On the Solutions of Equations for Nonlinear Resistive Networks

By A. N. WILLSON, JR.

(Manuscript received December 13, 1967)

Several theorems are proved concerning the solutions of equations that arise in the study of resistive nonlinear electrical networks. The first, an existence and uniqueness theorem, applies to equations describing an interesting class of networks which includes certain active and nonreciprocal networks for which the existence and uniqueness of solutions has not previously been established. A method of computing bounds on the location of the solutions is given, and two iterative techniques are presented for computing the solutions. It is proved that the iterative techniques converge for a subclass of the equations which also includes equations describing certain active and nonreciprocal networks. Finally, the rate of convergence of the iterative techniques is compared with that of another well-known iterative technique and some practical computational aspects are pointed out. Computations for two example problems, not reported here, were carried out to show the practicality of applying these iterative techniques to the equations of specific networks.

I. INTRODUCTION

In this paper we consider the solution of the equation

$$F(x) + Ax = B \tag{1}$$

where $x \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ is a point in the n-dimensional Euclidean space E^n,

$F(x) \equiv \begin{bmatrix} f_1(x_1) \\ \vdots \\ f_n(x_n) \end{bmatrix}$ is a nonlinear function mapping E^n into E^n, A is an

$n \times n$ matrix of real numbers, and $B \equiv \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ is an arbitrary point in E^n. We prove (Theorem 1) that there is a unique solution of (1) if:

(*i*) Each f_i is a strictly monotone increasing function mapping E^1 onto E^1,

and

(*ii*) The elements a_{ij} of the matrix A satisfy the inequality

$$a_{ii} \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} |, \quad \text{for} \quad i = 1, \cdots, n.$$

We then demonstrate a straightforward method of computing bounds on the location of this solution. Finally, we present two iterative techniques for computing the solution; and prove (Theorem 3) that the two additional assumptions:

(*iii*) Either all of the functions f_i are convex, or else all f_i are concave,

and

(*iv*) $a_{ij} \leqq 0$ if $i \neq j$,

are sufficient to guarantee that the iterations converge to the solution.

Equations of type (1) occur often in the study of nonlinear electrical networks. For example, if a linear n-port containing resistors, independent sources, and dependent sources has a two-terminal device whose V vs I curve is specified by $I_i = f_i(V_i)$, for $i = 1, \cdots, n$, connected across each port, then the port voltages may often be expressed as the solution of an equation of type (1). In this case the matrix A will be the y-parameter matrix of the n-port, the constant vector B will account for the presence of the independent sources, and the components of the vector x will be the desired port voltages.

II. ACTIVE AND NONRECIPROCAL n-PORTS

In case the n-port of the above example contains no dependent sources and the functions f_i satisfy condition (*i*) above, the existence and uniqueness of a solution of (1) follows immediately from the well-known result of R. J. Duffin.[1] In fact, with the additional assumption that the slope of each f_i is bounded by positive constants the computational technique of J. Katzenelson and L. H. Seitelman may be used to compute the solution.[2] This computational technique is based upon a theorem of I. W. Sandberg which relies upon the contraction-mapping fixed point theorem.[3]

Sandberg's theorem may, in fact, be used to prove the existence and uniqueness of a solution of (1), and to construct a convergent iteration process for computing this solution, whenever the matrix A is positive semidefinite* and the slope of each f_i is bounded by positive constants. Other theorems which do not require that the slopes of each strictly monotone increasing f_i be bounded by positive constants also exist. (For example, see Ref. 4.) These theorems guarantee existence and uniqueness of a solution of (1) whenever A is positive semidefinite but do not specify a procedure for computing it.

Suppose, however, that the matrix A is not positive semidefinite; that is, suppose the n-port in the above example is active. Then the above results no longer apply. It may often happen that the matrix A is not positive semidefinite but still satisfies condition (*ii*) above. The matrix

$$A = \begin{bmatrix} 1 & 1 \\ 5 & 7 \end{bmatrix},$$

for example, has this property. It is interesting to notice that in this case the matrix A will necessarily also be nonsymmetric (the corresponding n-port will be nonreciprocal). This follows from the fact that for symmetric matries A, condition (*ii*) implies that A is a dominant matrix[5] which, in turn, implies that A is positive semidefinite. It is for this class of active nonreciprocal n-ports that our work provides entirely new results. Even for the passive case, however, notice that our computational techniques do not require that the slopes of the functions f_i be bounded. Also, there is reason to believe that for certain problems our iteration schemes may converge more rapidly than the ones based upon the contraction mapping theorem. More is said about this in Section VII.

III. EXISTENCE AND UNIQUENESS

Before proving the existence and uniqueness theorem we first prove a lemma which is used many times in this and the following section.

