A New Method for Global Solution of Systems of Non-Linear Equations

LUKOMIR V. KOZIN

Department of Theoretical Electrodynamics, Faculty of Automation, Technical University of India, 1776

Submitted: 5 May 1977, accepted: 8 October 1977

Abstract: In this paper the problem of finding the set of all real solutions to a system of non-linear equations is considered. A general algorithm is suggested for solving the global solution problem. It is based on a transformation of the original system into a larger system of separable form. The global solution of the latter system is then found by a new iterative method which exploits the separability property of the system. An example illustrating the efficiency of the method is also provided.

1. Introduction

Let \( f: \mathbb{R}^n \rightarrow \mathbb{R}^n \) be a continuously differentiable function and let \( \mathbf{x}^n = (x_1, \ldots, x_n) \) be a given initial vector in \( \mathbb{R}^n \). The following global solution problem (GSP problem) has been considered in a number of publications.

**THE GS PROBLEM.** Given \( w \) and \( X^n \), find the set \( S(w, X^n) = \{ (y_1, \ldots, y_n) \} \) of all real solutions \((y_1, \ldots, y_n)\) to the system of equations

\[
y_i = f_i(x_{i-1}, \ldots, x_n), \quad i = 1, \ldots, n
\]

which are contained in \( X^n \), i.e., such that

\[
y_i \in X_i, \quad i = 1, \ldots, n.
\]

Currently, interval methods (method based on interval analysis techniques) seem to be the only methods which are capable of reliably solving the GSP problem (see (17)) and the references therein only. Indeed, an extension of the method to provide a set of small boxes \( x^{(n)}, n = 1, \ldots, p \) with \( f(x^{(n)}) \cap \mathbb{R}^n \), which contains the set \( S(x^{(n)}) \), is also known. However, this method has disadvantages mentioned above, namely, its computational complexity grows exponentially with the dimension of the system. Thus, for a system of 9 equations and a relatively small initial region \( X^n \) with \( X_i = [0, 10] \) a rather sophisticated method suggested in (17) requires billions of function evaluations in interval form to locate the unique zero of the GS problem considered. The authors' comment:
There were too many function evaluations in order to be proud of reaching the goal and to be really successful”, speaks for itself. The fact that known interval methods are rather inefficient for such moderate values of n is not difficult to explain. There seems to be a close reason for this.

1. Since known interval methods for solving the US problem are all one or another form of the interval bisection method, some interval extension $J[A]$ of the function $A(x)$ at $x$ is needed at each iteration of the method, here $X \subseteq X'$ denotes the box currently processed. As the initial iterations $X$ is large and $J[A]$ is generally rather a crude overestimation of the range $J[A]$ of $A(x)$ in $X$.

2. All methods of this group are associated with solving a linear interval system

$A'[X'] \cdot y = \phi$ (1.2)

with respect to $y$ for each iteration. Here $A(x)$ is a real matrix (ignoring for simplicity of exposition the interval arithmetic implementation of the method considered), $X$ is usually the center of $X$ while $A'[X']$ is an interval matrix what is either the interval extension $J[A]$ itself or in one way or another related to it. For instance, in methods using preconditioning (1)

$A'[X'] = B^T[B][X']$, $A(x) = B[y]$ (1.3)

where $B$ is some real matrix. In Hansen's method (2)

$B = [B(x)]^{-1}$ (1.3b)

Since the exact (optimal) interval solution $y$ of (1.2) is extremely hard to determine, an approximate interval solution $y'' \subseteq Y'$ is found, which is once again, rather a crude overestimation of $y$. Indeed, most often a coarse singular Gauss-Seidel procedure is used to compute $y''$ and it can be usually seen that the overestimation of $y$ by $y'$ becomes more and more pronounced as $n$ increases. Now let $A(x)$ denote the interval matrix associated with the log $A(x)$ i.e.

$A(x) = B(x)$

Obviously, this is the narrowest possible interval matrix for the current box which can replace $A(x)$ in (1.2). Furthermore, let $\tilde{Y}$ denote the optimal interval solution of the “best” linear interval system

$A[\tilde{Y}][x'] = \phi$ (1.4)

On account of inclusion monotonicity

$Y' \subseteq \tilde{Y} \subseteq Y''$ (1.4)

and the inclusion is proper and rather pronounced. Hence, initially, most obe

$X \subseteq Y''$ (1.4)
4. A NEW METHOD FOR GLOBAL SOLUTION OF SYSTEMS OF NONLINEAR EQUATIONS

and therefore the new box

\[ X' = X^* = X \]  

(1.7)

generated by the method for the next iteration is the same as the box \( X \) at the current iteration. In this case \( X \) is to be split into two subsets \( X_1 \) and \( X_2 \). One of them is used for further processing while the other is removed \( X \) and the iterative process continues on \( X \).

3. The above two factors contribute the third cause for low efficiency; there are too many splittings at the early stages of the complete process. Thus, in turn, gives rise to a long queue of boxes \( X' \) stored in \( L \), and awaiting processing. In fact, splitting stops and reducing the size of \( X \) by (1.7) starts only at that moment, when the width of \( X \) becomes so small that the first two factors associated with overestimation become insignificant.

4. Overestimation accounts for the ineffective functioning of the exclusion rule

\[ X' = X = \emptyset \]  

(1.8)

which deletes the current box \( X \) from further processing. Similarly to reducing, deleting occurs only at the last iterations when the width of \( X \) becomes small enough and the effects of overestimation become negligible.

