AN ITERATIVE PROCEDURE FOR SOLVING NONLINEAR EQUATIONS

by

Y-F Lam and J. D. McPherson

Memorandum No. ERL-M516

16 April 1975

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

AN ITERATIVE PROCEDURE FOR SOLVING NONLINEAR EQUATIONS

Y-F Lam

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

J. D. McPherson

Department of Electrical Science
University of Wisconsin - Milwaukee
Milwaukee, Wisconsin 53201

ABSTRACT

In this paper, we consider the problem of solving a system of nonlinear equations $f(s) = b$, where b is a known vector in R^m (the Euclidean m-dimensional space), $f: R^n \to R^m$ with m and n not necessarily equal. An iterative procedure with a great deal of flexibility is proposed. Various aspects of the proposed iteration equation are examined. In particular, we discuss the relationships between the converged value of the iterated sequence and a weighted least squares solution. We also show that the flexibility of the proposed iteration scheme may be employed to improve the condition of ill-conditioned matrices, to distribute weights (of importance) among equations and to study the error due to finite precision arithmethic. Finally, we show that under normal computation noise, the proposed iteration scheme is the best first order estimator of the solution based on the previously iterated point.

Research sponsored in part by National Science Foundation Grant GK-32236X1.

I. INTRODUCTION

In this paper we consider the problem of solving a system of nonlinear equations of the form

$$f(x) \triangleq \begin{bmatrix} f_1(x_1,x_2,\ldots,x_n) \\ f_2(x_1,x_2,\ldots,x_n) \\ \vdots \\ f_m(x_1,x_2,\ldots,x_n) \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \triangleq b \qquad (1)$$

where $x = [x_1,x_2,\ldots x_n]^T$ is an unknown column vector in R^n (Euclidean n-dimensional space), $b = [b_1,b_2,\ldots,b_m]^T$ is a known column vector in R^m and $f(\cdot)$ is a C^1 mapping from R^n to R^m. We say that $f: U \subset R^n \to R^m$ is C^k on U if

$$\frac{\partial^k f_i(x)}{\partial x_{j_1}\ \partial x_{j_2}\ \ldots\ \partial x_{j_k}}$$

is a continuous function of x in U for $i = 1,2,\ldots,m$ and $j_1,j_2,\ldots,j_k = 1,2,\ldots,n$.

It is well known that closed form solutions of nonlinear equations are not, in general, possible. In the special case when $n = m$ in (1), Newton's method and a wide variety of quasi-Newton methods [1-3] may be used. In this paper we will consider the general case where m and n in (1) are not required to be equal.

To find a solution of (1) we propose to utilize the iteration equation

$$x^{k+1} = x^k - [J^T(x^k)\ R\ J(x^k)]^{\dagger}\ J^T(x^k)\ R\ [f(x^k) - b] \qquad (2)$$

where $x^k = [x_1^k,x_2^k,\ldots,x_n^k]^T$ is the previously iterated point with x^0 being the initial approximation to the solution of (1), $J(x)$ is the Jacobian matrix of $f(\cdot)$ evaluated at the point x with the ijth element of J given by $\frac{\partial f_i(x)}{\partial x_j}$, R is a positive definite and symmetric matrix of order m, lending flexibility to the proposed iteration scheme of (2), and A^T and $A^{\dagger}$ respectively denote the transpose and the Moore-Penrose generalized inverse [4,5] of the

matrix A. Note that if A is non-singular, the inverse of A, denoted by A^{-1}, exists and under this condition $A^{\dagger} = A^{-1}$.

Various aspects of the iteration equation (2) are discussed in Section II. In particular, we show that the converged value of (2) minimizes a weighted least squares error function defined for (1). Furthermore, we show that the iteration equation (2) is strictly a downhill method. In Section II, we explore the results obtained by various judicious choices of the matrix R in (2). We will show that R may be used to better the condition of an ill-conditioned matrix and to assign weights to individual equations in a system of equations such that those equations which are considered more accurate or more important are weighted more heavily than the remaining equations in the system. We also show that under the usual computation noises, i.e. round off or truncation errors due to finite precision in the computation, (2), with a particular choice of R, is the best first order estimator of a solution of (1) based on a guessed point generated by whatever means possible, including by (2) itself.

Throughout this paper, we let $||A||$ denote the norm of the matrix A, i.e., the square root of the spectral radius of $A^T A$ [6]. For a vector, y, we shall let $||y||$ denote the Euclidean norm of y [6]. Finally, we always let I denote an identity matrix whose order is inferred from the context.

II. THE ITERATION EQUATION

In this section we discuss various properties of the iteration equation (2). We begin by studying the matrix product $J^T(x)\ R\ J(x)$ which is central to the iteration scheme. Next, we show that the converged value of (2) minimizes a weighted least squares error function defined for (1). Detailed computational procedures, resulting in two algorithms, are given in the following subsection. Section II is concluded with the introduction of a modified iteration equation

which requires fewer calculations per iteration than that of (2).

II.1 Properties of $J^T(x)\ R\ J(x)$

Observe that (2) involves a generalized inverse of the matrix $G(x) \triangleq J^T(x)\ R\ J(x)$. Note that G(x) is symmetric and positive semidefinite (possibly positive definite) since R is an arbitrary positive definite symmetric matrix. We will, in this subsection, quote some properties of positive semidefinite matrices which will be useful in our subsequent discussions and provide three methods to compute the generalized inverse of G(x).

II.1.1 Properties of Semidefinite Matrices

We give several results on positive semidefinite matrices which will be useful in the sequel.

Lemma 1 [7]

Let A be an n x n, symmetric, positive semidefinite matrix of rank r. Then there exists an orthogonal matrix Q such that

$$QA\ Q^T = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}$$

where A_1 is an r x r diagonal matrix with rank r.

Lemma 2 [7]

Let A be a symmetric positive semidefinite matrix of order n, C be an r x n matrix and

$$\hat{A} \triangleq C^T C + A. \tag{3}$$

Let $B = A^{\dagger}$. Then B is a positive semidefinite matrix of order n. In addition, if we define

$$\hat{B} = B - B\ C^T[I + C\ B\ C^T]^{-1}\ C\ B, \tag{4}$$

then $\hat{B} = \hat{A}^{\dagger}$ if and only if

$$N(A) \subseteq N(C) \tag{5}$$

where N(Z) denotes the null space of the matrix Z. Furthermore, if (5) holds, then

(i) $\hat{B}\,\hat{A} = BA$

(ii) $\hat{A}\,\hat{B} = AB$

(iii) $\hat{B}$ is positive semidefinite, and

(iv) $\hat{B} = \hat{A}^{\ddagger}$ if and only if $B = A^{\ddagger}$.

Lemma 3 [8]

Let M be a positive semidefinite symmetric matrix. Then M may be partitioned as

$$M = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$$

where A and B are positive semidefinite symmetric matrices. Moreover, the generalized inverse of M is given by

$$M^{\ddagger} = \begin{bmatrix} A^{\ddagger} + A^{\ddagger} C Q^{\ddagger} C^T A^{\ddagger} & -A^{\ddagger} C Q^{\ddagger} \\ -Q^{\ddagger} C^T A^{\ddagger} & Q^{\ddagger} \end{bmatrix}$$

where $Q = B - C^T A^{\ddagger} C$. Let M be of rank r_M, A be of rank r_A and let B be of order $r_M - r_A$. Then Q is nonsingular. Hence $Q^{\ddagger} = Q^{-1}$.

Lemma 4 [9]

Let M be a matrix of rank r which may be written in the form

$$M = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$$

where A is an r x r matrix with rank r, and A and B are symmetric. Then there exists a matrix P such that

$$M = \begin{bmatrix} A & A P^T \\ P A & P A P^T \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} A \begin{bmatrix} I & P^T \end{bmatrix}$$

where P is a matrix relating the dependence of the dependent rows to the independent rows of M. In addition,

$$M^{\ddagger} = \begin{bmatrix} I \\ P \end{bmatrix} [(I + P^T P) A (I + P^T P)]^{-1} \begin{bmatrix} I & P^T \end{bmatrix}.$$

Lemma 5 [10]

Let A be an n x n matrix. If A = BC where B and C are both matrices of rank r, then

$$A^{\ddagger} = C^T [C C^T]^{-1} [B^T B]^{-1} B^T = C^{\ddagger} B^{\ddagger}.$$

II. 1.2 Computation of the Generalized Inverse

Here we describe three methods to compute $G^{\ddagger}(x)$, the generalized inverse of G(x). Recall that G(x), defined by $G(x) \triangleq J^T(x) R J(x)$, is a positive semidefinite and symmetric matrix.