Lemma 1: Let the $n \times n$ matrix A of real numbers satisfy condition (ii) of Section I. For $j = 1, \cdots, n$ let p_j denote the jth component of $p \in E^n$. Let $k \in \{1, \cdots, n\}$ be chosen such that $|p_k| = \max \{|p_j| : j = 1, \cdots, n\}$. Then,

---

* The $n \times n$ matrix A is said to be positive semidefinite if $\langle x, Ax \rangle \geq 0$ for all x in E^n.

$$p_k > 0 \Rightarrow \sum_{j=1}^{n} a_{kj}p_j \geqq 0,$$

and

$$p_k < 0 \Rightarrow \sum_{j=1}^{n} a_{kj}p_j \leqq 0.$$

Proof:

$$a_{kk} \mid p_k \mid \geqq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid \cdot \mid p_k \mid \geqq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj}p_j \mid \geqq \left| \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}p_j \right|.$$

Thus,

$$a_{kk} \mid p_k \mid \geqq \pm \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}p_j \ .$$

But then,

$$p_k > 0 \Rightarrow a_{kk}p_k \geqq -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}p_j \Rightarrow \sum_{j=1}^{n} a_{kj}p_j \geqq 0,$$

and,

$$p_k < 0 \Rightarrow -a_{kk}p_k \geqq \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}p_j \Rightarrow \sum_{j=1}^{n} a_{kj}p_j \leqq 0. \qquad \square$$

Theorem 1: There exists a unique solution of (1) whenever conditions (i) and (ii) of Section I are satisfied.

Proof: We first prove that if a solution exists it is unique. Let x^1 and x^2 be solutions of (1). Then,

$$F(x^2) - F(x^1) = A(x^1 - x^2).$$

For $j = 1, \cdots, n$ let x_j^1 and x_j^2 denote the jth components of x^1 and x^2, respectively, and choose $k \in \{1, \cdots, n\}$ such that

$$\mid x_k^1 - x_k^2 \mid = \max \{ \mid x_j^1 - x_j^2 \mid : j = 1, \cdots, n \}.$$

If $x_k^1 > x_k^2$ then, by Lemma 1,

$$f_k(x_k^2) - f_k(x_k^1) = \sum_{j=1}^{n} a_{kj}(x_j^1 - x_j^2) \geqq 0.$$

If $x_k^1 < x_k^2$ then, by Lemma 1,

$$f_k(x_k^2) - f_k(x_k^1) = \sum_{j=1}^{n} a_{kj}(x_j^1 - x_j^2) \leqq 0.$$

Both of these conclusions contradict the fact that f_k is strictly monotone increasing. Thus, $x_k^1 = x_k^2$ and hence $x_j^1 = x_j^2$ for $j = 1, \cdots, n$. That is, the solution of (1) is unique, if it exists.

We prove existence of a solution by induction. For $k = 1, \cdots, n$ let

$$F_k(x) \equiv \begin{bmatrix} f_1(x_1) \\ \vdots \\ f_k(x_k) \end{bmatrix}, \qquad A_k \equiv \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \cdots & \cdots & \cdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}, \qquad B_k \equiv \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}.$$

Clearly, the matrix A_k satisfies condition (*ii*) of Section I. Also, it is clear that there exists a unique solution* of $F_1(x) + A_1x = B_1$ for every strictly monotone increasing function f_1 mapping E^1 onto E^1.

Assume that there exists a solution of $F_k(x) + A_kx = B_k$ for arbitrary strictly monotone increasing functions f_i, $i = 1, \cdots, k$ mapping E^1 onto E^1. Then, for every real number x_{k+1}, the equation

$$F_k(x) + A_kx + \begin{bmatrix} a_{1,k+1} \\ \vdots \\ a_{k,k+1} \end{bmatrix} x_{k+1} = B_k$$

has a (unique) solution; since for $i = 1, \cdots, k$ the function $f_i(x_i) + a_{i,k+1}x_{k+1}$ is strictly monotone increasing from E^1 onto E^1. Let the components of this solution be denoted by $x_i = m_i(x_{k+1})$ for $i = 1, \cdots, k$. We have thus defined k functions m_i on E^1.

We now prove that for every $x_{k+1}^1, x_{k+1}^2 \in E^1$,

$$| x_{k+1}^2 - x_{k+1}^1 | \geqq | m_j(x_{k+1}^2) - m_j(x_{k+1}^1) |, \quad \text{for} \quad j = 1, \cdots, k. \tag{2}$$

This inequality, incidentally, implies that each m_i is continuous.

Let $x_{k+1}^1, x_{k+1}^2 \in E^1$ and choose $l \in \{1, \cdots, k\}$ such that

$$| m_l(x_{k+1}^2) - m_l(x_{k+1}^1) | = \max \{ | m_j(x_{k+1}^2) - m_j(x_{k+1}^1) | : j = 1, \cdots, k \}.$$

Assume that $| m_l(x_{k+1}^2) - m_l(x_{k+1}^1) | > | x_{k+1}^2 - x_{k+1}^1 |$. Clearly, then, $m_l(x_{k+1}^2) - m_l(x_{k+1}^1) \neq 0$. If $m_l(x_{k+1}^2) - m_l(x_{k+1}^1) > 0$ then,

$$f_l[m_l(x_{k+1}^2)] - f_l[m_l(x_{k+1}^1)] > 0.$$

---

* We take the liberty of using the same symbol x to denote points in any of the spaces E^k, $1 \leq k \leq n$. No confusion should arise since the subscripts on F and A will make our choice clear.