It is to be noted that less effective reducing and deleting rules lead to a stronger clustering effect. An attempt to reduce the adverse effect of the first factor due to overestimation is associated with using interval slopes \( h \) [11], [12], [20] and interval derivatives in minimizing \( f(x) \). However, since this has enlarged the second factor for low efficiency, further experimental evidence showed that the overall improvement of the interval methods based on the use of interval slopes is still not satisfying, especially for more complex problems of larger \( n \) [9], [12].

In this paper, a new method for tackling the \( \Phi \)-problem seems to have considerably better numerical performance is suggested. It is based on the following approach. First, the original system (1.1) is transformed into a larger system of \( n' \) step-linear equations:

\[ f(x) = 0 \]  

(1.9a)

\[ x = x^{N} = \bar{x} \]  

(1.9b)

in such a way that the resulting system (1.9a) is in separable form, i.e.

\[ f(x) = \sum_{j=1}^{m} f_j(x - i \bar{x}) \]  

(1.10)

This transformation has recently become possible due to a result obtained in [19].

Thus a new interval method is applied to system (1.9). The global solution of (1.9) provides the global solution of the original problem (1.1). The new interval method
exploits in a rather efficient manner the separability property (1.10) and reduces, essentially, to solving the following linear system

$$A^{**Y} = A^{**Y}(X)$$

(1.11)

at each iteration $v$. Now, unlike (1.2), $A^{**Y}$ is a real matrix while $B^{**Y}(X)$ is an interval vector. In contrast to (1.2) the optimal interval solution $F$ to (1.11) is readily obtained:

$$F = A^{**Y}(X).$$

(1.12)

This is the main feature of the new interval method which accounts for the computational superiority over the previously known interval methods. Indeed, using $F$ (instead of $F^{**Y}$ as in the previous methods) results in a faster rate of convergence of the new method.

The paper is organized as follows. Section 2 presents the new method for solving systems of separable form. A numerical example of a system with $x = 0.01$ solved. The method is extended to systems of arbitrary form in Section 3. The overall efficiency of the resulting method is illustrated by the example considered in [17]. Concluding remarks are given in the last section.

2. Solving Separable Form Systems

2.1. The New Method

We consider the GS problem related to (1.9) with $F$ satisfying the separability property (1.10). Several methods (44, 51, 76-80, 121) have been proposed for solving this problem for the special case where $f_i(x_1) = e_i x_1$ is a linear function for all $i$, and $f(x)$ are assumed continuously differentiable (CD) functions.

A new method for globally solving the separable form system (1.9) will be presented here. In its present form, no restrictions on the functions $f_i(x)$ are imposed except for the requirement that they be continuous in $X$. Let $K$ denote the solution set, the current iteration. Then, a new interval approximation of a component $f_i(x)$ on $K$ is suggested. Unlike previous methods where the functions $f_i(x)$ are approximated at $K$, by enclosures using interval derivatives (44, 51, 76-80) or interval slopes (99, 121), the new approximation is chosen in the following form:

$$f_i(x) = R_i + a_i x_i, \quad x_i \in X_i$$

where $R_i = [b_i, b_i]$ is an interval while $a_i$ is a real number. Both $R_i$ and $a_i$ are to be determined such that the following inclusion property should hold

$$f_i(x) \subseteq R_i + a_i x_i, \quad x_i \in X_i$$

(2.1)

A simple and efficient procedure for finding $a_i$, $b_i$, and $R_i$ is suggested here for the case of CD functions. It is motivated by elementary geometrical considerations (Figure 1a) and can be readily adapted for the case of functions that are only continuous (Figure 1b).
PROCEDURE 1. First, compute

\[ \hat{y}_1 = f_1(x_1), \quad \hat{y}_2 = f_2(y_2). \]  \hspace{2cm} (2.3)

Then \( y_1 \) is defined as the slope

\[ y_1 = \frac{\hat{y}_2 - f_2(y_2)}{1 - f_1(x_1)}. \] \hspace{2cm} (2.4)

Let

\[ f_1(x_1) = b_1 + \alpha_1 \] \hspace{2cm} (2.5a)

be a straight line such that

\[ f_1(x_1) \leq \hat{y}_1(x_1), \quad x_1 \in X_1. \] \hspace{2cm} (2.5b)

Thus, \( b_1 \) is the lowest possible line of slope \( \alpha_1 \) bounding \( f_1 \) from above in \( X_1 \); see Figure 1a. Similarly, let the straight line

\[ b_2(y_2) = b_2 + \alpha_2 y_2 \] \hspace{2cm} (2.6a)

have the property

\[ f_2(y_2) \geq b_2(y_2), \quad y_2 \in X_2. \] \hspace{2cm} (2.6b)

Thus, \( b_2 \) is the highest possible line of slope \( \alpha_2 \) bounding \( f_2 \) from below in \( X_2 \); see Figure 1a.

The unknown constants \( b_1 \) and \( b_2 \) are determined as follows. Let \( p_1(x_1) \) denote

the derivative of \( f_1(x_1) \), i.e., \( p_1(x_1) = \frac{df_1}{dx_1} \). Then the following functional equation of one variable,

\[ p_1(x_1) = -\alpha_1 = 0 \] \hspace{2cm} (2.7)
is solved globally in $X_k$. Let $X_k^0, k = 1, 2, \ldots, K$, denote the $k$-th solution of (2.7) in $X_k$. Furthermore, let $X^0 = X_0$. Now compute the quantities

$$X_k^0 = f(X_k^0) - x_kX_k^0, \quad k = 0, 1, \ldots, K$$

(2.8)

As is easily seen from geometrical considerations (Figure 1a)

$$b_0 = \min\{X_k^0, \quad k = 0, 1, \ldots, K\}$$

(2.9a)

and

$$b_k = \max\{X_k^0, \quad k = 0, 1, \ldots, K\}.$$  

(2.9b)

Finally, the interval

$$B_0 = [b_0, b_1]$$

(2.10)

is found. Thus, the linear interval approximation (2.1) having the encasing property (2.2) has been determined.