II. 1.2.1 First Method

By Lemma 1, there exists, for each $x \in R^n$, an n x n orthogonal matrix Q(x), such that

$$G(x) = Q(x) \begin{bmatrix} A(x) & 0 \\ 0 & 0 \end{bmatrix} Q^T(x) \qquad (6)$$

where A(x) is a diagonal matrix with nonzero diagonal entries and, furthermore, if G(x) is of rank r, then A(x) is of order r. By Lemmas 3 and 4

$$G^{\ddagger}(x) = [Q^{-1}(x)]^T \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{-1}(x)$$

$$= Q(x) \begin{bmatrix} A^{-1}(x) & 0 \\ 0 & 0 \end{bmatrix} Q^T(x). \qquad (7)$$

Furthermore, $G^{\ddagger}(x)$ is positive semidefinite and symmetric. Notice that if $J(x)$ is of rank n,[1] then $G(x)$ is nonsingular and (7) implies

$$G^{\ddagger}(x) = G^{-1}(x) = Q(x)\ A^{-1}(x)\ Q^{T}(x). \tag{8}$$

II. 1.2.2 Second Method

For each $x \in R^{n}$, $G^{\ddagger}(x)$ may also be computed by performing simultaneous row and colum operations of $G(x)$ until $G(x)$ may be written in the form

$$G(x) = Q(x)\ M(x)\ Q^{T}(x)$$

where the upper left hand corner of $M(x)$ contains a nonsingular symmetric matrix of order equal to the rank of $G(x)$, and $Q(x)$ is an orthogonal permutation matrix containing zeros and ones. We write $Q(x)$ to stress the fact that for different values of x the permutation matrix $Q(x)$ may change, even though it is a constant matrix for each fixed value of x.

Since $G(x)$ is positive semidefinite and symmetric, $M(x)$ is also positive semidefinite and symmetric. Hence, by Lemma 3, we may partition $M(x)$ as follows:

$$M(x) = \begin{bmatrix} A(x) & C(x) \\ C^{T}(x) & B(x) \end{bmatrix} \tag{10}$$

where $A(x)$ is nonsingular, and A and B are symmetric matrices. By Lemma 4 we have

$$M(x) = \begin{bmatrix} I \\ P(x) \end{bmatrix} A(x) \quad [I \quad P^{T}(x)] \tag{11}$$

where $P(x)$ is, for each choice of x, a constant matrix, but which may vary for different choices of x. Lemma 4 implies

$$M^{\ddagger}(x) = \begin{bmatrix} I \\ P(x) \end{bmatrix} \left\{[I + P^{T}(x)\ P(x)]\ A(x)\ [I + P^{T}(x)\ P(x)]\right\}^{-1} [I \quad P^{T}(x)]. \tag{12}$$

Since Q(x) is orthogonal, repeated application of Lemma 5 yields

$$G^{\mp}(x) = Q(x)\ M^{\mp}(x)\ Q^T(x)$$

$$= S(x) \left\{ [I + P^T(x)\ P(x)]\ A(x)\ [I + P^T(x)\ P(x)] \right\}^{-1} S^T(x) \quad (13)$$

where

$$S(x) \triangleq Q(x) \begin{bmatrix} I \\ P(x) \end{bmatrix}. \quad (14)$$

II. 1.2.3 Third Method

Let $H(x^k) \triangleq G^{\mp}(x^k)$. Suppose we have already obtained the matrix $H(x^k)$. In order to continue the iteration scheme (2), we must compute $G^{\mp}(x^{k+1})$. It would be computationally advantageous to utilize Lemma 2 as an updating technique.

If $G(x^{k+1}) - G(x^k)$ is either positive semidefinite or negative semi-definite, we may write

$$G(x^{k+1}) = \pm C^T C + G(x^k) \quad (15)$$

since both $G(x^{k+1})$ and $G(x^k)$ are symmetric. If the null space of $G(x^k)$ is contained in the null space of C, then Lemma 2, after some algebraic manipulation, indicates that $H(x^{k+1}) \triangleq G^{\mp}(x^{k+1})$ may be computed by the following equation:

$$H(x^{k+1}) \triangleq H(x^k) \mp H(x^k)\ C^T\ [I \pm C\ H(x^k)\ C^T]^{-1}\ C\ H(x^k). \quad (16)$$

This updating technique is computationally very attractive if C is r x n with r being a small positive integer compared to n.

We note that the condition (15) is often satisfied. In particular, in circuit analysis with piecewise-linear elements and no controlled sources it has been shown [11,12] that (15) will always hold with C being a 1 x n matrix.

II. 2 Properties of the Converged Values of the Iteration Equation

In this subsection, we examine the relationship between the converged value

of the iteration equation (2) and a solution of the original system (1). In particular, we show that the converged value of (2) minimizes a weighted least squares error function defined for (1). Our discussion will cover the three cases $m > n$, $m < n$ and $m = n$ separately.

Define a weighted least squares error function, $e(x)$ by

$$e(x) \triangleq [f(x) - b]^T R [f(x) - b]. \tag{17}$$

Note that $e(x)$ may also be written as:

$$e(x) = [f(x) - b]^T R^{1/2} R^{1/2} [f(x) - b] = ||R^{1/2} [f(x) - b]|| \tag{18}$$

where $R^{1/2}$ is the square root matrix of R such that $R^{1/2} R^{1/2} = R$. $R^{1/2}$ is positive definite and symmetric [13].

It follows that

$$\nabla e(x) = 2 J^T(x) R [f(x) - b] \tag{19}$$

where $\nabla e(x) = [\frac{\partial e(x)}{\partial x_1}, \frac{\partial e(x)}{\partial x_2}, \ldots, \frac{\partial e(x)}{\partial x_n}]^T$ is the gradient vector of $e(x)$. We will call x^s a stationary point of $e(x)$ if $\nabla e(x^s) = 0$. In addition, if x^s is a minimum point of $e(x)$, then x^s is called a least squares solution of (1) with weight R.

II. 2.1 $m > n$

Before we proceed to discuss this case, several lemmas are required.

Lemma 6 [2]

If B is m x n with rank n, then $B\eta = 0$ if and only if $\eta = 0$.

Lemma 7

If $J(x)$ is an m x n matrix of maximal column rank (i.e. the rank of $J(x)$ is n for all x), then

$$[J^T(x) R J(x)]^{\dagger} = [J^T(x) R J(x)]^{-1}. \tag{20}$$

Indeed, $J^T(x)\ R\ J(x)$ is positive definite for all x.

Proof: Define $w(x,z) \triangleq J(x)z$. By Lemma 6, $w(x,z)$ is zero if and only if $z = 0$. Let $z \varepsilon R^n$ and let z be different from zero. Then $w(x,z) \neq 0$. Since R is positive definite, the quadratic form $w^T(x,z)\ R\ w(x,z)$ is greater than zero. But $w^T(x,z)\ R\ w(x,z) = z^T\ J^T(x)\ R\ J(x)\ z$. Thus $J^T(x)\ R\ J(x)$ is positive definite and hence nonsingular. Consequently $[J^T(x)\ R\ J(x)]^{\dagger} = [J^T(x)\ R\ J(x)]^{-1}$.

Q.E.D.

Lemma 8

Let J(x) be an m x n matrix of rank n. If the sequence $\{x^0, x^1, \ldots\}$ generated by (2) converges to a point x^*, then x^* is a stationary point of (17).

Proof: Assume the sequence generated by (2) converges and that x^* is the converged value. Then (2) implies that

$$x^* = x^* - [J^T(x^*)\ R\ J(x^*)]^{-1}\ J^T(x^*)\ R\ [f(x^*) - b]. \tag{21}$$

Thus

$$[J^T(x^*)\ R\ J(x^*)]^{-1}\ J^T(x^*)\ R\ [f(x^*) - b] = 0.$$

By Lemma 7, $J^T(x^*)\ R\ J(x^*)$ is positive definite and therefore its inverse is also positive definite. Hence,

$$J^T(x^*)\ R\ [f(x^*) - b] = 0. \tag{22}$$

In light of (19), (22) implies that $\nabla e(x^*) = 0$ and the conclusion follows.

Q.E.D.

Before we proceed to establish the main results for this case ($m > n$), several definitions are needed.

Define the matrix M(x) by

$$M(x) \triangleq J(x)\ J^T(x) \tag{23}$$

Note that M(x) is an m x m, positive semidefinite symmetric matrix. Define the space P(x*) by

$$P(x^*) \triangleq \{y \in R^m : [I - M^{\dagger}(x^*)\, M(x^*)]\, y = 0\}. \tag{24}$$

Clearly, $P(x^*)$ is a subspace of R^m. In addition, for every vector $y \in R^m$, define

$$y_p \triangleq M^{\dagger}(x^*)\, M(x^*)\, y \tag{25}$$

and

$$||y||_p \triangleq ||y_p|| = ||M^{\dagger}(x^*)\, M(x^*)\, y||. \tag{26}$$

Then $||y||_p$ is a semi-norm on R^m [13] and a norm on $P(x^*)$. In fact, $||y||_p$ is the Euclidean R^m norm of the projected vector of y onto the subspace $P(x^*)$.