Also, since the matrix A_{k+1} satisfies condition (*ii*) of Section I, letting

$$p_j = m_j(x^2_{k+1}) - m_j(x^1_{k+1}), \quad \text{for} \quad j = 1, \cdots, k,$$

$$p_{k+1} = x^2_{k+1} - x^1_{k+1},$$

we have, by Lemma 1,

$$\sum_{j=1}^{k} a_{lj}[m_j(x^2_{k+1}) - m_j(x^1_{k+1})] + a_{l,k+1}(x^2_{k+1} - x^1_{k+1}) \geqq 0.$$

Thus,

$$f_l[m_l(x^2_{k+1})] + \sum_{j=1}^{k} a_{lj}m_j(x^2_{k+1}) + a_{l,k+1}x^2_{k+1}$$

$$> f_l[m_l(x^1_{k+1})] + \sum_{j=1}^{k} a_{lj}m_j(x^1_{k+1}) + a_{l,k+1}x^1_{k+1}, \qquad (3)$$

which is a contradiction since the quantity on each side of this inequality is equal to b_l. If $m_l(x^2_{k+1}) - m_l(x^1_{k+1}) < 0$ then,

$$f_l[m_l(x^2_{k+1})] - f_l[m_l(x^1_{k+1})] < 0.$$

By applying Lemma 1 again, as above, one arrives again at (3) with $>$ replaced by $<$. This is also a contradiction. Thus, we must have

$$| x^2_{k+1} - x^1_{k+1} | \geqq | m_l(x^2_{k+1}) - m_l(x^1_{k+1}) |,$$

and hence (2) is proved

Now, consider the function

$$\sum_{j=1}^{k} a_{k+1,j}m_j(x_{k+1}) + a_{k+1,k+1}x_{k+1}. \qquad (4)$$

Let $x^1_{k+1}, x^2_{k+1} \in E^1$ with $x^1_{k+1} < x^2_{k+1}$. Then,

$$a_{k+1,k+1} \geqq \sum_{j=1}^{k} | a_{k+1,j} |$$

implies

$$a_{k+1,k+1}(x^2_{k+1} - x^1_{k+1})$$

$$= a_{k+1,k+1} | x^2_{k+1} - x^1_{k+1} | \geqq \sum_{j=1}^{k} (| a_{k+1,j} | \cdot | x^2_{k+1} - x^1_{k+1} |).$$

But, using (2),

$$a_{k+1,k+1}(x^2_{k+1} - x^1_{k+1}) \geqq \sum_{j=1}^{k} | a_{k+1,j}[m_j(x^2_{k+1}) - m_j(x^1_{k+1})] |$$

$$\geqq -\sum_{j=1}^{k} a_{k+1,j}[m_j(x_{k+1}^2) - m_j(x_{k+1}^1)],$$

which implies

$$\sum_{j=1}^{k} a_{k+1,j}m_j(x_{k+1}^1) + a_{k+1,k+1}x_{k+1}^1 \leqq \sum_{j=1}^{k} a_{k+1,j}m_j(x_{k+1}^2) + a_{k+1,k+1}x_{k+1}^2 .$$

That is, the function (4) is monotone increasing. Clearly (4) is continuous. It follows, therefore, that if f_{k+1} is a strictly monotone increasing function mapping E^1 onto E^1, then so is the function

$$f_{k+1}(x_{k+1}) + \sum_{j=1}^{k} a_{k+1,j}m_j(x_{k+1}) + a_{k+1,k+1}x_{k+1} .$$

Thus, there exists a unique solution of the equation

$$f_{k+1}(x_{k+1}) + \sum_{j=1}^{k} a_{k+1,j}m_j(x_{k+1}) + a_{k+1,k+1}x_{k+1} = b_{k+1} .$$

If x_{k+1}^0 denotes this solution then

$$x^0 \equiv \begin{bmatrix} m_1(x_{k+1}^0) \\ \vdots \\ m_k(x_{k+1}^0) \\ x_{k+1}^0 \end{bmatrix}$$

is the (unique) solution of

$$F_{k+1}(x) + A_{k+1}x = B_{k+1} .$$

Thus, we have proved that there exists a unique solution of (1). □

IV. BOUNDS ON THE SOLUTION

Having established the existence and uniqueness of a solution of (1) a natural question to arise is: What can one say about the location of this solution? It turns out that we can say quite a bit (again assuming that conditions (*i*) and (*ii*) of Section I are satisfied). One can, in fact, with little effort (compared with the effort required, in general, to actually compute the solution) determine a finite region R in E^n, in which the solution must lie. This region is the cartesian product of finite intervals $I_i \subset E^1$, for $i = 1, \cdots, n$, each of which has the property that if

$$x^0 \equiv \begin{bmatrix} x_1^0 \\ \vdots \\ x_n^0 \end{bmatrix}$$

is the solution of (1) then $x_i^0 \in I_i$ and, as

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \to 0,$$

the length of I_i, $l(I_i) \to 0$. Thus, when the off-diagonal elements of A are small, the region R will also be small.

In many applications it may be sufficient to know only that there exists a unique solution of (1) and to know the region R in which it must lie. If, however, one actually does want to compute the solution by some iterative technique, the knowledge of R should be useful in determining a starting point for the iteration. In fact, it will be shown that if the point x^* is the solution of

$$F(x) + \text{diag } [a_{11}, \cdots, a_{nn}]\, x = B, \tag{5}$$

then x^* is also in R and thus might be a reasonable starting point for an iterative computation of x^0.