Remark. For simplicity of notation and presentation, the quantities $x_0, x_1, x_2, x_3, x_4, x_5, \ldots$, $r_0, r_1, \ldots$, $C_0, C_1, C_2, \ldots$, $\Delta_0, \Delta_1, \ldots$ are assumed to be real numbers. However, in the actual implementation of the procedure, they must be computed using interval arithmetic. Therefore, these are in fact intervals although their widths are rather small. Rigorous bounds on all sorts of error (2.7) in $X_k$ can be computed using an appropriate interval algorithm (e.g., see [1], Ch. 7 and 8).

A simple illustrative example of the procedure for determination $b_0$ and $b_1$ will be considered now. Assume that $f(x)$ is a cubic equation in $x$. Then (2.7) is a quadratic equation. Let $x_0$ and $x_0'$ denote the corresponding zeros. Several cases are possible depending on whether both zeros belong to $X_k$ or not.

Case A. Both zeros belong to the interval $X_k$ as shown in Figure 1a. Now, computing $X_k^0, k = 0, 1$ by (2.8) we see that in this case

$$b_0 = b_k^0 = f(0) - x_k0'$$

while

$$b_1 = b_k^1 = f(0) - x_k0.$$  

Case B. The first zero belongs to $X_k$, while the second does not. In this case, we need to compute only $X_k^0$ and $X_k^1$ since now $k = 0, 1$; thus

$$b_0 = b_k^0 = f(0) - x_k0'$$

$$b_1 = b_k^1 = f(0) - x_k0.$$  

Case C. Only the second zero belongs to $X_k$. It is easily seen that in this case $b_k^0 = b_k^0$ and $b_1 = b_k^0$.  

A new method for global solution of systems of nonlinear equations.

Using the Procedure 1 a real matrix is introduced

\[ A = [-a_{ij}] \]  \hspace{1cm} (2.31)

and an interval vector \( B = (B_1, \ldots, B_n) \) is formed with

\[ B_i = \sum_{j=1}^{n} a_{ij} B_j, \hspace{1cm} 1 \leq i \leq n \]  \hspace{1cm} (2.32)

On account of (2.10), (2.11), (2.12), and (2.13):

\[ f(x) = \sum_{i=1}^{n} a_{ij} x_j + B_i, \hspace{1cm} x \in X, \hspace{1cm} i = 1, \ldots, n \]  \hspace{1cm} (2.33)

or in vector form:

\[ f(x) = A x + B, \hspace{1cm} x \in X \]  \hspace{1cm} (2.34)

If \( y \) is a solution of (2.11) in \( X \) then \( f(y) = 0 \) and by (2.14):

\[ 0 \leq -B y, \hspace{1cm} y \in X \]  \hspace{1cm} (2.15)

Now we can state the main result of the section.

**THEOREM 2.3.** All the solutions \( x \) to

\[ f(x) = 0 \]  \hspace{1cm} (2.16)

contained in \( X \) are also contained in the solution set \( S(x) \) of the system

\[ A x + B = 0, \hspace{1cm} x \in X \]  \hspace{1cm} (2.17)

where \( B \) is any real vector contained in \( B \).

The proof is straightforward on account of (2.16) and (2.17).

Since \( B \) is an interval vector the set \( S(x) \) is a convex polyhedron. Indeed, (2.17) is in fact a system of linear inequalities and 2n two-sided linear inequalities.

**THEOREM 2.2.** All the solutions \( y \) to (2.16) in \( X \) are also contained in the intersection

\[ P(X) = S(X) \cap X \]  \hspace{1cm} (2.18)

Since \( S(X) \) and \( X \) are convex polyhedra it is seen from (2.18) that \( P(X) \) is also a convex polyhedron.

**COROLLARY 2.1.** If \( P(X) \) is empty, i.e. if

\[ S(x) \cap X = \emptyset \]  \hspace{1cm} (2.19)

the system (2.16) has no solution in \( X \).
Let \( R(P, X) \) denote the interval hull of \( P(X) \), that is, the smallest interval vector (box) containing \( P(X) \). Consider the following iteration procedure
\[
X^{(k+1)} = R(P, X^{(k)}), \quad k \geq 0
\]  
(2.20)

Procedure (2.20) could be used for designing a method for finding all real solutions to (1.1) in \( \mathbb{X} \). Such an approach seems to be rather costly since linear programming problems are to be solved at each iteration to determine \( R(P, X^{(k)}) \). Therefore, a simpler and, presumably, more efficient procedure is suggested here.

Let \( R(S, X) \) denote the interval hull of \( S(X) \). Then it follows from (2.17) that \( R(S, X) \) is given by the formula
\[
R(S, X) = A^{-1}R
\]  
(2.21)

Let \( C = A^{-1} \) (assuming \( A \) is invertible) and \( Y = R(S, X) \). It follows from (2.21) that the components of \( Y = [E, T] \) of \( Y \) are given by the formulae
\[
E = \sum_{j=1}^{n} I_{D_{j}},
\]
(2.22)

with
\[
I_{D_{j}} = \begin{cases} 
-c_{j} & \text{if } c_{j} \geq 0,
-c_{j} & \text{if } c_{j} < 0.
\end{cases}
\]
(2.23)

\[
T = \sum_{j=1}^{n} I_{C_{j}},
\]
(2.24)

with
\[
I_{C_{j}} = \begin{cases} 
-c_{j} & \text{if } c_{j} \geq 0,
-c_{j} & \text{if } c_{j} < 0.
\end{cases}
\]
(2.25)

Now, the following iterative procedure can be used as the basis for solving the GS problem considered.