Now consider e(x) as defined in (17). If $[f(x) - b]$ is projected onto the subspace $P(x^*)$, then e(x) takes the value

$$\begin{aligned} e^p(x) &\triangleq [f(x) - b]_p^T\, R\, [f(x) - b]_p \\ &= ||R^{1/2}\, [f(x) - b]_p||. \end{aligned} \tag{27}$$

Theorem 1

Let J(x) be an m x n matrix with rank n. Suppose that the sequence $\{x^0, x^1, \ldots,\}$ generated by (2) converges to a point x^*. Then x^* is a least squares solution of (1) with weight R when the vectors in R^m are projected onto $P(x^*)$. That is, x^* is a minimum value of (27).

Proof: By Lemma 8, x^* is a stationary point of (18). We must show that there exists an $\varepsilon_0 > 0$ such that

$$e^p(x^* + \varepsilon z) \triangleq [f(x^* + \varepsilon z) - b]_p^T\, R\, [f(x^* + \varepsilon z) - b]_p \geq e^p(x^*) \tag{28}$$

whenever $|\varepsilon| < \varepsilon_0$ and for all $z \in R^n$ such that $||z|| = 1$.

Write

$$f(x^* + \varepsilon z) = f(x^*) + \varepsilon\, J(x^*) z + o(\varepsilon) \tag{29}$$

where

$$\lim_{\varepsilon \to 0} \frac{||o(\varepsilon)||}{\varepsilon} = 0 .$$

Substituting, we have, after some manipulation,

$$e^p(x^* + \varepsilon z) = e^p(x^*) + 2\,[f(x^*) - b]_p^T R\,[o(\varepsilon)]_p$$
$$+ 2\varepsilon\{[f(x^*) - b]_p^T R\,[J(x^*)z]_p + [J(x^*)z]_p^T R\,[o(\varepsilon)]_p\}$$
$$+ \varepsilon^2\,[J(x^*)z]_p^T R\,[J(x^*)z]_p + [o(\varepsilon)]_p^T R\,[o(\varepsilon)]_p. \quad (30)$$

Thus

$$e^p(x^* + \varepsilon z) = e^p(x^*) + 2\,[f(x^*) - b]_p^T R\,[o(\varepsilon)]_p$$
$$+ 2\varepsilon[f(x^*) - b]_p^T R\,[J(x^*)z]_p$$
$$+ \varepsilon^2[J(x^*)z]_p^T R\,[J(x^*)z] + \hat{o}(\varepsilon^2) \quad (31)$$

where $o(\varepsilon) \triangleq 2\varepsilon[J(x^*)z]_p^T R[o(\varepsilon)]_p + [o(\varepsilon)]_p^T R\,[o(\varepsilon)]_p$, and

$$\lim_{\varepsilon \to o} \frac{||\hat{o}(\varepsilon^2)||}{\varepsilon^2} = 0$$

Now consider the term

$$[f(x^*) - b]_p^T R\,[o(\varepsilon)]_p = [o(\varepsilon)]_p^T R\,[f(x^*) - b]_p. \quad (32)$$

Notice that $M(x^*)$ is a symmetric matrix. Hence $M^{\ddagger} M$ is a diagonal matrix with diagonal elements being zeros or ones. Thus

$$M^{\ddagger}(x^*)\,M(x^*)\,R = R\,M^{\ddagger}(x^*)\,M(x^*). \quad (33)$$

Since $y_p = M^{\ddagger}(x^*)\,M(x^*)y$, (32) may be written

$$[f(x^*) - b]_p^T R\,[o(\varepsilon)]_p = [o(\varepsilon)]_p^T\, M^{\ddagger}(x^*)\,J(x^*)\,J^T(x^*)\,R\,[f(x^*) - b]$$
$$= 0 \quad (34)$$

by virtue of (22).

Note that (22) implies that the term $[f(x^*) - b]_p^T R\,[J(x^*)z]_p$ must also vanish, just like the case of (34). Thus, we have,

$$e^p(x^* + \varepsilon z) = \varepsilon^p(x^*) + \varepsilon^2\,[J(x^*)z]_p^T R\,[J(x^*)z]_p + o(\varepsilon^2) \quad (35)$$

Since R is positive definite, there exists $\varepsilon_o > 0$ such that, for all $|\varepsilon| < \varepsilon_o$,

$$e^P(x^* + \varepsilon z) - e^P(x^*) > 0 \tag{36}$$

Therefore, x^* is a least squares solution of (1) with weight R when all vectors in R^m are projected onto $P(x^*)$.

Q.E.D.

Since $m > n$, (22) does not require $R[f(x^*) - b] = 0$. Indeed, (22) requires only that the vector $R[f(x^*) - b]$ be orthogonal to the matrix $J(x^*)$. Thus the solution to (22) is not necessarily unique. In fact, the null space of $J^T(x^*)$ is an (m-n) dimensional space -- $R[f(x^*) - b]$ may take on any value in this (m-n) dimensional space and still satisfy (22). In general, we have

$$R\,[f(x^*) - b] = \delta \tag{37}$$

where

$$\delta \triangleq [\, I - M^{\dagger}(x^*)\, M(x^*)]y \tag{38}$$

for some $y \in R^m$.

Equation (37) implies

$$f(x^*) = b + R^{-1}\,\delta \tag{39}$$

Thus, even if the sequence generated by (2) converges, the converged value x^*, which is a least squares solution of (1) with weight R, may <u>not</u> be an actual solution of (1). This conclusion, in general, is unavoidable since there are more equations than unknowns so that overspecification of constraints among the n variables may occur. Hence, there may not exist a vector $\bar{x} \in R^n$ such that $f(\bar{x}) = b$ is satisfied. In this case, a least squares solution is probably the best that may be hoped for.

One way to overcome this difficulty is to put weights on the equations of (1). For example, if the equations $j_1, j_2, \ldots, j_k$ are thought to be very

reliable or very important, if the equations $j_{k+1}, j_{k+2},\ldots,j_p$ are felt to be trustworthy and if the remainder of the equations $j_{p+1}, j_{p+2},\ldots,j_m$ are considered to be less reliable or trustworthy, we may then choose $R = [r_{ij}]$ to be a diagonal matrix with

$$r_{j_\alpha j_\alpha} > r_{j_\beta j_\beta} > r_{j_\gamma j_\gamma}$$

where $\alpha = 1,2,\ldots,k$; $\beta = k+1, k+2,\ldots, p$ and $\gamma = p+1, p+2,\ldots,m$. This point will be expanded upon in Section III.

It may occur that δ in (38) is zero. In this case x^* is an actual solution of (1). Consider the special case of a linear equation of the form $f(x) = Jx$ where J is an m x n matrix. Thus we consider the equation

$$Jx = b \tag{40}$$

and (39) becomes

$$Jx^* = b + R^{-1}\,\delta \tag{41}$$

and the general solution has the form [9]

$$x^* = J^{\dagger}\,(b + R^{-1}\,\delta) + J^{\dagger}\,[I - J\,J^{\dagger}]z \tag{42}$$

where z is an arbitrary vector in R^m. Note that (38) implies that δ is a vector in R^m orthogonal to the space S_M spanned by the column vectors $M(x^*)$. In view of (23), S_M is also spanned by the column vectors of J. Since R is nonsingular, the vector $R^{-1}\,\delta$ is also orthogonal to S_M. Thus we may write,

$$R^{-1}\,\delta = [I - J\,J^{\dagger}]\,\tilde{z} \tag{43}$$

for some $\tilde{z}$ in R^m. Hence (42) becomes

$$x^* = J^{\dagger}\,b + J^{\dagger}\,[I - J\,J^{\dagger}]y \tag{44}$$

where $y \triangleq z + \tilde{z}$ is an arbitrary vector in R^m since z in (42) is arbitrary.

Equation (44) is precisely the general solution of (40) [15], and the best approximation solution [16] of (40) is obtained by setting y in (44) to zero. Hence, (39) may be viewed as a logical generalization of the linear case.

II.2.2 $m < n$

Since $m < n$, the maximal rank of the matrix $J(x)$ is m. Recall that $G(x) \triangleq J^T(x)\ R\ J(x)$ and is therefore an $n \times n$ matrix. Thus, for this case, $G(x)$ is always singular.