The computation of bounds for the solution of (1) proceeds in two steps. First, one solves each of the equations

$$f_i(x_i) = b_i, \quad \text{for} \quad i = 1, \cdots, n. \tag{6}$$

Letting α_i denote the solutions of (6), and defining

$$\alpha = \max \{ | \alpha_i | : i = 1, \cdots, n \},$$

$$B' = \begin{bmatrix} \sum_{\substack{j=1 \\ j \neq 1}}^{n} | a_{1j} | \\ \vdots \\ \sum_{\substack{j=1 \\ j \neq n}}^{n} | a_{nj} | \end{bmatrix},$$

one then solves each of the equations

$$F(x) + \text{diag } [a_{11}, \cdots, a_{nn}]\, x = B - \alpha B', \tag{7a}$$

$$F(x) + \text{diag } [a_{11}, \cdots, a_{nn}]\, x = B + \alpha B'. \tag{7b}$$

Denoting the solutions of (7a) and (7b) by

$$\eta \equiv \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \quad \text{and} \quad \xi \equiv \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix},$$

respectively, one has $R = I_1 \times \cdots \times I_n$, where

$$I_i = [\eta_i , \xi_i], \quad \text{for} \quad i = 1, \cdots , n.$$

It is clear from the fact that each component of the vector $\alpha B'$ is a nonnegative number and from the monotone nature of the left-hand sides of (7) that x^*, the solution of (5), is (as claimed) always in R. It is also clear that, for $i = 1, \cdots , n$, as

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \to 0,$$

the ith components of both $B - \alpha B'$ and $B + \alpha B'$ approach b_i, and hence $\eta_i \to x_i^*$ and $\xi_i \to x_i^*$. Thus, $l(I_i) \to 0$. We now prove that the solution of (1) is in R.

Theorem 2: If R is constructed as described above, then the solution of (1) is contained in R whenever conditions (i) and (ii) of Section I are satisfied.

Proof: Let x^0 be the solution of (1) and let $k \in \{1, \cdots , n\}$ be chosen such that $| x_k^0 | = \max\{| x_i^0 | : i = 1, \cdots , n\}$. Then, by Lemma 1, if $x_k^0 > 0$, $\sum_{i=1}^{n} a_{ki} x_i^0 \geqq 0$ and hence,

$$0 = f_k(x_k^0) + \sum_{i=1}^{n} a_{ki} x_i^0 - b_k \geqq f_k(x_k^0) - b_k$$

or $f_k(x_k^0) \leqq b_k$.

Thus, because of the monotonicity of f_k,

$$| x_k^0 | = x_k^0 \leqq \alpha_k \leqq \alpha,$$

and hence $| x_i^0 | \leqq \alpha$ for $i = 1, \cdots , n$. Similarly, by Lemma 1, if $x_k^0 < 0$, $\sum_{i=1}^{n} a_{ki} x_i^0 \leqq 0$ and hence,

$$f_k(x_k^0) \geqq b_k ,$$

and thus

$$| x_k^0 | = -x_k^0 \leqq -\alpha_k \leqq \alpha,$$

and hence $| x_i^0 | \leqq \alpha$, for $i = 1, \cdots, n$. Thus, in any case, $| x_i^0 | \leqq \alpha$, for $i = 1, \cdots, n$.

Now, for all x with $| x_j | \leqq \alpha$ for $j = 1, \cdots, n$, and for each $i \in \{1, \cdots, n\}$ we have,

$$\alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | = \sum_{\substack{j=1 \\ j \neq i}}^{n} (| a_{ij} | \alpha) \geqq \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} x_j |,$$

which implies

$$\alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \geqq \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j ,$$

and

$$-\alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \leqq \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j .$$

Thus,

$$a_{ii} x_i - \alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \leqq \sum_{j=1}^{n} a_{ij} x_j \leqq a_{ii} x_i + \alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} |.$$

In particular, for $x = x^0$, we have

$$f_i(x_i^0) + a_{ii} x_i^0 - \alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \leqq b_i \leqq f_i(x_i^0) + a_{ii} x_i^0 + \alpha \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} |.$$

Comparing this result with (7) we have, as a consequence of the monotonicity of the functions on the left-hand sides of (7),

$$\eta_i \leqq x_i^0 \leqq \xi_i , \quad \text{for} \quad i = 1, \cdots, n.$$

Hence, $x^0 \in R$. □

Since in the above proof it was shown that $| x_i^0 | \leqq \alpha$ for $i = 1, \cdots, n$ it might seem to some readers that the intervals I_i might be reduced in length if we simply define them to be: $I_i = [-\alpha, \alpha] \cap [\eta_i, \xi_i]$. This, however, is unnecessary since it is easily shown that $-\alpha \leqq \eta_i \leqq \xi_i \leqq \alpha$, for $i = 1, \cdots, n$.

