в алгоритмах для решения нелинейных уравнений или систем нелинейных уравнений. В эти алгоритмы могут входить либо производные, либо <наклоны>. И в том и в другом случае улучшается скорость сходимости. Но при использовании наклонов это улучшение значительнее, чем в случае производных. Однако метод наклонов применим лишь для рациональных функций. Для многомерных задач мы вводим обобщенную интервальную арифметику в метод Хансена-Сенгупты. Скорость сходимости улучшается и в этом случае.

1 Introduction

The “slope method” is an algorithm using interval arithmetic for computing and bounding the zeros of nonlinear functions or systems of nonlinear functions. It was introduced by Krawczyk and Neumaier [1]. See also [2], [3]. It is closely related to the interval Newton method. It is even more closely related to a method described in [4]. Each method (in interval form) provides guaranteed bounds on all zeros in a given region.

We shall derive the slope method for one-dimensional problems in Section 2. In Section 3, we show how it can be incorporated into the Hansen-Sengupta method for multidimensional problems.

In Section 4, we describe a generalized interval arithmetic (g.i.a.) introduced by the author in [1]. We give an illustrative example of its use in Section 5. In Section 6, we show how g.i.a. can be introduced into the one-dimensional slope method.

In Section 7, we give a detailed discussion of how g.i.a. can be used in the multidimensional slope method and how the structure of g.i.a. can be exploited to extract enhanced convergence when solving nonlinear equations or systems.

Section 8 contains an illustrative example of the use of g.i.a. in solving nonlinear problems in one variable.

Numerical results for a two-dimensional problem using g.i.a. in a slope version of the Hansen-Sengupta method are given in Section 9.

The use of g.i.a. does not improve convergence of the interval Newton as greatly as it does that of the slope method. In Section 10, we show why this is the case; and note that the slope method is applicable to only rational functions.

In Section 11, we compare four possible combinations of methods using either a Newton or slope method and either ordinary interval arithmetic (o.i.a.) or g.i.a. Numerical results for the four methods are given in Section 12.

In Section 13, we compare the slope method to a procedure introduced by the author in [4] and show why the slope method is superior.

The last two sections contain discussions of miscellaneous topics.

We believe that everything in this paper relating to g.i.a. is new except for the introductory material in Section 4.

2 The slope method in one dimension

In this section, we shall introduce the slope method by considering a simple problem in the one-dimensional case. We shall describe how it can be used to compute and bound real zeros of rational functions.

Consider a simple power, $p(x) = x^m$, of x for some integer, m. Note that

$$p(y) - p(x) = y^m - x^m = (y-x)\sum_{k=0}^{m-1} x^k y^{m-1-k}. \tag{2.1}$$

Now consider a polynomial

$$f(x) = \sum_{m=0}^{n} a_m x^m.$$

Using (2.1), we obtain the identity

$$f(y) - f(x) = (y-x)g(x,y) \tag{2.2}$$

where

$$g(x,y) = \sum_{m=0}^{n} a_m \sum_{k=0}^{m-1} x^k y^{m-1-k}.$$

Let y be a zero of f. Then $f(y) = 0$ and from (2.2),

$$y = x - f(x)/g(x,y). \tag{2.3}$$

Suppose X is an interval containing the zero, y, of f. From (2.3),

$$y \in x - f(x)/g(x,X).$$

This relation forms the basis for an algorithm for computing the zero, y. Define

$$N(x,X) = x - f(x)/g(x,X). \tag{2.4}$$

If $y \in X$, then $y \in N(x,X)$. We replace the bound X on y by

$$X' = X \cap N(x,X).$$

Iterating this step produces a set of nested intervals converging to the zero, y. This is the *slope method*.

We have derived the slope method for the case in which $f(x)$ is a polynomial. It serves also for computing and bounding the zeros of any rational function. However, it is not applicable when $f(x)$ is irrational because $g(x, X)$ does not have a finite representation in this case.

The reader will note the close similarity of this procedure to the interval Newton method. Let $f'(X)$ denote the derivative of $f(x)$ evaluated with interval argument, X. If we replace $g(x, X)$ in the slope method by $f'(X)$, we obtain the interval Newton method. See Section 10 for a derivation of the interval Newton method. For more details, see, for example, [2] or [3].

The slope method generally converges in fewer steps than the standard interval Newton method. However, it requires more computing per step. The reason is that $g(x, X)$ is contained in $f'(x)$ and is generally a narrower interval than $f'(x)$; but it is a more complicated function.

We can illustrate this with a simple example. Let

$$f(x) = x^4.$$

Then

$$f'(X) = 4X^3 \quad \text{and} \quad g(x, X) = x^3 + x^2X + xX^2 + X^3.$$

We can rewrite $f'(X)$ as

$$f'(X) = X^3 + X^3 + X^3 + X^3.$$

Note that each term x^3, x^2X, and xX^2 in $g(x, X)$ is narrower than the corresponding term, X^3, of $f'(X)$. Therefore, $g(x, X)$ is narrower than $f'(X)$.

This result is general. The slope function, $g(x, X)$, contains the real quantity, x, instead of the interval, X, in various places. As a result, $g(x, X)$ is narrower than $f'(X)$.

The slope method and the interval Newton method share a number of common properties which make them extremely efficient and reliable algorithms. See [2].

The slope method extends naturally to the multidimensional case. See [2] or [3]. Thus, it is applicable for computing the zeros of rational functions of one or more variables.