Proceeding in an analogous fashion to the case $m > n$, define a subspace of R^n, $\hat{P}(x^*)$, by

$$\hat{P}(x^*) \triangleq \{z \in R^n: [I - G^{\ddagger}(x^*)\ G(x^*)]\ z = 0\} \tag{45}$$

and for each $z \in R^n$, define

$$z_{\hat{p}} \triangleq G^{\ddagger}(x^*)\ G(x^*)\ z \tag{46}$$

and

$$||z||_{\hat{p}} \triangleq ||z_{\hat{p}}|| = ||G^{\ddagger}(x^*)\ G(x^*)\ z||. \tag{47}$$

Thus $z_{\hat{p}}$ is the projection of z onto the subspace $\hat{P}(x^*)$ which is spanned by the column vectors of $G(x^*)$. Note that since $G(x)$ is symmetric, $G^{\ddagger}(x)\ G(x)$ is diagonal. Hence we have

$$G^{\ddagger}(x)\ G(x) = G(x)\ G^{\ddagger}(x).$$

Lemma 9

Let $\{x^0, x^1, ...\}$ be the sequence generated by (2) with the initial approximation to the solution of (1) being x^0. Suppose the sequence converges to a point x^*. Then x^* is a stationary point of the error function defined by (17) when the gradient is projected onto the subspace $\hat{P}(x^*)$, that is $[\nabla e(x^*)]_{\hat{p}} = 0$.

Proof: Since x^* is a converged value of (2), we have

$$x^* = x^* - [G(x^*)]^{\ddagger}\ J^T(x^*)\ R\ [f(x^*) - b]$$

which implies

$$[G(x^*)]^{\ddagger}\ J^T(x^*)\ R\ [f(x^*) - b] = 0. \tag{48}$$

Projecting $\nabla e(x)$ onto $\hat{P}(x^*)$, we have

$$[\nabla e(x)]_{\hat{P}} = G(x)\ [G(x)]^{\dagger}\ J^T(x)\ R\ [f(x) - b], \tag{49}$$

Thus, by (48),

$$[\nabla e(x^*)]_{\hat{P}} = 0. \tag{50}$$

Q.E.D.

Theorem 2

Suppose $J(x)$ is of rank m. If the sequence $\{x^0, x^1, \ldots\}$ generated by (2) converges to x^*, then x^* is a solution of (1). Hence, x^* is a stationary point of $e(x)$ and a least squares solution of (1).

Proof: By our assumptions, $[J^T(x^*)R]$ has dimension n x m and rank m, thus $[J^T(x^*)R]^{\dagger}\ J(x^*)\ R = I$ [9]. Furthermore, since x^* is a converged value of the sequence generated by (2), (48) may be written as

$$[J^T(x^*)\ R\ J(x^*)]^{\dagger}\ J^T(x^*)\ R\ [f(x^*) - b] = 0 \tag{51}$$

By Lemma 5, we have

$$[J^T(x^*)\ R\ J(x^*)]^{\dagger} = [J(x^*)]^{\dagger}\ [J^T(x^*)\ R]^{\dagger}.$$

Thus, in view of our earlier remark, (51) becomes

$$[J(x^*)]^{\dagger}\ [f(x^*) - b] = 0. \tag{52}$$

Note that $J(x^*)$ is itself of dimension m x n and rank m, so $[J(x^*)]^{\dagger}$ is of dimension n x m and rank m. Therefore, by Lemma 6, (52) implies

$$f(x^*) - b = 0. \tag{53}$$

That is, if x^* is a converged value of a sequence generated by (2), then x^* is a solution of (1). Clearly, (53) also implies $e(x^*) = 0$ and $\nabla e(x^*) = 0$.

Q.E.D.

Note that Theorem 2 neither asserts that the sequence generated by (2) may converge to only one point nor that the solution of (1) is unique. Indeed, there may exist an infinite number of x^* which satisfy (53). Since

we are dealing with the case $m < n$, there are more unknowns in (1) than equations constraining the unknowns. Thus, even in the linear case, we would not expect a unique solution.

II.2.3 $m = n$

In this case $J(x)$ and $G(x)$ are both square matrices of order n. Furthermore, $G(x)$ is nonsingular if and only if $J(x)$ is nonsingular. The proofs of Lemma 10 and Theorem 3 will not be given since they are parallel to the proofs of Lemma 8 (Lemma 9) and Theorem 1 (Theorem 2), respectively.

Lemma 10

Let $\{x^0, x^1, \ldots\}$ be the sequence generated by (2) with x^0 being the initial approximation to the solution of (1). Suppose the sequence converges to a point x^*. Then x^* is a stationary point of the error function defined by (17), where the gradient is projected onto the subspace $\hat{P}(x^*)$. If $J(x^*)$ is nonsingular, then $\nabla e(x^*) = 0$.

Theorem 3

Suppose $J(x)$ is of rank n. If the sequence $\{x^0, x^1, \ldots\}$ generated by (2) converges to x^*, then x^* is a solution of (1). Hence x^* is a stationary point of $e(x)$ and a least squares solution of (1) with weight R.

Just as in the previous two cases, the solution of (1) is not necessarily unique even when $m = n$. The difficulty here arises from considerations which are quite different from the previous cases. In this case there are, at most, a countably infinite number of x^* which satisfy (53). If a certain norm condition on $f(x)$ is satisfied, then x^* is the unique solution of (1) [17,18].

II.3. Computational Procedures

In this subsection we shall consider a computational procedure for (2) which resembles, in some respects, the procedure used in [1]. In addition,

we show that the iteration procedure either reduces the error function defined in (17) with each cycle of computation or else reaches a stationary point of (17). In the latter case, by Theorems 1, 2 and 3 we consider the procedure to have reached a (least squares) solution of (1) and hence the computation is terminated.

Consider the case when the iteration is at the k^{th} stage. Let the direction of search be denoted by p^k. Then

$$p^k \triangleq -[J^T(x^k)\ R\ J(x^k)]^{\dagger}\ J^T(x^k)\ R\ [f(x^k) - b]. \tag{54}$$

Denote the next iterate by

$$x^{k+1} = x^k + s_k p^k \tag{55}$$

where s_k is a scalar chosen either to satisfy

$$||f(x^k + s_k p^k) - b|| < ||f(x^k) - b|| \tag{56}$$

or such that the function

$$||f(x^{k+1}) - b|| = ||f(x^k + s_k p^k) - b||$$

$$= \min_{s>0} ||f(x^k + s\ p^k) - b||. \tag{57}$$

To see that it is always possible to satisfy (56) or (57) whenever $\nabla e(x^k) \neq 0$, consider $[-p^k]^T\ \nabla e(x^k)$. We have

$$[-p^k]^T\ \nabla e(x^k) = 2[f(x^k) - b]^T\ R\ J(x^k)\ \left\{[J^T(x^k)\ R\ J(x^k)]^{\dagger}\right\}^T.$$
$$\left\{J^T(x^k)\ R\ [f(x^k) - b]\right\}. \tag{58}$$

Since the matrix $J^T(x^k)\ R\ J(x^k)$ is positive semidefinite, its generalized inverse is also positive semidefinite. Thus, by (58),

$$[-p^k]^T\ \nabla e(x^k) \geq 0. \tag{59}$$

By treating the cases $m \geq n$ and $m \leq n$ separately, it can be shown that in both cases (58) may be written as:

$$[-p^k]^T \nabla e(x^k) = 2[f(x^k) - b]\{J^T(x^k)\}^{\dagger} J^T(x^k) R [f(x^k) - b] \quad (60)$$

Now, x^k is not a stationary point of e(x), i.e.

$$\nabla e(x^k) = J^T(x^k) R [f(x^k) - b] \neq 0. \quad (61)$$

Thus R $[f(x^k) - b]$ is not orthogonal to all the column vectors of $J(x^k)$. Since the matrix $[J^T(x^k)]^{\dagger} J^T(x^k)$ is a matrix which projects vectors in R^m onto the subspace spanned by the column vectors of $J(x^k)$, we have

$$[J^T(x^k)]^{\dagger} J^T(x^k) R [f(x^k) - b] \neq 0. \quad (62)$$

From (60) and (62) we see that the inequality in (59) is strict and therefore

$$[-p^k]^T \nabla e(x^k) > 0 \quad (63)$$

whenever $\nabla e(x^k) \neq 0$.

Thus p^k has components along the negative gradient of the error function of (17). Hence the iteration equation yields a downhill process. That is, there exists s_k such that (56) and (57) are possible whenever $\nabla e(x^k) \neq 0$.

The preceding discussion justifies the following two iteration procedures which will be presented in the form of algorithms.

Algorithm 1

Step 1: Let k = 0. Choose an initial estimate x^0.