V. EXAMPLE

We now give an example of the use of the above method for the computation of solution bounds. Consider the equation

$$\begin{bmatrix} f_1(x_1) \\ f_2(x_2) \end{bmatrix} + \begin{bmatrix} 5 & 4 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 15 \\ 13 \end{bmatrix},$$

where f_1 and f_2 are defined by

$$f_1(x_1) = \begin{cases} 4^{x_1} - 2, & x_1 \geqq 0 \\ \frac{1}{10}x_1 - 1, & x_1 < 0, \end{cases}$$

and

$$f_2(x_2) = \begin{cases} x_2 + 9, & x_2 \geqq 3 \\ 4x_2, & -3 < x_2 < 3 \\ x_2 - 9, & x_2 \leqq -3. \end{cases}$$

Figure 1 shows the graphs of f_1 and f_2. Since we know that the region R will be small if the off-diagonal terms of A are small enough, we have intentionally chosen an example in which these terms are rather large.

The computation of α by solving (6) may be done by inspection for this example. One finds that $4^{\alpha_1} = 17$ implies that α_1 is slightly greater than 2, and since $\alpha_2 = 4$ we have $\alpha = 4$. Using this result in (7) one readily computes

$$\eta = \begin{bmatrix} 0 \\ 0.125 \end{bmatrix}, \qquad \xi \approx \begin{bmatrix} 2.23 \\ 3.2 \end{bmatrix}.$$

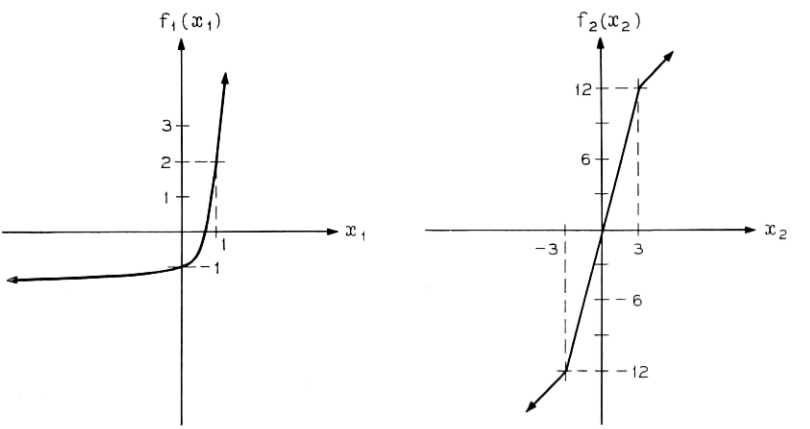

Fig. 1 — The nonlinear functions f_1 and f_2 for the example.

Actually, it is easily verified that the solution of this example is $x^0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

VI. COMPUTATION OF THE SOLUTION

For $i = 1, \cdots, n$ we denote by $f_i'(x_i)$ the right-hand derivative of f_i at the point $x_i \in E^1$. For each $x \in E^n$ we denote by $F'(x)$ the following matrix:

$$F'(x) = \text{diag } [f_1'(x_1), \cdots, f_n'(x_n)].$$

It is easy to prove that if F satisfies condition (*i*) of Section I then $F'(x)$ exists for all $x \in E^n$. Also, it is clear that each element of the main diagonal of $F'(x)$ is nonnegative for all $x \in E^n$. (Each element is in fact positive if, in addition, F satisfies condition (*iii*) of Section I.) Finally, we note that

$$F^{-1}(y) \equiv \begin{bmatrix} f_1^{-1}(y_1) \\ \vdots \\ f_n^{-1}(y_n) \end{bmatrix}$$

is defined for all $y \in E^n$, assuming again that F satisfies condition (*i*) of Section *I*.

The following two iteration schemes are proposed for the computation of the solution of (1):

Scheme 1: For given $x^1 \in E^n$ the sequence $x^1, x^2, x^3, \cdots$ of points in E^n is constructed by use of the formula

$$x^{k+1} = [F'(x^k) + A]^{-1}(B - F(x^k) + F'(x^k)x^k). \tag{8}$$

Scheme 2: For given $x^1 \in E^n$ the sequence $x^1, x^2, x^3, \cdots$ of points in E^n is constructed by use of the formula

$$x^{k+1} = [F'(F^{-1}(y^k)) + A]^{-1}(B - y^k + F'(F^{-1}(y^k))F^{-1}(y^k)), \tag{9}$$

where $y^k = -Ax^k + B$.

In order to explain the origin of (8) and (9) we make the following observations: If for $i = 1, \cdots, n$ (x_i^k, y_i^k) is a given point in E^2, and if we draw the graph of each of the functions f_i, then each of the points in the sets $\{(x_i^k, f_i(x_i^k)): i = 1, \cdots, n\}$ and $\{(f_i^{-1}(y_i^k), y_i^k): i = 1, \cdots, n\}$ lies on the graph of the corresponding function f_i. Suppose we now replace (approximate) each f_i by the straight line which is tangent to it at the corresponding point in one of the above sets.* Choosing the

* Our definition of *tangent* coincides with the usual one, except that the right-hand derivative is used at those points where the derivative fails to exist.

first set of points we approximate F by

$$\hat{F}(x) \equiv F'(x^k)x + F(x^k) - F'(x^k)x^k.$$

Choosing the second set gives

$$\tilde{F}(x) \equiv F'(F^{-1}(y^k))x + y^k - F'(F^{-1}(y^k))F^{-1}(y^k).$$

If we now define $y^k = -Ax^k + B$ and compute the solution of the equation

$$\hat{F}(x) + Ax = B$$

and call it x^{k+1}, we obtain (8). Calling x^{k+1} the solution of

$$\tilde{F}(x) + Ax = B,$$

yields (9).