Unfortunately, if f is irrational, then $g(x, X)$ is not a finite sum. Because of this, the slope method cannot be used by itself to compute zeros of irrational functions. However, a hybrid method can be used in which the slope method is used for the "rational part" of f and the interval Newton method is used for the "irrational part". See [2]. For an alternative method, see Section 13.

Note that the slope method could be used in a non-interval setting. In this case, of course, it would not provide error bounds on the computed zeros. In this paper, we always assume it is implemented in interval arithmetic.

3 The multidimensional slope method

Let x and y be vectors of n components. In the multidimensional slope method, we seek zeros of a vector functions, $f(x) : \mathbb{R}^n \to \mathbb{R}^n$. We assume that each component of f is a rational function of the components of x. Therefore, each component of f can be expanded in a form which is an extension of (2.2). See [2] or [3]. Putting these expansions together in vector form, we obtain an equation of the form

$$f(y) = f(x) + J(x, y)(y - x). \tag{3.1}$$

Here, $J(x, y)$ is a matrix corresponding to the Jacobian of f. Its elements are rational functions of the components of x and y.

Assume that y is a zero of f. Then $f(y) = 0$ and, from (3.1),

$$f(x) + J(x, X)(y - x) = 0 \tag{3.2}$$

where we have replaced y in $J(x, y)$ by the box X in which we seek a zero of f.

Note that, to be correct, we should write (3.2) to indicate that the vector 0 is contained in the left member. However, here (and elsewhere) we conform to common usage and write the relation as an equation.

To compute (with error bounds) the zeros of f in this multidimensional case, we shall use the Hansen-Sengupta method. See [2] or [6] for details. It makes no difference therein whether J is a usual Jacobian or the "slope Jacobian".

In the Hansen-Sengupta method we precondition this equation by multiplying by an approximate inverse, B, of the center of $J(x,X)$. Thus, (3.2) becomes

$$M(x,X)(y-x)=r(x) \tag{3.3}$$

where $M(x,X)=BJ(x,X)$ and $r(x)=-Bf(x)$.

We solve the i-th equation of (3.3) for the i-th component of $y-x$. Before doing so, we replace each component of y except for the i-th by the (most recently computed) corresponding component of the box in which we seek a zero of f. Thus, we obtain an interval

$$N_i(x,X)=x_i+P_i/M_{ii} \tag{3.4a}$$

where

$$P_i=r_i(x)-\sum_{j=1}^{i-1}M_{ij}(X_j'-x_j)-\sum_{j=i+1}^{n}M_{ij}(X_j-x_j). \tag{3.4b}$$

The new bound for y_i is obtained as

$$X_i'=X_i\cap N_i(x,X). \tag{3.5}$$

The new box, X', whose components are given by (3.5), contains any zero of f which is in X.

This step is done for each $i=1,\ldots,n$ and the process is iterated until the new box is sufficiently small.

Note that M_{ii} may contain the value zero. If so, the division in (3.4a) may create a gap in the i-th component of the current box. When this occurs, the gap can be used to split the box into sub-boxes. See [2] for details.

In Section 7, we modify this procedure to introduce g.i.a.

4 Generalized interval arithmetic

In this section, we shall briefly describe the generalized interval arithmetic (g.i.a.) derived in [5].

The original intent in the derivation of g.i.a. (see [5]) was simply to provide sharper error bounds than those obtained using o.i.a. Thus, for example, we might replace o.i.a. in the interval Newton method by g.i.a.

However, it turned out to have other virtues as well. In [5], we showed how it could reduce the growth of multidimensional intervals (the so-called wrapping effect) due to rotations. We shall show below how it can be used to improve the convergence of the slope method or the interval Newton method.

Consider an interval, X, of width $2s$. Denote the midpoint of X by x. Any point in X can be expressed as $x+u$ for some value of u satisfying $-s \leq u \leq s$. When we evaluate a function using g.i.a., we retain terms linear in u and bound higher order terms.

For example, note that a point in X^2 can be expressed as $(x+u)^2 = x^2+2xu+u^2$. To simplify exposition, we shall write

$$X^2 = x^2 + 2xu + u^2$$

although the left member is an interval and the right member is a single number in that interval. This should create no confusion and we shall continue to use this notation.

Since $-s \leq u \leq s$, it follows that $0 \leq u^2 \leq s^2$. Therefore, we express X^2 as

$$[x^2, x^2+s^2] + 2xu.$$

Thus, the result is linear in u and contains $(x+u)^2$ for all $u \in [-s, s]$.

In general, when we evaluate a function of X, at each step, we combine (by addition, subtraction, multiplication, or division) two intermediate functions, say $f(X)$ and $g(X)$. Each of these functions will be expressed as linear functions of u. Denote $f(X) = A + Bu$ and $g(X) = C + Du$ where A, B, C, and D are intervals.

Consider the four arithmetic operations for combining $f(X)$ and $g(X)$. Let $h(X)$ be the result and express $h(X)$ as $E + Fu$. For addition or subtraction,

$$f(X) \pm g(X) = (A \pm Bu) + (C \pm Du) = (A \pm C) + (B \pm D)u.$$

In this case, $E = A \pm C$ and $F = B \pm D$.

For multiplication, we have

$$f(X)g(X) = (A+Bu)(C+Du) = AC + (AD+BC)u + BDu^2.$$

We bound the term u^2 by $[0, s^2]$. Therefore, $f(X)g(X) = E + Fu$, where

$$E = AC + BD[0, s^2] \quad \text{and} \quad F = AD + BC.$$

For division, we have

$$\frac{f(X)}{g(X)} = \frac{A + Bu}{C + Du} = \frac{A}{C} + \frac{(BC - AD)u}{C(C + Du)}.$$

In the denominator of the last term, we replace u by its bound $[-s, s]$. Thus, we obtain

$$E = \frac{A}{C} \quad \text{and } F = \frac{BC - AD}{C(C + D[-s, s])}.$$

In general, the interval, X, will be a datum. For example, it may be the interval in which we seek the root of a polynomial. More generally, we shall have more than one datum interval. For example, we may compute the inverse of a matrix for which each element is an independent interval.