**PROCEDURE 2.** Let \( X^{(0)} \) be a convex box. Using Procedure 1 determine \( x^{(0)} \) and \( y^{(0)} \) corresponding to \( y^{(0)} \). By formulae (2.22) and (2.23) compute \( x^{(k)} \). The iterative procedure is then defined as follows
\[
x^{(k+1)} = y^{(k)}, \quad k \geq 0.
\]  
(2.26)

The procedure may result in three outcomes.

A. The sequence \( x^{(k)} \) converges to a solution \( x^{*} \) as \( k \rightarrow \infty \). Actually, the iterations are stopped whenever the width of \( x^{(k)} \) becomes smaller than a constant \( \epsilon \) (arbitrarily with respect to \( x \)). Let \( x^{(k)} \) be approximated by the convex set \( \mathcal{C} \) of \( x^{(k)} \) and \( x^{*} = \text{sub} \in \mathcal{C} \). If
\[
\|x^{*} - x^{(k)}\| \leq \epsilon,
\]  
(2.27)

then \( x^{*} \) is accepted as a solution to (1.9).

B. The sequence \( x^{(k)} \) is bounded, and the iterations are resumed.

C. The sequence \( x^{(k)} \) is not bounded, and the iterations are terminated.

(\( \epsilon \) is the accuracy of \( x^{*} \) with respect to the system of equations; \( x^{*} \) is the solution of (1.9)).
A NEW METHOD FOR GLOBAL SOLUTIONS OF SYSTEMS OF NONLINEAR EQUATIONS 133

B. At some \( i \):
\[ x^{(\infty)} = x^{(i)} = \mathbf{R} \quad (2.28) \]

Since \( F(x, x^{(i)}) \) is lower triangular, it follows from Corollary 2.5 that system (2.26) has no solution as \( x^{(i)} \) (2.28) becomes valid. In this case \( x^{(i)} \) is discarded from further consideration.

C. The sequence \( x^{(i)} \) converges to a fixed interval \( \mathbf{X} \). In practice, the procedure is stopped whenever the reduction in the volume \( V(x^{(i)}) \) of the current box \( x^{(i)} \) is smaller than a constant \( \epsilon \), i.e., if \( V(x^{(i)}) / V(x^{(0)}) \) is smaller than a constant \( \epsilon \). In this case \( x^{(i)} \) is split along its wider side into two boxes \( x^{(i)'} \) and \( x^{(i)''} \), respectively. The right box is moved into a list for further processing. The left box is retained \( x^{(i)} \) and the iterative procedure (2.22) to (2.23) is repeated.

\section{Procedures 2 to 5}

The convergent procedure can be improved if, in addition to operator \( F \), a componentwise operator \( H \) is introduced. Then, whenever a reduction of a component \( x^{(i)} \) occurs, this will be used immediately for reducing (if possible) the remaining components \( x^{(i)}' \), \( i = 1, \ldots, n \).

To introduce the componentwise algorithm we need to modify Procedure 1 and Procedure 3 to Procedures 2 and 4, respectively.

\section{Procedures 6 to 8}

For reasons to become clear later, it is expedient to introduce two new real numbers \( R_1 \) and \( R_2 \), and to some \( h_1 \) and \( h_2 \) in the \( j \)-th row of \( R_1 \) (lower and upper end points) and \( h_2 \) (upper end points), respectively.

To introduce Procedure 6, formula (2.26) is written in componentwise form
\[ x^{(i)} \times x^{(i)} = x^{(i)} \quad (2.28) \]

Now, according to the idea to get the most out of the reduction of \( x^{(i)} \), we check whether
\[ x^{(i)} < x^{(i)} \quad (2.30) \]

If (2.30) holds (i.e., if the reduction is greater than some threshold \( \epsilon \)), i.e.,
\[ x^{(i+1)} < x^{(i)} \quad (2.31a) \]

or
\[ x^{(i+1)} < x^{(i)} \quad (2.31b) \]


\section{Procedures 9 to 11}


\section{Procedures 12 to 14}


\section{Procedures 15 to 17}


\section{Procedures 18 to 20}
A NEW METHOD FOR GLOBAL UTILITY OF SYSTEMS OF NONLINEAR EQUATIONS

THEOREM 2.3. Let \( f : D \subseteq \mathbb{R}^m \rightarrow \mathbb{R}^m \) be a continuously differentiable function in the domain \( D \) and \( X^{(0)} \subseteq D \) be arbitrary. The iteration operator

\[
K(X^{(k)}) = R(X^{(k)}) = \alpha^{-1}(A(X^{(k)})) \quad k \geq 0
\]

where \( A(X^{(k)}) \) and \( R(X^{(k)}) \), defined in Procedure 1, correspond to the current box \( X^{(k)} \). Then, if at some \( k \)th iteration

\[
R(X^{(k)}) \subseteq X^{(k)}
\]

the above inclusion implies the existence of a solution in (2.15) in \( X^{(k)} \).