The above remarks have a very meaningful interpretation for problems arising from nonlinear electrical networks of the type described in Section I. Iteration Scheme 1 implements the following procedure: Given the vector x^k of port voltages for the linear n-port, replace each two-terminal nonlinear device with a linear Thévenin's "equivalent" circuit whose V vs I curve is a straight line, tangent to the given curve at the point $(x_i^k, f_i(x_i^k))$. Compute the port voltages in the resulting linear network to obtain x^{k+1}.

Iteration Scheme 2 has a similar interpretation; this time, however, the vector of port currents, $y^k = -Ax^k + B$, is used to determine the linear equivalent circuit replacing the nonlinear devices at each step.

In view of the above remarks it is apparent that if one has some facility for solving linear network problems (a computer program, for example) then it might easily be adapted to solve many nonlinear problems as well.

We finally remark that the use of the first iteration scheme is, in essence, the same as using the Newton-Raphson technique to compute the root of (1).

We now prove a theorem which specifies conditions which are sufficient to ensure convergence of each of the above iteration schemes. We emphasize, however, that these iteration schemes will converge for many problems in which the conditions of the theorem are not satisfied—especially if a good enough starting point is provided.

In the following we denote the origin in E^n by θ and, for the points $x, y \in E^n$, the notation $x \leqq y$ means $x_i \leqq y_i$ for $i = 1, \cdots, n$. The

relations $x < y$, $x \geqq y$, $x > y$ are defined similarly. We also make use of the concept of a *matrix of monotone kind.*[6] The matrix A is said to be of monotone kind if $x \in E^n$, $Ax \geqq \theta \Rightarrow x \geqq \theta$. It is easy to show that A is of monotone kind if and only if A^{-1} contains only nonnegative elements. It is also easy to show that if A is of monotone kind and $x, y \in E^n$ with $Ax \leqq y$, then $x \leqq A^{-1}y$. Ref. 6 shows that if the strict inequality $>$ holds in condition (*ii*) of Section I, then conditions (*ii*) and (*iv*) are sufficient to ensure that A is of monotone kind.

Theorem 3: For an arbitrary starting point x^1, both of the above iteration schemes will converge to the solution of (1) if conditions (i) through (iv) of Section I are satisfied.

Proof: We give here only the proof for the second iteration scheme, assuming that all of the functions f_i are convex. The other three cases are quite similar and it will be apparent to the reader how this proof may easily be modified to take care of them.*

We first remark that the iteration scheme is well defined. The fact that for every $y^k \in E^n$, $F'(F^{-1}(y^k))$ is a diagonal matrix containing all positive numbers on the main diagonal, and the fact that A satisfies conditions (*ii*) and (*iv*) of Section I, assures us that the matrix $[F'(F^{-1}(y^k)) + A]$ is nonsingular (it is, in fact, of monotone kind—see Ref. 6, p. 376).

Let x^1 be an arbitrary point in E^n. Then, since for $i = 1, \cdots, n$ and $k = 2, 3, 4, \cdots$ each of the points (x_i^k, y_i^k) lies on some straight line, tangent to the corresponding function f_i, and since each f_i is strictly monotone increasing and convex, we have that $F^{-1}(y^k) \leqq x^k$ for $k = 2, 3, 4, \cdots$. We now show that $F^{-1}(y^k) \leqq x^k$ implies that $x^{k+1} \leqq x^k$. Obviously,

$$F'(F^{-1}(y^k))(x^k - F^{-1}(y^k)) \geqq \theta.$$

But, by definition, $Ax^k + y^k - B = \theta$; hence,

$$F'(F^{-1}(y^k))(x^k - F^{-1}(y^k)) + Ax^k + y^k - B \geqq \theta,$$

which implies

$$[F'(F^{-1}(y^k)) + A]x^k \geqq B - y^k + F'(F^{-1}(y^k))F^{-1}(y^k).$$

But then, since $[F'(F^{-1}(y^k)) + A]$ is a matrix of monotone kind,

* After this manuscript had been completed, the author became aware of J. S. Vandergraft's paper (Ref. 7). With a certain amount of reformulation, the (monotone) convergence of the first iteration scheme, when all f_i are convex, can be shown to follow, in essence, from his Theorem 5.1.

$$x^k \geqq [F'(F^{-1}(y^k)) + A]^{-1}(B - y^k + F'(F^{-1}(y^k))F^{-1}(y^k)),$$

or, $x^k \geqq x^{k+1}$. Thus, the sequence $x^2, x^3, x^4, \cdots$ has the property

$$x^2 \geqq x^3 \geqq x^4 \geqq \cdots .$$

We now show that for $k = 2, 3, 4, \cdots$, $x^k \geqq x^0$, where x^0 is the solution of (1). For each x^k, $k = 2, 3, 4, \cdots$, there is some point $p \in E^n$ $(p \equiv F^{-1}(y^{k-1}))$ such that

$$Ax^k - B = F'(p)p - F'(p)x^k - F(p). \tag{10}$$

Furthermore, from the convexity of each f_i, it is clear that for every pair of points $q^1, q^2 \in E^n$,