Step 2: If $||f(x^k) - b|| < \varepsilon_1$, stop. If not, go to step 3.

Step 3: If $||\nabla e(x^k)|| < \varepsilon_2$, terminate the computation.

Otherwise, go to step 4.

Step 4: Compute [3]

$$p^k = -[J^T(x^k) R J(x^k)]^{\dagger} J^T(x^k) R [f(x^k) - b] \quad (64)$$

by solving

$$J^T(x^k) R J(x^k) p^k = - J^T(x^k) R [f(x^k) - b] \quad (65)$$

Step 5: Let $x^{k+1} = x^k + s_k p^k$

where s_k is chosen to satisfy either (56) or (57).

Step 6: Increment k by 1 and go to step 2.

In general (56) is computationally more efficient and the program is also simpler.

It is possible to utilize the information gained from the fact that $[-p^k]^T \nabla e(x^k)$ indicates the descending direction of the error function to modify step 5 in Algorithm 1. The resulting algorithm, which we shall call Algorithm 2, is the same as Algorithm 1 except that steps 5 and 6 are replaced by steps 5' and 6':

Step 5': Let $S^k = [s_{ij}^k]$ be a diagonal matrix such that

$$s_{ii}^k = s_i^k \text{ if } -p_i^k [\nabla e(x^k)]_i > 0 \tag{66}$$

or

$$s_{ii}^k = 0 \text{ if } -p_i^k [\nabla e(x^k)]_i \leq 0 \tag{67}$$

for $i = 1, 2, \ldots, n$ and where the s_i^k are chosen such that

$$||f(x^k + S^k p^k) - b|| < ||f(x^k) - b|| \tag{68}$$

or

$$||f(x^k + S^k p^k) - b|| = \min_{S \in \delta} ||f(x^k + S p^k) - b|| \tag{69}$$

where δ is the set of diagonal matrices where entries satisfy (66) and (67).

Step 6': Let $x^{k+1} = x^k + S^k p^k$.

As an illustration of the use of the preceding algorithm, consider the following system of equations:

$$x_1^2 - 3x_2 = 34$$

$$x_1 + x_2^2 = 14$$

$$x_1 x_2 = -15.$$

This system was first solved by Algorithm 1 with $x^0 = [0 \quad 0]^T$ and $R = I$. The results of this calculation are summarized in Table I. Step 5 was implemented using (56).

Table I

k	$[x^k]^T$	$[p^k]^T$	s_k	$e(x^k)$
0	[0 0]	[14 -11.33]	.25	1577
1	[3.5 -2.83]	[6.65 3.90]	.25	207.52
2	[5.16 -1.86]	[-.77 -1.77]	1	61.36
3	[4.39 -3.62]	[1.12 .79]	1	27.99
4	[5.51 -2.84]	[-1.52 - .93]	.5	24.35
5	[4.75 -3.31]	[.75 .64]	.5	5.56
6	[5.13 -2.98]	[-.39 - .21]	.5	1.61
7	[4.93 -3.09]	[.22 .20]	.5	.47
8	[5.04 -2.99]	[-.13 - .07]	.5	.16
9	[4.98 -3.03]	[.08 .07]	.5	.05
10	[5.01 -3.00]			.02

The system was then solved by Algorithm 2 with, once again, $x^0 = [0 \quad 0]^T$ and R = I. The results of this calculation are summarized in Table II. Step 5' was implemented using (68).

Table II

k	$[x^k]^T$	$[p^k]^T$	$[\nabla e(x^k)]^T$	S^k	$e(x^k)$
0	[0 0]	[14 -11.33]	[-28 204]	$\begin{bmatrix} .25 & 0 \\ 0 & .25 \end{bmatrix}$	1577
1	[3.5 -2.83]	[6.65 3.9]	[-219.27 143.13]	$\begin{bmatrix} .25 & 0 \\ 0 & 0 \end{bmatrix}$	207.52
2	[5.16 -2.83]	[-.53 - .44]	[18.45 4.07]	$\begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}$	2.13
3	[4.9 -3.06]	[.32 .21]	[-15.88 2.40]	$\begin{bmatrix} .25 & 0 \\ 0 & 0 \end{bmatrix}$	.74
4	[4.98 -3.06]	[.06 .09]	[1.24 -5.75]	$\begin{bmatrix} 0 & 0 \\ 0 & .5 \end{bmatrix}$	.15
5	[4.98 -3.01]	[.07 .04]	[-3.16 .36]	$\begin{bmatrix} .25 & 0 \\ 0 & 0 \end{bmatrix}$	.03
6	[5.00 -3.01]				.01

II. 4 A Modified Iteration Equation

The algorithms in subsection II.3 require the computation of the generalized inverse of $G(x^k)$ or equivalently, the computation of p^k of (65) at each new iterated point. The computation of p^k generally dominates the complexity of the entire computation.

In this section, we consider a modified iteration equation:

$$x^{k+1} = x^k - [J^T(x^0)\ R\ J(x^0)]^{\dagger} J^T(x^0)\ R\ [f(x^k) - b] \tag{70}$$

We now give a theorem which assures the convergence of the sequence generated by (70) under certain conditions.

Theorem 4

Let $f: R^n \to R^m$ be a C^2 map with $m \geq n$ and let x^0 be the initial approximate solution of (1). Let α, β, γ_1, γ_2 and γ_3 be positive constants such that

1. $$||H(x^0)|| \leq \alpha \tag{71}$$

   where $$H(x^0) \triangleq [J^T(x^0)\ R\ J(x^0)]^{\dagger}, \tag{72}$$

2. $$||H(x^0)\ F(x^0)|| \leq \beta \tag{73}$$

   where $$F(x) \triangleq J^T(x^0)\ R\ [f(x) - b], \tag{74}$$

3. $||R|| \leq \gamma_1$; $||J(x^0)|| \leq \gamma_2$; $||J'(x)|| \leq \gamma_3$, and $J(x)$ is of rank n for all x in $N(x^0)$ with

$$N(x^0) \triangleq \{x \in R^n: ||x - x^0|| \leq \rho(h)\beta\} \tag{75}$$

where

$$\rho(h) = \frac{1 - \sqrt{1 - 2h}}{h} \tag{76}$$

and $J'(x)$ is the derivative of $J(x)$ or the second derivative of $f(x)$ [6], and

4. $$h \triangleq \alpha\ \beta\ \gamma < 1/2$$

   where $$\gamma \triangleq \gamma_1\ \gamma_2\ \gamma_3. \tag{77}$$

Then the sequence $\{x^0, x^1, \ldots\}$ generated by the modified iteration equation

$$x^{k+1} = x^k - [J^T(x^0)\ R\ J(x^0)]^{\dagger}\ J^T(x^0)\ R\ [f(x^k) - b] \tag{78}$$

converges to a point x^* in $N(x^0)$ with the rate

$$||x^{k+1} - x^*|| \leq q^{k+1}\ ||x^0 - x^*|| \tag{79}$$

where

$$q \triangleq 1 - \sqrt{1 - 2h} < 1. \tag{80}$$

Proof: By condition 3, $J(x^0)$ is of rank n, so the matrix

$$G(x^0) \triangleq J^T(x^0)\ R\ J(x^0) \tag{81}$$

is, by Lemma 7, nonsingular. Thus we have,

$$H(x^0) = G^{-1}(x^0). \tag{82}$$

Let us define

$$g(x) \triangleq x - H(x^0)\ J(x^0)\ R\ [f(x) - b]. \tag{83}$$

Differentiating (83), we have

$$g'(x) = I - H(x^0)\ J(x^0)\ R\ J(x) \tag{84}$$

and

$$g''(x) = -H(x^0)\ J(x^0)\ R\ J'(x) \tag{85}$$

where $g'(x)$ and $g''(x)$ denote the first and second derivatives of $g(\cdot)$, respectively, evaluated at the point x.

Note that

$$g'(x^0) = I - H(x^0)\ J(x^0)\ R\ J(x^0) = I - H(x^0)\ G(x^0) = 0 \tag{86}$$

by virtue of (82). Furthermore, whenever $x \epsilon N(x^0)$,

$$||g''(x)|| \leq ||H(x^0)||\ \ ||J(x^0)||\ \ ||R||\ \ ||J'(x)|| \leq \alpha\ \gamma \tag{87}$$

From (78) and (83), we see that

$$x^{k+1} = g(x^k) \tag{88}$$

and, by condition 2,

$$||x^1 - x^0|| = ||H(x^0)\ F(x^0)|| \leq \beta. \tag{89}$$

Using a Taylor's series expansion for $\rho(h)$, we find

$$\rho(h) = \frac{1 - \sqrt{1 - 2h}}{h} = \frac{1 - [1 - h - \frac{h^2}{2!} - \frac{3h^3}{3!} - \frac{15h^4}{4!} \cdots\cdots]}{h}$$

$$= 1 + \frac{h}{2!} + \frac{3h^2}{3!} + \frac{15h^3}{4!} + Q\ (h) \tag{90}$$

where Q(h) contains terms of the form $a_r h^r$ with $a_r > 0$ for all r. Hence

$$\rho(h) \geq 1 \qquad \text{for } 0 \leq h \leq 1/2. \tag{91}$$

Thus, we may write (89) as

$$||x^1 - x^0|| \leq \beta \leq \rho(h)\beta. \tag{92}$$

Therefore, $x^1 \epsilon N(x^0)$.