Proof. Let \( t \) be a solution in (2.15). Then (2.15) can be transformed into the fixed point format

\[
x = C(x) + P(x)
\]

where \( C : X^{(k)} \rightarrow R^n \) and \( P : X^{(k)} \rightarrow R^n \) are to be determined. We choose

\[
C(x) = C^{(0)} = x^{-1}(P(x))
\]

(2.37a)

\[
C(x) = C^{(0)} = x - h \beta(x)
\]

(2.37b)

Obviously, for any \( x \in X^{(k)} \)

\[
P(x) = R(X^{(k)})
\]

Thus, if (2.36) holds, then \( P \) maps \( X^{(k)} \) into itself. Therefore, by the Banach fixed point theorem \( P \) has a fixed point in \( X^{(k)} \), and hence \( t(x) = 0 \) is a solution in \( X^{(k)} \).

The theorem can be extended in a straightforward manner to the comprehensive version of the present method as introduced by Procedure 5.

Now we shall consider the convergence rates of the sequence (2.26) towards a solution. We need the following well-known facts from interval analysis. For an interval \([a,b] \) (\( a \leq b \)), and an interval vector \( X = (X_1, \ldots, X_n) \) define the widths of \([a,b] \) and \( X \) as follows

\[
\omega([a,b]) = b - a
\]

(2.38a)

\[
\omega(X) = \max_i \omega(X_i)
\]

(2.38b)

If

\[
Y = AX
\]

where \( A \) is a real matrix with elements \( a_{ij} \) then

\[
\omega(Y) = \sum_{i=1}^{n} \omega(a_{ij}X_j)
\]

(2.40a)

and

\[
\omega(Y) \leq \|A\| \omega(X)
\]

(2.40b)
where
\[ |k| = \max \sum |a_i| \]  \hspace{1cm} (2.4k)

Additionally, we shall make use of the following lemma.

**Lemma 2.1.** Let \( f(x) \) be a CD function in \( X'' \). Furthermore, let \( B_0 \) be determined as in Procedure 1. Then for a narrow enough interval \( X \subset X'' \) (neglecting terms of order higher than 2),
\[ u(B_0) = B_0 \cdot u^2(x_0) \]  \hspace{1cm} (2.4l)

**Proof.** For simplicity of notation the subscripts will be temporarily dropped. If \( X \) is narrow enough, equation (2.7) will have a unique solution \( x' \) in \( X \). With no loss of generality, assume that \( u(x) \) is convex in \( X \), i.e. \( f(x') > 0 \) in \( X \). In this case
\[ \bar{u} = f(x) - a_0 \quad b = f(x') - a_0' \]

But \( v(x) \) can be expressed approximately as
\[ H(x) = f(x') + f(x)(x'-x) + \frac{1}{2} f''(x')(x'-x)^2 \]

Taking into account that \( f'(x) = u \) we have
\[ w(x) = \bar{u} - \frac{1}{2} f''(x)(x' - x)^2 \]

We can relate \( x' - x \) to \( w(x) \) as follows,
\[ x' - x = u(x) - x = w(x) \]

Hence
\[ w(x) = \frac{1}{2} f''(x) u^2(x) + bh^2(x) \]

which completes the proof. \( \square \)

Now we are in a position to state the following theorem.

**Theorem 2.4.** Suppose that
\[ X^{(0)} \subset X^{(1)} \]  \hspace{1cm} (2.4m)

holds for all \( k \geq 0 \), i.e. that the iterative procedure (2.26) converges to a solution \( x^* \) and the Jacobian \( J(x) \) is non-singular in \( X^{(0)} \). Then the convergence rate towards \( x^* \) is quadratic.
Proof: It follows from (2.21) and (2.41a) that
\[ w(t) = \sum_{j=1}^{n} \mu_j w_j(t) \]  
(2.43)
By using (2.12) and (2.41)
\[ w_j(t) = \sum_{i=1}^{n} \rho_i u_i(t) \]  
(2.44)
Hence
\[ w(t) = \sum_{j=1}^{n} \sum_{i=1}^{n} \rho_i u_i(t) \]  
(2.45a)
with
\[ \rho_i = \sum_{j=1}^{n} \mu_j \beta_j(i) \]  
(2.45b)
Now it will be shown that
\[ w(t) \leq |H| \mu(t) \]  
(2.46a)
where
\[ H = \{ (R_j) \} \]  
(2.46b)
Indeed
\[ w(t) = \max w(t) = \max \sum_{i=1}^{n} R_i u_i(t) \leq \max \sum_{i=1}^{n} |R_i| \mu_i(t) \]  
(2.47)
Now (2.46) follows from (2.47), (2.50b), and (2.46b).

Due to the invertibility of \( A(t) \), matrix \( C \) has bounded elements \( c_{ij} \). As \( f \) is a C0 function, \( f(x) \) is in \( X^0 \) and the coefficients \( \beta_j \) are also bounded. Hence (as seen from (2.55)) \( |f(t)| \leq \tilde{f} \in X^0 \). Finally,
\[ w(t) \leq \tilde{w}(t) \]  
(2.48)
which completes the proof of the theorem. \( \square \)

Now we shall consider the question of uniqueness.

THEOREM 2.5 If at some iteration \( h_0 \) of the iterative procedure (2.26) the condition
\[ \sum_{i=1}^{n} R_i u_i(t) < \tilde{w}(t) \]  
(2.49)
is satisfied and the real functions \( f(x) \) are all strictly monotone in \( X^{0(\cdot)} \), then:
a) there is a unique solution \( u \) to \( f(u) = 0 \) in \( X^{0(\cdot)} \).
b) the modified procedure

\[ x^{(k+1)} = x^{(k)} - \beta^{(k)} \gamma^{(k)} \], \quad k \geq 0 \tag{2.50} \]

where \( C^{(k)} = (\gamma^{(k)})^{-1} \) is a corrector matrix while \( \beta^{(k)} \) is computed as in Procedure 1 using, however, the same slopes \( a^{(k)}_t \) at each iteration, converges to \( x^* \).