$$F(q^1) \geqq F(q^2) + F'(q^2)(q^1 - q^2).$$

In particular,

$$F(x^0) \geqq F(p) + F'(p)(x^0 - p).$$

Hence,

$$F'(p)p - F(p) + F(x^0) \geqq F'(p)x^0$$

which implies

$$F'(p)(p - x^k) - F(p) + F(x^0) \geqq F'(p)(x^0 - x^k).$$

Using (10) we have, therefore,

$$Ax^k - B + F(x^0) \geqq F'(p)(x^0 - x^k).$$

But, $F(x^0) = -Ax^0 + B$, hence

$$A(x^k - x^0) \geqq F'(p)(x^0 - x^k)$$

or,

$$[F'(p) + A](x^k - x^0) \geqq \theta.$$

But then, since $[F'(p) + A]$ is of monotone kind, $x^k - x^0 \geqq \theta$, or $x^k \geqq x^0$. Thus, we have shown that each sequence $x_i^2, x_i^3, x_i^4, \cdots$ is a bounded monotone sequence and hence the sequence $x^2, x^3, x^4, \cdots$ converges to some point x^* in E^n. We now prove that $x^* = x^0$; that is, we show that x^* satisfies (1).

Let $y^* = -Ax^* + B$. Then, as $k \to \infty$, $x^k \to x^*$ and $y^k \to y^*$. Thus, $F^{-1}(y^k) \to F^{-1}(y^*)$ and each element of the matrix $F'(F^{-1}(y^k))$ approaches the corresponding element of $F'(F^{-1}(y^*))$. Now, from (9), we have

$$Ax^{k+1} + F'(F^{-1}(y^k))x^{k+1} = Ax^k + F'(F^{-1}(y^k))F^{-1}(y^k)$$

which implies

$$F'(F^{-1}(y^k))(F^{-1}(y^k) - x^{k+1}) = A(x^{k+1} - x^k)$$

and hence

$$F'(F^{-1}(y^k))(F^{-1}(y^k) - x^* = A(x^{k+1} - x^k) - F'(F^{-1}(y^k))(x^* - x^{k+1}).$$

But as $k \to \infty$, $(x^{k+1} - x^k) \to \theta$ and hence $A(x^{k+1} - x^k) \to \theta$; also, $(x^* - x^{k+1}) \to \theta$ and hence $F'(F^{-1}(y^k))(x^* - x^{k+1}) \to F'(F^{-1}(y^*))\theta = \theta$. Thus, as $k \to \infty$,

$$F'(F^{-1}(y^k))(F^{-1}(y^k) - x^*) \to \theta$$

which implies

$$F^{-1}(y^k) - x^* \to \theta$$

or

$$F^{-1}(y^k) \to x^*$$

and therefore

$$y^k \to F(x^*).$$

Hence, $y^* = F(x^*)$, and thus,

$$F(x^*) + Ax^* = B.$$

Thus, the iteration converges to the solution of (1). □

Although Theorem 3 states that both of our iteration schemes will converge for the same class of problems, only one of the schemes might converge for some problems for which all of the conditions (*i*) through (*iv*) of Section I are not satisfied. Also, for some problems a prior knowledge of the region in which the solution lies might dictate the choice of one iteration scheme over the other. For example, if it is known that some of the functions f_i are quite steep in the neighborhood of the solution then perhaps F^{-1} may be evaluated in this region more accurately than F. In this case Scheme 2 might be preferred to Scheme 1.

VII. SPEED OF CONVERGENCE

Section II mentions that in certain situations our iteration schemes may converge to the solution of (1) more rapidly than those based upon the contraction-mapping fixed point theorem. To illustrate this property we have chosen to compare the rate of convergence of Sandberg's iteration scheme to that of our schemes.[3]

If we define the operator G mapping E^n into E^n by

$$G(x) \equiv F(x) + Ax,$$

then, as a special case of Sandberg's Theorem I, we have the result: If there are positive constants k_1 and k_2 such that

$$\langle G(x) - G(y), x - y \rangle \geqq k_1 \| x - y \|^2, \tag{11}$$

and

$$\| G(x) - G(y) \|^2 \leqq k_2 \| x - y \|^2, \tag{12}$$

for all $x, y \in E^n$, then there is a unique solution of (1) and the solution is given by $\lim_{k \to \infty} x^k$, where x^1 is an arbitrary point in E^n, and

$$x^{k+1} = \frac{k_1}{k_2} [B - G(x^k)] + x^k,$$

for $k = 1, 2, 3, \cdots$. The proof of this theorem consists of showing that the mapping

$$H(x) \equiv \frac{k_1}{k_2} [B - G(x)] + x$$

is a contraction.