If the data consist of n intervals, X_i, of respective widths, w_i, we express each of them as $X_i = x_i + u_i$ where $u_i \in [-s_i, s_i]$. As we combine these data intervals, we form intermediate generalized intervals of the form

$$f(X) = A_0 + \sum_{i=1}^{n} A_i u_i \quad \text{and} \quad g(X) = B_0 + \sum_{i=1}^{n} B_i u_i \tag{4.1}$$

where A_i and B_i $(i = 0, \ldots, n)$ are intervals.

Let “op” denote one of the operations of addition, subtraction, multiplication, or division in g.i.a. Denote

$$h(X) = C_0 + \sum_{i=1}^{n} C_i u_i \tag{4.2}$$

where $h(X) = f(X)$ op $g(X)$. The rules for computing $h(X)$ are obtained in much the same way as for the single datum case.

For addition, we have

$$C_i = A_i + B_i \qquad (i = 0, \ldots, n). \tag{4.3}$$

For subtraction,

$$C_i = A_i - B_i \qquad (i = 0, \ldots, n). \tag{4.4}$$

For multiplication, let us first consider the case $n = 2$. That is,

$$\begin{aligned} f(X_1, X_2) &= A_0 + A_1 u_1 + A_2 u_2 \\ g(X_1, X_2) &= B_0 + B_1 u_1 + B_2 u_2. \end{aligned}$$

We have

$$\begin{aligned} h(X_1, X_2) &= f(X_1, X_2) g(X_1, X_2) \\ &= A_0 B_0 + (A_0 B_1 + A_1 B_0) u_1 + (A_0 B_2 + A_2 B_0) u_2 \\ &\quad + A_1 B_1 u_1^2 + A_2 B_2 u_2^2 + (A_1 B_2 + A_2 B_1) u_1 u_2. \end{aligned}$$

As in the one datum case, we replace u_1^2 $(i = 1,\ 2)$ by its bounding interval $[0, s_i^2]$.

We must decide what to do with the product $u_1 u_2$. We can replace either u_1 by the interval $[-s_1, s_1]$ which bounds it or we can replace u_2 by $[-s_2, s_2]$. It the first case, we retain a term linear in u_2. In the second case, we retain a term linear in u_1. In general, it is not obvious which is the better policy. For a case in which we do know which alternative to use, see Section 14.

For multiplication for general n, we want the product of $f(X)$ and $g(X)$ where they are of the form given by (4.1). We obtain the product $h(X)$ in the form (4.2) where

$$C_0 = A_0 B_0 + \sum_{i=1}^{n} A_i B_i [0, s_i^2] \tag{4.5a}$$

and

$$C_i = A_0 B_i + B_0 A_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} A_i B_j [-s_j, s_j] \qquad \text{for } i = 1,\ \ldots,\ n. \tag{4.5b}$$

Here we have arbitrarily replaced u_j by $[-s_j, s_j]$ in the product $u_i u_j$.

For division, we obtain $C_0 = A_0 / B_0$ and

$$C_i = (B_0 A_i - A_0 B_i)/D \tag{4.6}$$

where

$$D = B_0 \Big(B_0 + \sum_{i=1}^{n} B_i [-s_i, s_i] \Big).$$

Once we have obtained the desired function as a linear function of $u_1, \ldots,$ u_n, we "reduce" it to an interval by replacing u_i by its bounding interval $[-s_i, s_i]$ $(i = 1,\ \ldots,\ n)$.

5 An example using g.i.a.

As an example of g.i.a., consider the function

$$f(X_1, X_2) = (1 + X_1 X_2)/(X_1 + X_2)$$

where $X_1 = [0.8, 1]$ and $X_2 = [1, 1.2]$. We express a point in X_1 as $x_1 + u_1$ where $x_1 = 0.9$ is the midpoint of X_1 and $u_1 \in [-0.1, 0.1]$. We express a point in X_2 as $x_2 + u_2$ where $x_2 = 1.1$ is the midpoint of X_2 and $u_2 \in [-0.1, 0.1]$.

Using (4.5), we obtain

$$X_1 X_2 = 0.99 + [1, 1.2]u_1 + 0.9u_2.$$

From this result and (4.3),

$$1 + X_1 X_2 = 1.99 + [1, 1.2]u_1 + 0.9u_2. \tag{5.1}$$

Using (4.3),

$$X_1 + X_2 = 2 + u_1 + u_2. \tag{5.2}$$

From (5.1) and (5.2) we obtain, using (4.6)

$$f(X_1, X_2) = \frac{1.99}{2} + \left[\frac{0.01}{4.4}, \frac{0.41}{3.6}\right] u_1 + \left[\frac{-0.19}{3.6}, \frac{-0.19}{4.4}\right] u_2.$$

We "reduce" this result to an interval by replacing u_1 by its bounding interval $[-0.1, 0.1]$ and replacing u_2 by its (same) bounding interval. We obtain

$$f(X_1, X_2) = \left[\frac{3.522}{3.6}, \frac{3.642}{3.6}\right] = [0.978, 1.012].$$

If we had used ordinary interval arithmetic (o.i.a.), we would have obtained $f(X_1, X_2) = [0.818, 1.223]$. The g.i.a. procedure gives a sharper result in this case. However, for other input intervals or other functions, o.i.a. may obtain a better result than g.i.a. The latter tends to be superior when the data intervals are narrow and when the data variables occur many times in the function being evaluated.