Proof. Let for simplicity of notation \( C = C^{(0)} \) and \( \beta = 0 \). On account of (2.49) and (2.50)

\[ x^{(k)} \leq \beta^{(k)} C^{(k)} x^{(k)} \tag{2.51} \]

so

\[ w(x^{(k)}) \leq w(x^{(k)}) \tag{2.51a} \]

Taking into account (2.51) and the monotonicity of \( f(x) \), it is easily seen that

\[ \beta^{(k)} C^{(k)} x^{(k)} \leq x^{(k)} \tag{2.52} \]

(Indeed, it follows from (2.51) that \( x^{(k)} \leq x^{(k)} \) for each \( k \). As is readily seen, the above inclusion and the strict monotonicity of \( f(x) \) in \( x^{(k)} \) imply that \( \beta^{(k)} C^{(k)} x^{(k)} \leq x^{(k)} \) for all \( k \) and \( j \).)

These inclusions lead to (2.52); on account of (2.12) \( \beta^{(k)} \) and \( x^{(k)} \)

\[ \beta^{(k)} = (\beta^{(k)})^{(k)} C^{(k)} x^{(k)} \tag{2.52a} \]

Thus, we have shown that procedure (2.50) generates a nested sequence \( (x^{(k)}) \) of descending widths \( w(x^{(k)}) \). By Theorem 2.3 \( x^* \) exists and is unique because \( w(x^{(k)}) \to 0 \) as \( k \to \infty \). This completes the proof of Theorem 2.3. \( \square \)

2.3. A Numerical Example

The numerical performance of the present method has been tested on several systems of equations of the form

\[ f(x) = 0 \tag{2.53a} \]

where

\[ g(x) = g(x), \quad i = 1, \ldots, n \tag{2.53b} \]

and \( f \) is a constant matrix with \( \sigma \) ranging from 2 to 20. In all the cases \( g(x) \) are continuously differentiable functions. The method has been implemented in its basic form.

To illustrate the improved numerical efficiency of the new method an example (modified in [9] and [18]) will be presented here. The system (2.53) is now given by

\[ g(x) = 2.5x_i^2 - 10.5x_i + 11.8, \quad i = 1, \ldots, 10, \]

\[ H = (A^i_j) \quad \text{with} \quad a_i^i = -1, \tag{2.53c} \]

\[ s = (-1, -2, -3, -4, -5). \tag{2.53d} \]
The initial box $X^0$ is defined by

$$x_i, c_i \in [-1, 1], \quad i = 1, \ldots, n.$$ 

The accuracy $\epsilon$ has been chosen to be $10^{-4}$.

Two interval methods were applied in [9] to solve the GS problem associated with (2.55). The first method denoted by $M_1$ is a variant of a method due to Aulds and Hertleiner [1]. It is based on the use of interval derivatives.

The second method designated as $M_2$ is an improved version which uses interval slopes. This paper's method denoted by $M_3$ has also solved the problem considered and has found within the same accuracy $\epsilon = 10^{-2}$ all the 9 solutions contained in $X^0$. However, the data in Table 1, concerning the number of iterations $N$, required to solve the GS problem considered reveals that the present method is vastly superior to the other two interval methods as regards computer time.

On account of the fast convergence rate the new method has also improved characteristics as regards memory volume requirements. Indeed, the maximum number of boxes $n_b$, stored during computation reached the value of 3 for method $M_3$ while $n_b$ was maintained higher for $M_1$ and $M_2$. It should be stressed that no clustering effect has been observed in solving the present example by $M_3$. In contrast, among the two previous methods, even the better method $M_2$ generated dozens of clustering boxes, thus requiring much bigger memory volume.

In practice (at the early stage of various design problems) one does not always need to solve the GS problem completely; most often, it suffices to only find the number of solutions $N$ contained in $X^0$. This simple problem can be efficiently solved by a modification of the present method denoted $M_3A$. The variant $M_3A$ is based on Theorem 2.4. The iterative procedure (2.2h) is terminated whenever condition (2.4h) is satisfied and is resumed by removing a new box from the list if $L$ is not empty. The corresponding number of iterations for the example considered is given in the last column of Table 1.

3. Solving General Form Systems

3.1. Transformation to Separable Form

In this section, the method presented in the previous section will be extended to systems of general form. More specifically, the system's components and are assumed to be factorizable functions [80], i.e., functions that are composed of four arithmetic operators (+, −, ·, ÷), unary operators (sin, exp, log, sqrt, abs, etc.) and the power operation ($^c$).
The approach herein adopted is to transform the general form system (1.1) to the separable form (3.1) (12). The theoretical basis for such a transformation is a famous theorem due to Kotelnikov's work [10] published in early 1957. However, its proof is not constructive and only recently has a simple algorithm been proposed [19] to convert factorial-like functions into separable functions automatically by computer. To maintain completeness, several basic facts from [19] will be briefly presented here.

Let $f_1$ and $f_2$ be subfunctions of $y$ containing at least one variable. Consider the following three cases:

1. $f = f_1 \cdot f_2$.
   
   
   
   

2. $f = f_1 + f_2$.
   
   
   

3. $f = f_1^2 + f_2^2$.
   
   
   

If both $f_1$ and $f_2$ contain only one and the same variable then $f$ is obviously separable in all the three cases.