It is interesting to observe that if the inequalities (11) and (12) are satisfied then positive constants k_3 and k_4 exist, such that

$$\langle G(x) - G(y), x - y \rangle \leqq k_3 \| x - y \|^2, \tag{13}$$

and

$$\| G(x) - G(y) \|^2 \geqq k_4 \| x - y \|^2, \tag{14}$$

for all $x, y \in E^n$. In fact, a simple application of the Schwarz inequality to (11) and (12) yields (13) and (14) with $k_3 = (k_2)^{\frac{1}{2}}$ and $k_4 = k_1^2$. Now (13) and (14) are of the same form as (11) and (12), except that the inequalities are reversed. Thus, if one uses (13) and (14) in the proof of Sandberg's theorem, reversing all inequalities, one obtains:

$$\| H(x) - H(y) \|^2 \geqq K \| x - y \|^2,$$

where,

$$K = 1 - 2(k_1^2/k_2)^{\frac{1}{2}} + (k_1^2/k_2)^2.$$

It is readily seen that if $4k_1^2 < k_2$, then K is positive. If we let x^0 denote the solution of (1), and hence $H(x^0) = x^0$, we have, for $k = 1, 2, \cdots$,

$$\| x^{k+1} - x^0 \|^2 = \| H(x^k) - H(x^0) \|^2 \geqq K \| x^k - x^0 \|^2.$$

Thus, $(K)^{\frac{1}{2}}$ represents, in this case, a lower bound on the rate of convergence of the iteration scheme. It is true that $(K)^{\frac{1}{2}}$ is always in the interval (0,1), for indeed Sandberg has proved that the sequence x^k does converge to x^0. However, as k_1 becomes small, and as k_2 becomes large, K approaches 1 and the sequence converges quite slowly. For (1) the largest value that may be used for k_1 and the smallest value that may be used for k_2 will many times be dictated by the positive constants which are bounds on the slopes of the functions f_i. If, for example, the slopes of the f_i become so large for large x_i, and so small for large negative x_i that one must choose $k_1 = 10^{-1}$ and $k_2 = 10^2$, then one easily computes $(K)^{\frac{1}{2}} \approx 0.99$. Thus, no matter how close any iterate is to the solution, the next iterate will be no more than about one percent closer.

It is of course true that Sandberg's iteration scheme is applicable to a much more general class of problems than we consider in this paper. If, however, for any problem to which it is applied, the constants k_1 and k_2 must be restricted such that k_1/k_2 is quite small, then the rate of convergence will always be adversely affected. In the Katzenelson-Seitelman application of Sandberg's iteration scheme, their "heuristic refinement" (see Ref. 2) attempts to overcome this difficulty.

Although the classes of equations to which our iteration schemes and the Katzenelson-Seitelman algorithm may be applied are not identical, in those cases where both techniques may be used the advantage that our schemes offer is now clear. From (8) and (9) one easily obtains

$$x^{k+1} - x^0 = [F'(x^k) + A]^{-1}(F(x^0) - F(x^k) - F'(x^k)(x^0 - x^k)),$$

and

$$x^{k+1} - x^0 = [F'(F^{-1}(y^k)) + A]^{-1}\,(F(x^0) - y^k - F'(F^{-1}(y^k))(x^0 - F^{-1}(y^k))),$$

respectively. These equations show that $\| x^{k+1} - x^0 \|$ will be small (even if $\| x^k - x^0 \|$ is rather large) so long as for $i = 1, \cdots, n$,

$$\frac{f_i(x_i^0) - f_i(x_i^k)}{x_i^0 - x_i^k} \approx f_i'(x_i^k),$$

for Scheme 1, or

$$\frac{f_i(x_i^0) - y_i^k}{x_i^0 - f_i^{-1}(y_i^k)} \approx f_i'(f_i^{-1}(y_i^k)),$$

for Scheme 2. That is, as soon as the kth iterate comes close enough to the solution that each of the functions f_i is approximately linear, the rate of convergence of our iterations becomes quite rapid. In fact, the rate of convergence increases without bound as the iterates approach the solution. It is also clear that if each of the functions f_i is piecewise linear then our iterations will converge in a finite number of steps.

From the standpoint of computational efficiency it is, of course, the amount of time required to compute an approximate solution that is the major concern. For those problems to which both our iteration schemes and the Katzenelson-Seitelman algorithm may be applied, it can happen that our methods might still be slower than theirs even in the case when the convergence rate of our methods is faster. This can happen because, for some problems, the equation with which we are concerned may be of a higher order than theirs, and also because we must compute the inverse of a matrix at each iteration step. On the other hand, it is clear that for many problems, even from the standpoint of total computation time, our techniques will be more efficient.

VIII. ACKNOWLEDGMENT

The author is grateful to I. W. Sandberg for encouragement and many helpful conversations.

REFERENCES

1. Duffin, R. J., "Nonlinear Networks II," Bull. Amer. Math. Soc., *53* (October 1947), pp. 963–971.
2. Katzenelson, J. and Seitelman, L. H., "An Iterative Method for Solution of Networks of Nonlinear Monotone Resistors," IEEE Trans. Circuit Theory, *CT-13*, No. 3 (September 1966), pp. 317–323.
3. Sandberg, I. W., "On the Properties of Some Systems That Distort Signals-I," B.S.T.J., *42*, No. 5 (September 1963), pp. 2033–2046.
4. Minty, G. J., "Two Theorems on Nonlinear Functional Equations in Hilbert Space," Bull. Amer. Math. Soc., *69*, No. 5 (September 1963), pp. 691–692.
5. Weinberg, L., *Network Analysis and Synthesis*, New York: McGraw-Hill, 1962.
6. Collatz, L., *Functional Analysis and Numerical Mathematics* (tr. from German), New York: Academic Press, 1966.
7. Vandergraft, J. S., "Newton's Method for Convex Operators in Partially Ordered Spaces," SIAM J. Numerical Anal., *4*, No. 3 (September 1967), pp. 406–432.