6 Use of g.i.a. in the slope method (the one-dimensional case)

We now consider the slope method in one dimension. The basic step in this method is to compute a new bound, $N(x, X)$, on a zero of f from a bounding interval, X, as

$$N(x, X) = x - f(x)/g(x, X).$$

See (2.4). Here x is the midpoint of X. The slope function $g(x, X)$ is defined in Section 2.

Denote $U = [-s, s]$ where s is the halfwidth of X. We express a point in X as $x + u$ where $u \in U$. We evaluate $g(x, X)$ using g.i.a. and obtain

$$g(x, X) = C + Du \tag{6.1}$$

for some intervals C and D. If we now replace u by U in (6.1), we obtain

$$N(x, X) = x - f(x)/(C + DU). \tag{6.2}$$

We can evaluate the right member using interval arithmetic.

However, we can do better than this as we shall see in Section 7.

Before discussing the details, we shall consider the problem of computing zeros of multidimensional functions. The details to complete the computation in the one-dimensional case can be subsumed into the corresponding details for the multidimensional case.

7 Use of g.i.a. in the slope method (the multidimensional case)

In the multidimensional case, the slope method involves the use of equation (3.4). When we introduce g.i.a., the elements of the matrix, M, will be expressed as generalized intervals.

When we solve for the i-th new (interval) variable, we replace the other variables, y_j, by their bounding intervals, X_j, for $j = 1, \ldots, n$, $j \neq i$. We also "reduce" the variables, u_j, for $j \neq i$. This is actually the same process since we first replaced y_j by $x_j + u_j$ to introduce u_j.

We can think of the process as first replacing y_j by $x_j + u_j$ and then replacing u_j by $X_j - x_j$. Therefore, we can compute $u_j(y_j - x_j)$ as the square of an interval instead of the product of two different intervals. This yields sharper results.

When we use g.i.a. to "evaluate" the right member of equation (3.4), we obtain an intermediate result of the form

$$N_i(x, X) = x_i + (A + Bu_i)/(C + Du_i) \tag{7.1}$$

where A, B, C, and D are intervals.

Note that equation (6.2) is of the same form as equation (7.1) except that $B = 0$ for the former. That is, when we use the slope method in one variable to parallel the interval Newton method, we get a relation of essentially the same form as when we use the slope method to parallel the Hansen-Sengupta method in the multidimensional case. Since (7.1) is more general than (6.2), we shall discuss only the former.

For simplicity, we now drop the subscript, i, and write (7.1) in the form

$$N(x, X, u) = x + (A + Bu)/(C + Du). \tag{7.2}$$

We shall discuss this relation as if we were seeking a zero in the one-dimensional case. This simplifies the language since we will not have to keep repeating that we are seeking a *component* of a zero.

If we replace u in (7.2) by its bounding interval, U, and evaluate the result using o.i.a., we obtain an interval which contains any zero of f which is in the original interval, X. This follows from the derivation of the slope method. This would be the straightforward way to proceed.

Note, however, that if we consider the single point, u, in $U = X - x$, then $N(x, X, u)$ is a bound on the result we would get if we sought a zero of f in the degenerate interval $x + u$. Therefore, if $x + u$ is not in the interval $N(x, X, u)$, then $x + u$ is not a zero of f.

Using this fact, we can often prove that some (or all) of the initial interval, X, does not contain a zero of f. We generally obtain more information in this way than by simply "reducing" as described above. We proceed as follows.

We first consider the case in which $0 \notin C + Du$ for the values of u of interest. In this case $N(x, X, u)$ is a finite interval for a given value of u.

From (7.2), $x + u \notin N(u)$ if

$$u < (A + Bu)/(C + Du) \tag{7.3}$$

or

$$u > (A + Bu)/(C + Du). \tag{7.4}$$

Hence, if either (7.3) or (7.4) is satisfied, then $x + u$ is not a zero of f.

Next consider the case in which $0 \in C + Du$ for values of u of interest. We must now use extended interval arithmetic to divide $A + Bu$ by $C + Du$. Extended interval arithmetic (in which we are allowed to divide by an interval containing zero) was derived independently in [7] and [6]. For a discussion of the restricted part of this arithmetic that is needed here, see, for example, [2].

In both $0 \in A + Bu$ and $0 \in C + Du$, then using extended interval arithmetic to evaluate the right member of (7.2), we obtain the entire real line as the interval result. In this case, we follow the standard procedure for the o.i.a. case and split the interval into two or more subintervals.

If $0 \in C + Du$ but $0 \notin A + Bu$, then the right member of (7.2) is computed to be the entire real line but with a gap missing. In this case, $x + u$ is not a zero of f if u is in the gap. For details, see below.

We now consider how to determine the values of u for which $x+u$ cannot be a zero of f. Denote

$$R(u) = (A + Bu)/(C + Du).$$

From (7.2),

$$N(x, X, u) = x + R(u).$$

First, we need to express the endpoints of the numerator interval, $A+Bu$, and the endpoints of the denominator interval, $C + Du$, of $R(u)$ explicitly. To do so, we consider the cases $u < 0$ and $u > 0$ separately. The numerical values of the endpoints of A, B, C, and D will be known. Therefore, the endpoints of $A + Bu$ and $C + Du$ can be explicitly expressed as linear functions of u.