If $f_1$ contains only one variable and $f_2$ contains only another variable then the functions (3.1) to (3.3) can be easily transformed into separable form as follows:

The transformation of (3.1) is:

$$f = f_1 \cdot f_2 = \frac{1}{2}(f_1^2 - (f_2^2 - (f_1 f_2)^2)),$$

(3.4)

The second case is reduced to the first by letting

$$f_2 = 1 / f_1,$$

and applying (3.4).

For the third case the transformation suggested in [19] is:

$$f = f_1 f_2 \rightarrow$$

$$y_1 = f_1 + \log f_1$$

$$y_2 = \left((\log f_1^2) - (\log f_2^2)\right) / 2,$$

(3.5)

It should be mentioned that (3.5) is valid only if $f_1 > 0$ for all values of its argument.

If both $f_1$ and $f_2$ contain more than one variable then we first introduce auxiliary variables and apply the above approach. To illustrate this possibility consider formula (3.3). In this case:

$$f = f_1 \cdot f_2$$

$$f_1 = f_1$$

$$f_2 = f_2$$

$$y_1 = f_1$$

$$y_2 = f_2$$

$$y_3 = y_1 y_2$$

$$f = (y_1 y_2) \cdot (y_1 y_2)$$

$$f = (y_1^2 - (y_2^2 - (y_1^2 - (y_2^2))^2)) / 2.$$
3.2. An Illustrative Example

We take up the system considered in [17]

\[\begin{align*}
    u_k(x) &= 0, \quad k = 1, \ldots, 4, \\
    \beta_k(x) &= 0, \quad k = 1, \ldots, 4, \\
    y(x) &= 0, \\
\end{align*}\]

where \( x \in \mathbb{R}^n \)

\[\begin{align*}
    u_k(x) &= 1 - x_k + \left[ e^{x_k^2} - x_1 - x_2 - \cdots - x_{k-1} - 1 \right] \\
    \beta_k(x) &= 1 - x_k + \left[ e^{x_k^2} - x_1 - x_2 - \cdots - x_{k-1} - 1 \right] \\
    y(x) &= x_n - x_n, \\
\end{align*}\]

The numerical constants \( x_k \) are given in [17]. As in [17], the problem is to establish computationally that (3.7) has a unique solution in a given initial box \( A^n \) with sides:

\[\begin{align*}
    x_k^0 &= (x_k^l, x_k^u), \quad k = 1, \ldots, n, \\
\end{align*}\]

where \( n = 9 \).

To reduce (3.7) to separable form we first introduce the auxiliary variables:

\[\begin{align*}
    x_{k+9} &= e^{x_k^l} - e^{x_k^u} + e^{x_k^l} - e^{x_k^u}, \\
    \lambda_k &= x_k - x_k^l - e^{x_k^l} - e^{x_k^u} + 1, \\
    \gamma_k &= x_k^l - x_k^u, \\
    \sigma_k &= x_k^u, \\
    \rho_k &= 1 - x_k, \\
\end{align*}\]

Using (3.7) and (3.8) we obtain the following system of 20 equations:
\[ \begin{align*} 
& x_{kn} - x_{n} = k x_{n} y_{k} - x_{n} y_{k} = 0, \quad k = 1, \ldots, 4; \\
& y_{k} y_{n} - y_{k} y_{n} = 0, \quad k = 1, \ldots, 4; \\
& x_{kn} + x_{kn} = x_{kn} + x_{kn} = 0, \quad k \neq 1, \ldots, 4. 
\end{align*} \]

Since equation (3.10) contains the products \( x_{kn} y_{k} \) and \( x_{kn} y_{n} \), the form is called separable.

One way to get a system of separable form is to minimize the products in (3.11), using (3.4) and (3.6). Thus, a final system of separable form consisting of a total of 43 equations will be obtained.

An alternative approach to reach separability is suggested here. It is based on the following idea. Consider the product

\[ (x - x_{s})(y - y_{s}) \]

where \( X \) and \( Y \) are intervals, \( x_{s} \) and \( y_{s} \) are the centres of \( X \) and \( Y \), respectively, then

\[ \begin{align*} 
& 2 a_{k} + 2 a_{n} = a_{kn} + a_{kn} + a_{kn} + a_{kn} = a_{kn} + a_{kn} = a_{kn} + a_{kn}. 
\end{align*} \]

When \( a \in X \) and \( b \in Y \), the central variables \( a \in X \) and \( b \in Y \), where \( R_{a}, R_{b} \), are the nulls of \( X \) and \( Y \). Let \( R = R_{a} \cap R_{b} \), it follows from (3.10) that

\[ \begin{align*} 
& x \in a_{kn} + a_{kn} + a_{kn} + a_{kn}. 
\end{align*} \]

Thus, the product \( x \) has been enclosed by an interval expression in separable form, i.e.,

\[ \begin{align*} 
& x \in a + b \quad \text{w.r.t.} \quad R_{a}, R_{b}. 
\end{align*} \]

An appealing feature of the alternative approach is the fact that it converts the separable form into separable form without introducing new variables and equations. Indeed, using the new approach the original system (3.7) has been transformed into a separable form system (3.11) of only 20 equations (rather than to 43 equations if the standard approach of [19] were applied).