Let $x \epsilon N(x^0)$, then (86), (88) and (89) imply

$$\begin{aligned} ||g(x) - x^0|| &\leq ||g(x) - x^1|| + ||x^1 - x^0|| \\ &\leq ||g(x) - g(x^0)|| + \beta \\ &= ||g(x) - g(x^0) - g'(x^0)(x - x^0)|| + \beta. \end{aligned} \tag{93}$$

Using a Taylor's series expansion on g(x), (93) becomes

$$||g(x) - x^0|| \leq ||\tfrac{1}{2} g''(\bar{x})|| \;\; ||x - x^0||^2 + \beta. \tag{94}$$

where $\bar{x}$ is some point on the line segment $L(x,x^0)$ joining x and x^0. Since $N(x^0)$ is closed and convex, x^0 and x are in $N(x^0)$. Thus (87) applies at the point x and (94) becomes

$$\begin{aligned} ||g(x) - x|| &\leq 1/2 \; \alpha\gamma\rho^2(h)\beta^2 + \beta \\ &= 1/2 \; h\rho^2(h)\beta + \beta \\ &= \rho(h)\beta. \end{aligned} \tag{95}$$

Therefore, $g(x) \epsilon N(x^0)$. Summarizing, we have shown that if $x \epsilon N(x^0)$, then $g(x) \epsilon N(x^0)$.

Consider the sequence $\{x^0, x^1, \ldots\}$ generated by (78). Now $x^0 \epsilon N(x^0)$ by definition; hence $x^1 = g(x^0) \epsilon N(x^0)$. Similarly, $x^2 = g(x^1) \epsilon N(x^0)$. Proceeding inductively, we conclude that every element in the sequence $\{x^0, x^1, \ldots\}$ is in $N(x^0)$. $N(x^0)$ is a compact set in R^n; therefore there exists a subsequence $\{x^{\ell_1}, x^{\ell_2}, x^{\ell_3}, \ldots\}$ of the original sequence $\{x^0, x^1, \ldots\}$ and a limit point x^* [6] such that

$$\lim_{j\to\infty} x^{\ell_j} = x^*. \tag{96}$$

That is, x^* satisfies the equation

$$x^* = g(x^*). \tag{97}$$

Let x be a vector in $N(x^0)$. Then by (86) and (97) we have

$$\begin{aligned} ||g(x) - x^*|| &= ||g(x) - g(x^*)|| \\ &= ||g(x) - g(x^*) - g'(x^0)(x - x^*)||. \end{aligned} \tag{98}$$

By repeated application of the mean value theorem [6], (98) may be written as:

$$||g(x) - x^*|| \leq ||g''(\tilde{x})|| \quad ||\hat{x} - x^0|| \quad ||x - x^*|| \tag{99}$$

for some $\tilde{x}$ and $\hat{x}$ where $\hat{x}$ is a point on the line segment $L(x,x^*)$ and $\tilde{x}$ is a point on the line segment $L(\hat{x},x^0)$. Since x and x^* are both in $N(x^0)$ so is $\hat{x}$ and, consequently, $\tilde{x}$ is also in $N(x^0)$. Thus (87) applies and (99) becomes

$$||g(x) - x^*|| \leq \alpha\,\gamma\,||\hat{x} - x^0|| \quad ||x - x^*||. \tag{100}$$

Since $\hat{x}$ is on $L(x,x^*)$, we may write

$$\hat{x} = \theta x + (1 - \theta)x^* \text{ for some } \theta\varepsilon[0,1]. \tag{101}$$

Thus

$$\begin{aligned} ||\hat{x} - x^0|| &= ||\theta(x - x^0) + (1 - \theta)(x^* - x^0)|| \\ &\leq \max\{||x - x^0||,\ ||x^* - x^0||\}. \end{aligned} \tag{102}$$

Since both x and x^* are in $N(x^0)$, (102) implies

$$||\hat{x} - x^0|| \leq \rho(h)\beta. \tag{103}$$

From (100) and (103), we have

$$||g(x) - x^*|| \leq \alpha\,\beta\,\gamma\,\rho(h)\ ||x - x^*|| = q\ ||x - x^*||. \tag{104}$$

Since we have shown that every element of the sequence $\{x^0, x^1, \ldots\}$ is in

$N(x^0)$, (104) applies to each point in the sequence; thus we have,

$$||g(x^k) - x^*|| \leq q\ ||x^k - x^*||, \quad k = 0,1,2,\ldots . \tag{105}$$

Using (88) and (105), we find

$$\begin{aligned} ||x^{k+1} - x^*|| &\leq q\ ||x^k - x^*|| \\ &\leq q\ ||x^k - x^*|| \leq q^2\ ||x^{k-1} - x^*||. \ . \ . \\ &\leq q^{k+1}\ ||x^0 - x^*|| \end{aligned} \tag{106}$$

Thus the entire sequence $\{x^0, x^1, \ldots\}$ converges to a limit point x^* with the rate

$$||x^{k+1} - x^*|| \leq q^{k+1}\ ||x^0 - x^*||. \tag{107}$$

Q.E.D.

Theorem 4 suggests an algorithm which may be used to implement the modified iteration equation (70). This algorithm has the attractive feature that it does not call for the computation of the generalized inverse of the matrix $G(x^k)$ at each new iterate.

Algorithm 3

Step 1: Let $k = 0$. Choose an initial estimate x^0.

Step 2: If $[f(x^0) - b] \leq \varepsilon$, terminate the computation. The solution is x^0. If $[f(x^0) - b] > \varepsilon$, go to step 3.

Step 3: Check to see if the conditions of Theorem 4 are satisfied at x^0. If they are not satisfied go to step 1. If the conditions are satisfied go to step 4.

Step 4: Compute $T \triangleq [J^T(x^0)\ R\ J(x^0)]^{\dagger}\ J^T(x^0)\ R$.

Step 5: Compute $x^k = x^{k-1} - T\ [f(x^{k-1}) - b]$.

Step 6: If $[f(x^k) - b] \leq \varepsilon$, terminate the computation, x^k is the solution. If $[f(x^k) - b] > \varepsilon$, increment k by 1 and go to step 5.

As we have indicated, the algorithms in subsection II.3 (implementations of the iteration equation (2)), either reduce the norm $||f(x^k) - b||$ or else have reached a stationary point of the error function. Since the iteration equation (2) is quite similar to a Newton-Raphson iteration equation, we may say with confidence that when the initial guess is near enough to the actual solution of (1), the convergence of the algorithms in II.3 is quadratic.

In Theorem 4, the modified iteration equation of (70) converges only linearly. However, the amount of computation per iteration required by (70) is much less than that required by (2). This compensates for the slow convergence rate of (70). Thus, implementing (70) will be superior to using (2) on many occasions. Clearly, it is possible to take advantage of the strengths of each algorithm by using a combination of algorithms 1, 2, and 3.

III. USES OF THE R MATRIX

III.1 Conditioning the Jacobian Matrix

Let $\delta(G)$ be the largest absolute value among the entries of $G = [g_{ij}]$, i.e.,

$$\delta(G) = \max_{i,j = 1,\ldots,n} |g_{ij}|.$$

A nonsingular matrix G is said to be ill-conditioned if $\delta(G^{-1})$ is very large compared to $\delta(G)$. This condition normally occurs when the determinant of G is very small.[4]

Consider first the case with $J^T(x)J(x)$ nonsingular. The matrix R in (2) may be used to great advantage in conditioning the triple product $G(x) \triangleq J^T(x)RJ(x)$. For example, if $\det J^T(x)J(x) = \varepsilon^2$ where ε is small, we may choose a positive definite symeetric matrix R such that $\det R = \varepsilon^{-2}$. In this manner, $\det G(x) = 1$ and this will imply that $\delta(G^{-1})$ is comparable to $\delta(G)$.