We could simplify this process by choosing x to be an endpoint of the interval X. As a result, u would always be of one sign. We have not used this alternative.

After determining the endpoints of $A+Bu$ and $C+Du$, we must divide $A+Bu$ by $C+Du$ with u still unspecified except in sign. That is, we want the endpoint of $R(u)$ as an explicit function of u. In order to do this division, we must know the signs of the endpoints of $A+Bu$ and $C+Du$.

But, these sign depend on the value of u. They may change as u varies over U. Therefore, we must compute the values of u at which the endpoints of $A+Bu$ and $C+Du$ change sign. We can then break up the interval $U=[-s,s]$ bounding u into separate subintervals to be treated individually.

The values of u where these endpoints are zero constitute four points at which we have to subdivide U (if they are in U). The value $u=0$ (which is always in U) is another point at which we had to subdivide U earlier. Depending on how many of these points are in U (and distinct), we need to subdivide U into two to six subintervals and treat each separately.

Recall that, for one-dimensional problems, $B=0$. Therefore, in this case, we need to subdivide U into at most four subintervals. If we had chosen x to be an endpoint of X, then u would be of one sign and we would need to subdivide U into at most three subintervals for one-dimensional problems and into at most five subintervals in for multidimensional problems.

If, for a particular subinterval of values of u, we find that $0 \notin C+Du$, then we obtain an explicit result for R of the form $R(u)=[Q(u),Q'(u)]$ where $Q(u)$ and $Q'(u)$ are of the form

$$Q(u)=(a+bu)/(c+du) \qquad \text{and} \qquad Q'(u)=(a'+b'u)/(c'+d'u).$$

In this case, $x+u$ is not a zero of f if either

$$u<(a+bu)/(c+du) \qquad \text{or} \qquad u>(a'+b'u)/(c'+d'u). \tag{7.5}$$

If $0 \in C+Du$, but $0 \notin A+Bu$, we obtain a result of the form

$$R(u)=[-\infty,P(u)] \cup [P'(u),\infty] \tag{7.6}$$

where $P(u)$ and $P'(u)$ are of the form

$$P(u)=(a+bu)/(c+du) \qquad \text{and} \qquad P'(u)=(a'+b'u)/(c'+d'u).$$

In this case, $x+u$ is not a zero of f if u is in the gap between the two semi-infinite intervals; that is, if

$$u>(a+bu)/(c+du) \qquad \text{and} \qquad u<(a'+b'u)/(c'+d'u). \tag{7.7}$$

Examining (7.5) and (7.6), we see that we want to solve a relation of the form

$$(a+bu)/(c+du) < u \quad \text{or} \quad > u$$

or a similar relation involving the primed quantities.

In order to get the explicit form of the endpoints of $R(u)$, we shall have restricted our attention to a particular subinterval of U in which $c+du$ and $c'+d'u$ are each of one sign only. Consider the case in which $c+du>0$. Then the relation $R(u)>u$ becomes $Q(u)>u$ and can be rewritten

$$a+bu > u(c+du).$$

In general, knowing the sign of $c+du$ and of $c'+d'u$, the relation $R(u)>u$ (that is, $Q(u)>u$) or $R(u)<u$ (that is, $Q'(u)<u$) can be rewritten as a specific quadratic relation of the form

$$a+bu < u(c+du) \qquad \text{or} \qquad a+bu > u(c+du) \tag{7.8}$$

or else a similar relation involving the primed variables.

We solve this quadratic relation for the values of u which satisfy it. The solution values of interest are those which are in the subinterval of U being considered.

We now list the various cases that can occur. We first consider the case in which $d=0$ and/or $d'=0$. If $d=0$, then $Q(u)$ is a linear function of u and we have

$$Q(u)>u \quad \text{if} \begin{cases} a/c>0 \quad \text{and} \quad c=b, \\ u>a/(c-b), c\neq b, \quad \text{and} \quad b/c>1, \\ u<a/(c-b), c\neq b, \quad \text{and} \quad b/c<1. \end{cases} \tag{7.9}$$

If $d'=0$, then

$$Q'(u)<u \quad \text{if} \begin{cases} a'/c'<0 \quad \text{and} \quad c'=b', \\ u<a'/(c'-b'), c'\neq b', \quad \text{and} \quad b'/c'>1, \\ u>a'/(c'-b'), c'\neq b', \quad \text{and} \quad b'/c'<1. \end{cases} \tag{7.10}$$

For all other cases, we assume $d\neq 0$ and $d'\neq 0$. Let v denote the discriminate of the quadratic in (7.8); i.e.,

$$v=(c-b)^2+4ad.$$

Similarly, define

$$v' = (c' - b')^2 + 4a'd'.$$

If $v < 0$, then

$$Q(u) > u \qquad \text{if} \quad u < -c/d. \tag{7.11}$$

If $v' < 0$, then

$$Q'(u) < u \qquad \text{if} \quad u > -c'/d'. \tag{7.12}$$

If $bc - ad = 0$, then $Q(u) = a/c$. Therefore,

$$Q(u) > u \qquad \text{for } u < a/c. \tag{7.13}$$

If $b'c' - a'd' = 0$, then $Q'(u) = a'/c'$. Therefore,

$$Q'(u) < u \qquad \text{for } u > a'/c'. \tag{7.14}$$

The remainder of the cases depend on the roots of the quadratic in (7.8) (or of the corresponding quadratic for the primed variables). We have already considered the case in which the discriminate is negative (in which case the roots are complex and of no interest). We now have the case in which they are real.

Let r and s denote the roots ordered so that $r \leq s$. For the primed variables, denote the roots by r' and s' where $r' \leq s'$.