Two algorithms have been elaborated to solve the GI problem associated with (3.7). They are based on the simplex MAO of the present method and on the form of the
method using the uniqueness test from Section 2.2. To present these algorithms we need to distinguish between the original variables $a_i$, $1 \leq i \leq n$ and the auxiliary variables $a_i$, $10 \leq i \leq 20$. So we introduce the real vectors

\[ a = (a_1, \ldots, a_n), \]

\[ b = (b_1, b_2) \]

and the interval vectors $X_a, X_b$, and $X$. Now (3.4) can be written in the form

\[ a = f(X_b). \]

Hence

\[ X_a \subseteq f(X_b), \]  

(3.14)

where $f(X_b)$ denotes the range of $f$ under $f$. Since each original variable occurs only once in (3.3), $f(X_b)$ can be computed by a single interval computation according to a well-known theorem from interval analysis.

**ALGORITHM A1**

**Step 1 (Initialization).** Using the initial box $X^{(0)}$ given by (3.3) and relation (3.14), the corresponding initial auxiliary vector $X^{(0)}_a$ is computed. Thus, the initial vector

\[ X^{(0)} = (X^{(0)}_a, X^{(0)}_b) \]

is formed.

**Step 2.** The vector MUA is applied to the box $X^{(0)}$.

The second algorithm is a modification of A1; therefore, only the relevant differences will be noted.

**ALGORITHM A2**

The modifications associated with this algorithm are, essentially, related to the case where the current box $X$ is reduced in size.

At the current iteration, $X$ is transformed by Procedure 1 to yield vector $Y$. Now a new vector $X'$ is obtained

\[ X' = Y \cap X. \]

It is partitioned into two parts $X'_a$ and $X'_b$ corresponding to the original and auxiliary variables. Using $X'_b$, (3.14) we obtain

\[ X'_a = f(X'_b). \]

Next we find the intersection

\[ X'_a = X'_a \cap X'_b. \]
to form the vector
\[ X^1 = (a1, a2) \]
corresponding to the new iteration \( r = 1 \). Finally, \( X^1 \) is entered \( X \) and the iterative process continues as in A1.

It is also to be noted that unlike A1 the width \( w \) of the current box \( X^1 \) needed to assess the accuracy by comparison with \( a1 \) is determined on the basis of the original vector \( X2 \). Similarly, only \( X1 \) is used in assessing the reduction of the box.

The system (3.7) is known to have a solution which is approximately
\[ x^* = (0.9, 0.45, 1.2, 8, 5.1, 12) \]
The initial box \( X0 \) for the original variables was chosen to be centred at \( x^* \), that is, each component \( x_{ik}^* \) was expressed in the form
\[ X_{ik}^0 = x_{ik}^* + r \cdot a_i \]
Table 2 provides data about the number of iterations needed by A1 and A2 to establish uniqueness of the solution \( x^* \) in \( \mathbb{R}^n \) when \( r = 0.05 \).

It is worthwhile mentioning that the sophisticated version of the interval Newton method from [17] (using elaborate valiant iteration strategies) establishes the existence and uniqueness of the solution to (3.7) in the box dropping for simplicity only 5 decimal places
\[
\begin{array}{c|c}
0.0000 & 0.0000 \\
0.4400 & 0.4500 \\
1.7902 & 1.8000 \\
1.9902 & 2.0000 \\
2.9902 & 3.0000 \\
4.9902 & 5.0000 \\
1.9902 & 2.0000 \\
\end{array}
\]
Using \( X0 \), we can compute its width \( w = 0.00196 \), or approximately 0.001. It is seen that with \( r = 0.05 \) the width \( w = 0.1 \) of the box \( X1 \) is approximately 50 times larger in comparison to that of the box \( X0 \).

4. Conclusion

In this paper, the problem of finding (within preset accuracy) the set of all real solutions to a system of nonlinear equations (3.4) contained in a given box \( X0 \).
A NEW METHOD FOR GLOBAL SOLUTION OF SYSTEMS OF NONLINEAR EQUATIONS

the GS problem has been considered. The only assumption about the nonlinear functions involved is that they are continuous in \( \mathbb{R}^n \).

A new method for solving the GS problem has been suggested. It is based on a transformation of the original system (1.1) into a new system (1.9) of separable form (1.10). The latter system is then solved for each of the two limits by a new algorithm which exploits the separability property (1.10). More specifically, each function \( f_k(x) \) from (1.10) is approximated by a linear function \( f_k(x) \) defined by (2.1). Thus, at each iteration of the method a linear interval system (1.11) with a real matrix \( A \) is solved. This, in turn, distinguishes the present method from the other known interval methods where a much more complex linear interval system having an interval matrix and a real right-hand side vector is to be solved.

Experimental data show that, as regards computer time and memory volume requirements, the present method exceeds considerably the other known methods for solving the global solution problem considered. It should also be mentioned that so far no deterioration has been observed.

There seems to be several possibilities for further improvement of the numerical efficiency of the new method. Since the equations form of system (1.9) is, generally, much larger than the original system but has, at the same time, a rather sparse structure, one approach should appeal to sparse matrix techniques in implementing the computer version of the method. Work is presently in progress to accomplish this objective. Using such an approach it would be interesting to compare the relative efficiency of the two possible formulations to separability: the separable form suggested and employed in this paper or the standard separability form from (1.9). Another possibility is to incorporate into the present method ideas from the conjugate projection method. Preliminary limited experimental results indicate that such an approach seems to be rather promising. Finally, it remains to investigate the convergence properties of the alternative iterative procedure (2.3) which is based on linear programming implementation; it may turn out that such an approach could lead to a method of improved overall numerical efficiency.

Acknowledgments

This work was supported by the Bulgarian Research Foundation under Contract No. 79-K-5500.

The author is grateful to one of the referees for his valuable questions and remarks.

References