For example, let f(x) be a linear function of x. In particular let

$$f(x) \triangleq Jx \triangleq \begin{bmatrix} .2 & 6 \\ 2 & 6.00001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{108}$$

J is an ill-conditioned matrix since

$$J^{-1} = \begin{bmatrix} 300000.5 & -300000 \\ -100000 & 100000 \end{bmatrix} \tag{109}$$

and $\delta(J^{-1}) = 300000.5$ is large compared to $\delta(J) = 6.00001$. This occurs since $\det J = 2 \times 10^{-5}$ is very small compared to the entries of J.

Now if we consider $G \triangleq J^T R J$, we find

$$G = \begin{bmatrix} 2 & 6 \\ 2 & 6.00001 \end{bmatrix} \begin{bmatrix} r_1 & r_2 \\ r_3 & r_4 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 2 & 6.00001 \end{bmatrix}$$

$$= \begin{bmatrix} 4a & 12a + 2\times10^{-5}b \\ 12a + 2\times10^{-5}c & 36a + 6\times10^{-5}(b+c) + 10^{-10}d \end{bmatrix}$$

where $a \triangleq r_1 + r_2 + r_3 + r_4$, $b \triangleq r_2 + r_4$, $c \triangleq r_3 + r_4$ and $d \triangleq r_4$.

Then,

$$\det G = 4 \times 10^{-10} (r_1 r_4 - r_2 r_3).$$

In order to make G well-conditioned, we choose

$$\det R = r_1 r_4 - r_2 r_3 = \varepsilon^{-2} = 1/4 \times 10^{10}.$$

A suitable choice for this example is $r_1 = r_4 = 1/2 \times 10^5$ and $r_2 = r_3 = 0$.

With this choice we find

$$\det G = 4 \times 10^{-10} \left[\tfrac{1}{4} \times 10^{10}\right] = 1,$$

$$G = \begin{bmatrix} 4 \times 10^5 & 12 \times 10^5 + 1 \\ 12 \times 10^5 + 1 & 36 \times 10^5 + 6 + .5 \times 10^{-5} \end{bmatrix}$$

and thus

$$G^{-1} = \begin{bmatrix} 36 \times 10^5 + 6 + .5 \times 10^{-5} & -(12 \times 10^5 + 1) \\ -(12 \times 10^5 + 1) & 4 \times 10^5 \end{bmatrix} .$$

Clearly $\delta(G) \approx \delta(G^{-1})$. In general, this makes G a better conditioned matrix than the J matrix of (108).

Consider now the case where $J^T(x)\ J(x)$ is singular. Let the rank of $J^T(x)\ J(x)$ be r and let

$$\max_S \left[J^T(x)\ J(x)\right]_r = \varepsilon^2 \tag{110}$$

where $\left[J^T(x)\ J(x)\right]_r$ denotes the determinant of an r x r submatrix of $J^T(x)\ J(x)$ and S is the set of all possible r x r submatrices of $J^T(x)\ J(x)$.

Define

$$M(x) \triangleq \left[J^T(x)\ J(x)\right]^{\dagger} \left[J^T(x)\ J(x)\right]. \tag{111}$$

Then M(x) is a diagonal matrix with diagonal elements being either zero or one.

Choose $R = [r_{ij}]$ in the following manner:

$$r_{ii} = \eta \text{ if } [M(x)]_{ii} = 1$$

$$r_{ii} = 1 \text{ if } [M(x)]_{ii} = 0$$

where $\eta^r = \varepsilon^{-2}$. It may be shown in a manner similar to the case with $J^T(x)\ J(x)$

nonsingular, that $\delta([J^T(x)\ R\ J(x)]^{\dagger})$ is approximately the same order of magnitude as $\delta([J^T(x)\ R\ J(x)])$.

Thus, R may be chosen to better the condition of the triple product $J^T(x)\ R\ J(x)$.

III. 2 Weighting Equations

Call $[f_k(x^*) - b_k]$ the residue of the k^{th} equation of the system (1). Suppose that the equations $f_1(\cdot) = b_1$, $f_2(\cdot) = b_2, \ldots, f_m(\cdot) = b_m$ are not all equally important in the sense that relatively large residues in some component equations are more tolerable than in others. Thus an a priori decision is made that certain component equations of $[f(x) - b]$ may have larger residue values than others. To implement this decision, we choose R = diag $[r_1, r_2, \ldots, r_m]$ judiciously. The selection scheme is straightforward: Pick r_j to be relatively large if the residue of the j^{th} equation may not be allowed to be large and pick r_j to be relatively small if the j^{th} residue may be large. Since the iteration equation (2) will attempt to converge to a weighted least squares solution of (1), those r_j which are large will have a greater influence on the converged value than those that are small. Such situations may occur when modeling physical phenomena, such as in power systems.

For example, let (1) be the following system:

$$f_1(x) = x_1^2 + x_2^2 + 2 = 0$$
$$f_2(x) = x_1 + 4x_2 + 7 = 0$$
$$f_3(x) = 2x_1 + 9x_2 + 1 = 0.$$

If we wish to have a solution which will minimize $|f_1(x)|$ (which means, of course, that we would like to have $f_1(x)$ solved as exactly as possible), we may choose $R = [r_{ij}]$ to be given by $r_{11} = 10^5$, $r_{22} = r_{33} = 1$ and $r_{ij} = 0$ for $i \neq j$. Then (17) becomes

$$e(x) = r_{11}(x_1^2 + x_2^2 + 2)^2 + r_{22}(x_1 + 4x_2 + 7)^2 + r_{33}(2x_1 + 9x_2 + 1)^2. \quad (112)$$

Equation (112) is minimized when $\bar{x}_1 = \bar{x}_2 = 0$. Thus $e(\bar{x}) \approx 4 \times 10^5$. Note that the values of $\tilde{x}_1 = 59$, $\tilde{x}_2 = -13$ exactly satisfy the second and third equations but results in an error function having the value $e(\tilde{x}) \approx 13 \times 10^{11}$.

III.3. Computation Noise Considerations

Round off and truncation errors due to finite precision are often of interest. It is possible to choose the matrix R such that the iteration of (2) will give the minimum variance in locating the next iterate value. The variance is a minimum under the assumptions that the variance is due to the finite precision effect and that only first order estimators will be allowed.

For simplicity, this subsection will be written for the special case m = n. We will also assume that J(x) is nonsingular. The extension to the general case of $m \neq n$ and J(x) singular is straightforward but cumbersome. In those cases where interpretations are necessary in order to make the extensions to the general case, we will state the interpretations explicitly.

Suppose we are given a point x^k. Let

$$y^k \triangleq F(x^k) - N^k \quad (113)$$

where $F(x) \triangleq f(x) - b$ and $N^k = [N_1^k, N_2^k, \ldots, N_m^k]$ is the error vector in the computation of $f(x^k)$ by (2) due to finite precision effects. We will assume that $N_1^k, N_2^k, \ldots, N_m^k$ are statistically independent; this is normally the case in practice.

We begin by deriving the consistency constraint for the first order estimator. Let

$$x = x^k + W^k F(x) \quad (114)$$

be an estimate of x^* based on x^k. A Taylor expansion of $F(x)$ shows that (114) may be written as

$$x = x^k + W^k F(x^k) + W^k J(x^k)\ (x - x^k) + o(x - x^k) \tag{115}$$

where $o(x - x^k)$ represents the terms of order two and higher in $||x - x^k||$. Considering the first order term only, (115) becomes

$$x = x^k + W^k F(x^k) + W^k J(x^k)\ (x - x^k). \tag{116}$$

If [5]

$$W^k J(x^k) = I, \tag{117}$$

then (116) may be written as

$$x^k = x^k + W^k F(x^k). \tag{118}$$

Note that (118) is of the form of (114) with $x = x^k$. Thus, for consistency, we require the estimator to satisfy (117) and we call (117) the consistency constraint.

We may now directly consider the problem of estimating x^* based on x^k. By (110), we have, for a first order estimate,

$$x^* = x^k + W^k F(x^*). \tag{119}$$

Expanding $F(x^*)$ in a Taylor's series about x^k and using (113), we find

$$x^* = W^k y^k + W^k J(x^k)\ (x^* - x^k) + W^k N^k + x^k. \tag{120}$$

Thus $\hat{N}^k = W^k N^k$ represents the estimation error term due to the error in evaluating $f(\cdot)$ of (1) at the point x^k. To find the best first order estimate of x^* it is necessary to minimize the covariance matrix of the estimation error.