For $bc - ad < 0$ (which incidentally implies that the roots r and s are real),

$$Q(u) > u \quad \text{for } -c/d < u < s \quad \text{and for } u < r. \tag{7.15}$$

For $b'c' - a'd' < 0$

$$Q'(u) < u \quad \text{for } r' < u < -c'/d' \quad \text{and for } u > s'. \tag{7.16}$$

Now assume $bc - ad > 0$ and denote

$$q = (bc - ad)^{1/2}.$$

Then

$$Q(u) > u \qquad \begin{array}{l} \text{for } s < u < -c/d \text{ and for } u < r \\ \text{when } d > 0 \text{ and } b + c < -2q \\ \text{or when } d < 0 \text{ and } b + c > 2q; \end{array} \tag{7.17}$$

$$Q(u) > u \qquad \begin{array}{l} \text{for } r < u < r' \text{ and for } u < -c/d \\ \text{when } d > 0 \text{ and } b + c > 2q \\ \text{or when } d < 0 \text{ and } b + c < -2q. \end{array} \tag{7.18}$$

Similarly, assume $b'c' - a'd' > 0$ and denote

$$q' = (b'c' - a'd')^{1/2}.$$

Then

$$Q'(u) < u \qquad \begin{array}{l} \text{for } r' < u < s' \text{ and for } u > -c'/d' \\ \text{when } d' > 0 \text{ and } b' + c' < -2q' \\ \text{or when } d' < 0 \text{ and } b' + c' > 2q'; \end{array} \tag{7.19}$$

$$Q'(u) < u \qquad \begin{array}{l} \text{for } -c'/d' < u < r' \text{ or for } u > s' \\ \text{when } d' > 0 \text{ and } b' + c' > 2q' \\ \text{or when } d' < 0 \text{ and } b' + c' < -2q'. \end{array} \tag{7.20}$$

It may appear that the computations we have described in this section involve a lot of work. It is a lot of work for the programmer because of the many details involved. However, the computational effort is rather small compared to that for other parts of the overall algorithm.

It can be shown that, for one-dimensional problems, the procedure we have described is cubically convergent to simple zeros and quadratically convergent to double zeros. Obtaining this enhanced convergence is certainly worth the relatively small amount of extra computing.

Unfortunately, proofs of rates of convergence is very lengthy because of the many special cases which occur in the arithmetic. We shall not take up the necessary space here.

8 An illustrative example

In this section, we give an example to illustrate the steps of the g.i.a. procedure. Suppose we are using this procedure to find the zeros of the polynomial

$$f(x) = x^5 - 8x^3 + 6x^2 + 7x - 6 = (x+3)(x-2)(x+1)(x-1)^2. \tag{8.1}$$

The slope function for f can be written

$$\begin{aligned} g(x, x+u) &= u^4 + 5xu^3 + (10x^2 - 8)u^2 \\ &\qquad + (10x^3 - 24x + 6)u + 5x^4 - 24x^2 + 12x + 7. \end{aligned}$$

Suppose that in the course of solving this problem the algorithm generates the subinterval $[0.038, 1.83]$ and seeks a root therein. Thus, $x = 0.934$ and u is restricted to the interval $U = [-0.896, 0.896]$.

We use o.i.a. to evaluate f in order to bound rounding errors and obtain $f(0.934) = [-0.0354, -0.0353]$. We use g.i.a. to evaluate g and obtain

$$g(0.934, 0.934 + u) = [1.07, 2.31] + u[-8.27, -4.51].$$

To conserve space, we record results to only three significant digits. More were used in the computations. Also, henceforth, we shall omit the arguments when they are clear from the context.

To express g more explicitly, we must known the sign of u. We first consider the case $u \leq 0$. Thus, we restrict u to the subinterval $[-0.896, 0]$ of U. In this case,

$$g = [1.07 - 4.51u,\ 2.31 - 8.27u].$$

Note that $g > 0$ since $u \leq 0$. Knowing this, we are able to express $N(x, X, u)$ from (2.4) explicitly as

$$N = 0.934 + [Q, Q']$$

where

$$Q = 0.0353/(2.31 - 8.27u) \quad \text{and} \quad Q' = 0.0353/(1.07 - 4.51u). \tag{8.2}$$

Since $Q > 0$ (because $u \leq 0$), we have $Q > u$. Therefore, the point $x+u$ is not in N for any u in the current subinterval $[-0.896, 0]$ of U. That is, there is no zero of f in $[0.038, 0.934]$.

Since we have eliminated all of the current subinterval, we do not try to make use of the right endpoint function, Q'.

We now consider the case $u \geq 0$; that is, $0 \leq u \leq 0.896$. In this case,

$$g = [1.07 - 8.27u,\ 2.31 - 4.51u].$$

Note that the left endpoint of g is zero for $u = 0.13$ and the right endpoint is zero for $u = 0.509$. Therefore, we subdivide the subinterval $[0, 0.896]$ into the smaller subintervals $[0, 0.13]$, $[0.13, 0.509]$, and $[0.509, 0.896]$.

Consider the first of these subintervals, $[0, 0.13]$. Therein, $g > 0$ and the condition $Q > u$ is

$$0.0353/(2.31 - 4.51u) > u.$$

From (7.17), $Q > u$ for $0 \leq u < 0.0158$. Therefore, there is no zero of f for these values of u.

We now consider the condition $Q' < u$ for the subinterval $[0, 0.13]$ of U. We have just learned that there is no zero for $u \in [0, 0.0158]$. Hence, we need only consider the subinterval $[0.0158, 0.13]$.

We find that the discriminant, v', is negative. Hence, from (7.12), $Q' < u$ for $u > 0.13$. Thus we gain no useful information about the current subinterval.

Next, we consider the subinterval $[0.13, 0.509]$ of U. In this interval, $0 \in g$, and $R(u)$ is given by (7.6). We need both $P < u$ and $P' > u$ in order to assert that $x + u$ is not a zero of f. The condition $P < u$ in explicit form is

$$0.0353/(1.08 - 8.27u) < u.$$

The discriminant of the quadratic (see (7.8)) is negative and, from (7.12) (with P in place of Q'), we find that $P < u$ for the entire subinterval.

The condition $P' > u$ is

$$0.0353/(2.3 - 4.52u) > u.$$

From (7.17) (with P' in place of Q), this holds for $0.493 < u < 0.51$. Therefore, u is in the gap in $R(u)$ for $0.493 < u < 0.51$. It follows that there is no zero of $f(x)$ for u anywhere in the current subinterval, $[0.493, 0.509]$.

Finally, we consider the subinterval of U for $0.509 \leq u \leq 0.896$. We find that Q and Q' are the same as for the case $0 < u < 0.13$. From (7.17), we find that the condition $Q > u$ gives no useful information. We have

$$Q' = 0.0353/(1.08 - 8.27u)$$

and, from (7.17) (with Q' in place of Q), we find that $Q' < u$ for $u > 0.13$. That is, there is no zero of f for any value of u in $[0.509, 0.896]$.

Combining the information from all the four subintervals, we find that any zero of f in the original interval, $[0.038, 1.83]$, must be in $[0.95, 1.06]$. As it must, the new interval contains the same zero of f at $x = 1$ as did the original interval.

In this one iteration of our algorithm, we have reduced the width of the interval of search for this zero by 94%. This is typical. The procedure makes good progress even before the high rate of asymptotic convergence takes over.

9 A two-dimensional example

As another numerical example, we consider a simple two-dimensional problem introduced by Hansen and Sengupta [6]. In this problem, we seek the real and imaginary parts of the complex zeros of $(z^2 - 4i)(z - 1.7)$. Thus, we want the zeros of the function

$$f(x) = \begin{bmatrix} x_1^3 - 3x_1x_2^2 - 1.7x_1^2 + 1.7x_2^2 + 4x_2 \\ x_2^3 - 3x_1^2x_2 + 3.4x_1x_2 + 4x_1 - 6.8 \end{bmatrix}.$$

The zeros are

$$\begin{bmatrix} 1.7 \\ 0 \end{bmatrix}, \begin{bmatrix} 2^{1/2} \\ 2^{1/2} \end{bmatrix}, \begin{bmatrix} -2^{1/2} \\ -2^{1/2} \end{bmatrix}.$$

We used essentially the same algorithm as Hansen and Sengupta [6] except that we used slopes and inserted the g.i.a. procedure described in Section 7. The original Hansen-Sengupta algorithm required 81 steps to obtain the zeros with the error guaranteed to be less than 10^{-6}. Using g.i.a. as described above, we obtained the zeros with an error bound of 10^{-8} in 31 steps.

One must keep in mind, however, that the g.i.a. procedure involves more work per step. The saving in effort using g.i.a. depends on how much work is required to evaluate the function f and its derivatives.

The use of g.i.a. greatly improves convergence for this example and other problems of low dimension. From the nature of the procedure, it is obvious that convergence will be improved for larger problems. However, we have no experience on such problems.

10 Interval Newton methods

In order to discuss the topic we wish to consider in this section, it seems best to derive the interval Newton methods for the one-dimensional case.

Suppose $f(x)$ is continuously differentiable in an interval, X. Let x and y be in X. From the mean value theorem

$$f(y) = f(x) + f'(t)(y - x) \tag{10.1}$$

where t is between x and y and, hence, t is in X. If y is as zero of f, then $f(y) = 0$ and

$$y = x - f(x)/f'(t). \tag{10.2}$$

The interval Newton method is obtained by replacing t by its bound, X, in this relation and using the fact that any zero, y, of f in X is also in $x - f(x)/f'(X)$.

For the slope method, the relation corresponding to (10.2) is

$$y = x - f(x)/g(x, y). \tag{10.3}$$

Here, the left member, y, is the same quantity which occurs in $g(x, y)$ in the right member. But, in (10.2), y occurs on one side of the relation while a different quantity, t, occurs in the other.

This difference is crucial when we use g.i.a. When we replaced X by a single representative point, $x + u$, in the slope method, we are merely reverting to the case in which an occurrence of y in $g(x, y)$ is represented by $x + u$.

But we cannot do this for the interval Newton method because, in this case, X is bounding t, not y.

All is not lost, however. Instead of replacing X by a single point when using the procedure in Section 7, we can replace X by a subinterval of X known to contain t. Thus, when considering a single point $x + u$, we can replace X by the smallest interval (call it X') containing x and $x + u$. This subinterval, X', is of width less than or equal to half the width of X. This provides improved convergence.

Where the slope and Newton methods differ computationally is in (7.3) and (7.4). When these inequalities are used in the Newton method, we must replace u by $X' - x$ in the right members. We omit the details of the ensuing procedure.

The remarks we have made in this section hold equally well for the multidimensional case.