Let C^k denote the covariance matrix of the estimation noise derived from the evaluation of $f(\cdot)$ of (1). Then C^k is a good measure of the error due to finite precision arithmetics and registers. We have

$$\begin{aligned} C^k &= E[\hat{N}^k (\hat{N}^k)^T] \\ &= E[W^k N\ N^T (W^k)^T] \\ &= W^k\ Q[W^k]^T \end{aligned} \tag{121}$$

where $E(\cdot)$ denotes the expected value and $Q \triangleq E(N\ N^T)$ is the covariance matrix of the computation error in evaluating $f(\cdot)$. Since we assume statistical independence among the N_j^k for $j = 1,2,\ldots,n$, Q is a diagonal matrix. If all registers are of the same precision, then Q may be written as $\alpha\ I$. In any case, Q is symmetric and positive definite.

With this background we may now give the main result of this subsection.

Theorem 5

Let x^* be the solution of (1). Then the iteration equation (2) is the best first order estimator of x^* based on x^k.

Proof: To minimize the error in the estimation we wish to minimize $E(N\ N^T)$ subject to the estimator's consistency, $W^k\ J(x^k) = I$. Using the Lagrangian multiplier method [22], we define a scalar function

$$L(W^k,\Lambda) \triangleq W^k\ Q[W^k]^T - 2\ [W^k\ J(x^k) - I]\Lambda \tag{122}$$

where 2Λ is the Lagrange multiplier matrix. To find a minimum of $E[\hat{N}\ \hat{N}^T]$, we must find Λ and W^k such that

$$\frac{\partial L}{\partial W^k} = 0 \tag{123}$$

and

$$\frac{\partial L}{\partial \Lambda} = 0 \tag{124}$$

where $\left(\frac{\partial L}{\partial W^k}\right)_{ij} \triangleq \left[\frac{\partial L}{\partial W^k_{ij}}\right]$ and $\left[\frac{\partial L}{\partial \Lambda}\right]_{ij} \triangleq \frac{\partial L}{\partial \Lambda_{ij}}$.

Equations (123) and (124) imply that

$$Q[W^k]^T = J(x^k)\Lambda \tag{125}$$

and

$$W^k\ J(x^k) = I. \tag{126}$$

Transposing (125) and noting a property of Q, i.e. Q^{-1} exists, we have

$$W^k = \Lambda^T J^T(x^k)\, Q^{-1}. \tag{127}$$

Postmultiplying (127) by $J(x^k)$ and using (126), (127) becomes

$$\Lambda^T J^T(x^k)\, Q^{-1}\, J(x^k) = I.$$

This implies

$$\Lambda^T = [J^T(x^k)\, Q^{-1}\, J(x^k)]^{-1} \tag{128}$$

In the general case when $m \neq n$ or $J(x^k)$ is singular, then

$$\Lambda^T = [J^T(x^k)\, Q^{-1}\, J(x^k)]^{\dagger} \tag{129}$$

Using (127) and (128), we find

$$\begin{aligned} W^k &= [J^T(x^k)\, Q^{-1}\, J(x^k)]^{\dagger}\, J^T(x^k)\, Q^{-1} \\ &= [J^T(x^k)\, R\, J(x^k)]^{\dagger}\, J^T(x^k)\, R. \end{aligned} \tag{130}$$

where $R \triangleq Q^{-1}$. Note that R is positive definite and symmetric since Q is positive definite and symmetric.

Let x^{k+1} be the first order estimate of x^*. By (114), the first order estimate in the mean of x^* is given by

$$x^{k+1} = W^k y^k + x^k = W^k [f(x^k) - b] + x^k. \tag{131}$$

Using (130), (131) becomes

$$x^{k+1} = [J^T(x^k)\, R\, J(x^k)]^{\dagger}\, J^T(x^k)\, R\, [f(x^k) - b] + x^k. \tag{132}$$

Note that (132) is precisely our iteration equation (2). Thus, if R in (2) is chosen to be Q^{-1}, the inverse of the covariance matrix of computation error in evaluating $f(\cdot)$, then (2) is actually the best first order estimation of x^* based on x^k.

Q.E.D.

IV. CONCLUDING REMARKS

In this paper we have introduced an iteration equation which may be used to solve a system of m nonlinear equations in n unknowns. The iteration equation presented has a great deal of flexibility due to the utilization of the R matrix. We have shown that the R matrix may be chosen: (1) to improve the conditioning of an ill-conditioned matrix; (2) to assign weights among equations in a system of equations; and (3) in the presence of the usual computation noise, to give the minimum variance in the next iterate value. We have also presented two algorithms which serve as computational procedures for the iteration equation. In addition, a modified iteration equation has been introduced. This equation has the advantage of requiring fewer calculations per iteration. An algorithm to implement this modified equation has also been presented. Furthermore, a theorem giving sufficient conditions to insure the convergence of the modified iteration equation has been established.

We note that each step of the proposed iteration of (2) will either decrease the value of $e(x^k)$ or reach the termination state, i.e. $\nabla e(x^k) = 0$. Since $\nabla e(x^k) = 0$ only gives a local minimum there is no way that we can be sure that $e(x^k)$ is the global minimum and thus, the best possible solution.

V. REFERENCES

1. C. G. Broyden, A class of methods for solving nonlinear simultaneous equations, Math. Comp., 19 (1965), 577-593.

2. F. J. Zeleznik, Quasi-Newton methods for nonlinear equations, J. Asso. Comp. Mach., 15 (1968), 265-271.

3. J. E. Dennis, Jr., Toward a unified convergence theory for Newton-like methods, in Nonlinear Functional Analysis and Applications, edited by L. B. Rall, Academic Press, New York, 1971.

4. E. H. Moore, Generalized analysis, Part I, Mem. Amer. Phil. Soc. 1, (1935), 197-209.

5. R. Penrose, A generalized inverse for matrices, Proc. Cambridge Phil. Soc. 51 (1955), 406-413.

6. J. Diendonne, Foundation of Modern Analysis, Academic Press, New York, 1960.

7. T. O. Lewis and T. G. Newman, Pseudoinverse of positive semidefinite matrices, SIAM J. Appl. Math. 16, No. 4 (1968), 701-703.

8. C. A. Rohde, Generalized inverses of partitioned matrices, SIAM J. Appl. Math. 13, No. 4 (1965), 1033-1035.

9. B. Noble, A method for computing the generalized inverse of a matrix, SIAM J. Num. Anal. 3, No. 4 (1966), 582-584.

10. T. Greville, Some applications of the pseudoinverse of a matrix, SIAM Rev. 2 (1960), 15-22.

11. L. O. Chua, Efficient computer algorithms for piecewise-linear analysis of resistive nonlinear networks, IEEE Trans. CT. CT-18, No. 1 (1971), 73-85.

12. T. Fujisawa, E. S. Kuh and T. Ohtsaki, A sparse matrix method for analysis of piecewise-linear resistive networks, IEEE Trans. CT. CT-19, No. 6 (1972), 571-585.

13. S. Lang, Analysis II, Addison-Wesley, Reading, Mass., 1969.

14. L. O. Chua and Y.-F. Lam, Foundations of nonlinear n-port network theory, Purdue University, Lafayette, Indiana, Tech. Rep. TR-EE 71-20, June, 1971.

15. K. Eisemann, A heuristic description of generalized matrix inversion, IEEE Trans. CT. CT-20, No. 5 (1973), 481-487.

16. R. Penrose, On best approximate solution of linear matrix equation, Proc. Cambridge Phil. Soc. 52, (1956), 17-19.

17. R. S. Palais, Natural operators on differential forms, Trans. Amer. Math. Soc. 92 (1959), 125-141.

18. L. O. Chua and Y.-F. Lam, Global homeomorphisms of vector-valued functions, J. Math. Anal. Appl. 39, No. 3 (1972), 600-624.

19. C. D. Meyer and R. J. Painter, Note on a least squares inverse for a matrix, J. Assn. Comp. Mach. 17, No. 1 (1970), 110-112.

20. A. Ralston, A First Course in Numerical Analysis, McGraw-Hill, New York, 1965.

21. N. Morrison, Introduction to Sequential Smoothing and Prediction, McGraw-Hill, New York, 1969.

22. D. J. Wilde and C. S. Beightler, Foundations of Optimization, Prentice-Hall, New York, 1967.

Footnotes

1. $J(x)$ being of rank n implies $n \leq m$.

2. The proof of this lemma appears elsewhere; see, for example, [14].

3. If p^k is desired rather than $[J^T(x^k)\ R\ J(x^k)]^{\dagger}$, an efficient way to compute p^k from (65) is given by [19]. However, if $[J^T(x^k)\ R\ J(x^k)]^{\dagger}$ is also required, then the updating equation (16) may be used. If (16) is efficiently programmed then a straightforward computation of (64) is also desirable. This is particularly attractive in piecewise-linear circuits (which may contain controlled sources since $J(x)$ is not required to be symmetric) [11,12] where C in (16) is 1 x n.

4. For a more complete description, see [20].

5. Equation (117) is called the exactness constraint in estimation theory [21].