11 A hierarchy of methods

There are four methods of interest in this paper:

1. The standard interval Newton method using the derivative of f and o.i.a.

2. The interval Newton method using the slope function and o.i.a.

3. The standard interval Newton method using the derivative of f and g.i.a. See Section 10.

4. The interval Newton method using the slope function and g.i.a. See Section 7.

The slope function has a narrower interval “value”, in general, than the derivative evaluated over the same interval. Therefore method (2) (respectively, method (4)) will generally require fewer iterations than method (1) (respectively, method (3)). However, method (2) (method (4)) requires slightly more work per iteration than method (1) (method(3)).

Similarly, the use of g.i.a. rather than o.i.a. causes method (3) to take fewer iterations than method (2). In each case, the reduction in number of iterations is at the expense of more work per iteration.

The use of g.i.a. has a greater effect in reducing the number of iterations than does replacing the standard interval Newton method by the slope method. However, it requires quite a bit more work per step.

For the four methods, fewer iterations, but more work per step, are required as we go down the list. A proper comparison would be to determine run times for each method on a battery of test problems. We have not done so.

Such a comparison should be done using the best possible implementation of each method. The correct implementation of g.i.a. would be to use operator overloading in a language such as C++. We do not have such an implementation and cannot make a reasonable evaluation of the methods at this time.

However, we make the following guess based on limited experience: The run time will tend to decrease for methods farther down the list. We expect method (4) to be distinctly superior in efficiency; especially for problems with multiple zeros.

12 Another example

We now consider a numerical example comparing the methods discussed in Section 11.

We seek all the zeros of the polynomial given in (8.1). Note that $f(x)$ has three simple zeros at $x = -3$, -1, and 2 and one double zero at $x = 1$.

We chose the initial interval to be $X = [-4, 4]$. We solved this problem by the four methods using a final interval width tolerance of 0.000001. The number of iterations for each method is given in the following table.

Method	1	2	3	4
Number of iterations	93	54	30	16

13 Some remarks on the slope method

A method derived by the author in [4] yields same basic algorithm as the slope method. At one time, the author believed that they were, in fact, the same, differing only in that the slope method is a more organized version. In [4], he implied that they were the same method. We shall now show that they are not, and that the slope method is distinctly superior in most regards.

To conserve space, we shall not describe the method in [4]. Thus, our comments will be of limited interest to the reader. However, we feel that the comments will be of interest to a few specialists.

The derivation of the author's version requires that the point of expansion, x, be in X. (Compare the remarks in Section 9.) This is not the case for the slope method. As we have seen in Section 9, g.i.a. can improve the slope method more greatly if we do not need to have $x \in X$.

The author's method does have one advantage over the slope method. The latter cannot be used when f is irrational except for rational parts of f which might occur. However, the author's procedure is applicable for irrational functions. It provides an advantage over the standard interval Newton method.

14 Alternative Jacobians

We noted in Section 4 that when doing g.i.a. in the multidimensional case, we must decide how to linearize a product such as $u_i u_j$. We can replace u_i by its bounding interval, U_i, and have a result which is linear in u_j or we can replace u_j by U_j and have a result which is linear in u_i.

Sometimes, it is worthwhile to do both. Consider the procedure in Section 7 applied to a multidimensional problem. Assume we are solving the problem by the Hansen-Sengupta method using g.i.a.

When solving the i-th equation of (3.3) for the i-th variable, we replace all other variables by the interval bounding them. That is, we replace any variable, u_j, for $j \neq i$ by the interval, U_j. Therefore, when solving for the i-th variable, it would have been best to do all previous g.i.a. steps by replacing u_j in $u_i u_j$ by U_j instead of replacing u_i.

The coefficient matrix, M, in (3.3) is obtained as $M = BJ$ where J is the Jacobian or "slope Jacobian" of the vector function, f, whose zeros we seek. For each row of M that we compute in the manner just described, we have to compute a different Jacobian, J. This is obviously too much extra effort for problems of high dimension. However, it seems to be worthwhile for problems of low dimension.

15 A final note

It our g.i.a., we retain terms linear in u and linearize higher order terms. It is possible to extend g.i.a. to retain quadratic (or even higher) terms in a similar way. It is also possible to compute different terms to different orders.

Consider the problem of solving systems of nonlinear equations using the slope version of the Hansen-Sengupta method as described in Section 7. Suppose we retain quadratic terms when computing the off-diagonal terms of the matrix, M, in Section 3 and retain only linear term when computing the diagonal terms. Then the numerator of the function $R(u)$ in Section 7 will be a quadratic in u and the denominator will be linear in u. Therefore, the procedure discussed in Section 7 can be used with only minor modifications.

We have not tried using this alternative.

References

[1] Krawczyk, R. and Neumaier, A. *Interval slopes for rational functions and associated centered forms.* SIAM J. Numer. Anal. **22** (1985), pp. 604–616.

[2] Hansen, E. *Global optimization using interval analysis.* Marsel Dekker, New York, 1992.

[3] Neumaier, A. *Interval methods for systems of equations.* Cambridge University Press, Cambridge, 1990.

[4] Hansen, E. *Interval forms of Newton's method.* Computing **20** (1978), pp. 153–163.

[5] Hansen, E. *A generalized interval arithmetic.* In: Nickel, K. (ed.) "Interval Mathematics", Springer-Verlag, New York, 1975, pp. 7–18.

[6] Hansen, E. and Sengupta, S. *Bounding solutions of systems of equations using interval analysis.* BIT **21** (1981), pp. 203–211.

[7] Hanson, R. *Interval arithmetic as a closed arithmetic system on a computer.* Jet Propulsion Lab. Report 197, 1968.

[8] Kahan, W. *A more complete interval arithmetic.* Lecture notes for a summer course at the University of Michigan, 1968.

654 Paco Drive
Los Altos, CA 94024
USA