A FIRST ORDER ITERATION PROCESS FOR SIMULTANEOUS EQUATIONS*

H. A. Luther and W. F. Stewart†

Introduction

The judgment of an iterative solution of simultaneous equations properly entails more than the rate of convergence. An item of concern is whether or not we can arrive at a solution point whose approximate location is known.

Let the "region of influence" of a given solution point be the set of all starting values which (for a given iterative technique) have that solution point as the limiting value of the iteration. Ideally, perhaps, we might desire that every starting point yield some solution and that every region of influence be simply connected.

It is known that, for example, the Newton-Raphson technique is a second order process with favorable rate of convergence when well-started. It is also known that the regions of influence are not normally simply-connected.

The present procedure was devised as part of a program for studying methods of altering the regions of influence rather than rates of convergence.

The equations considered can be linear or nonlinear. By using

*This research was supported in part by the National Aeronautics and Space Administration under Research Grant No. NGR-44-001-024.

†Texas A&M University, College Station, Texas.

first order partial derivatives, there is created a technique which has the Newton-Raphson quality of converging to every solution point whose Jacobian is nonvanishing. The method is computationally simpler than the Newton-Raphson method; it is, however, first order rather than second order. An arbitrary function can be used parametrically.

The study is far from complete. The present approach seems faulty in that in practice the convergence rate is too slow. Because of this a method of accelerating convergence is included.

<u>Description of the Process</u>

Let

$$(1) \qquad f_i(x_1, x_2, \cdots x_n), \qquad 1 \le i \le n,$$

be real functions defined for an open region R. Let $X = [x_1\ x_2\ \cdots\ x_n]^t$ denote a column matrix, and write $f_i(x_1, x_2, \cdots x_n) = f_i(X)$. Let $Y = [y_1\ y_2\ \cdots\ y_n]^t$ be a point of R such that the following conditions are satisfied:

a. For $1 \le i \le n$, $f_i(Y) = 0$;

b. In a neighborhood of Y the partial derivatives $\frac{\partial f_i}{\partial x_j}$ are continuous $(1 \le i, j \le n)$;

c. For $f_{ij}(X) = \frac{\partial}{\partial x_j} f_i(X)$, the n x n matrix $J(X) = [f_{ij}(X)]$ is nonsingular at Y.

Let the eigenvalues of $\{1/Sp[J^t(Y)\ J(Y)]\}J^t(Y)\ J(Y)$ be $\lambda_1, \lambda_2, \cdots \lambda_n$. They are positive, and $\sum_{i=1}^{n} \lambda_i = 1$. Let them be so ordered that $\lambda_i \geq \lambda_{i+1}$. Let d(X) be a function over R arbitrary except that it is continuous in a neighborhood of Y and except that

$$0 < d(Y) < 2/\lambda_1 \ . \tag{2}$$

For $1 \leq i \leq n$ define functions $h_i(X)$ by

$$h_i(x) = \{d(X) \sum_{\ell=1}^{n} f_\ell(X) \; f_{\ell i}(X)\}/Sp\left(J^t(X) \; J(X)\right) \ . \tag{3}$$

Let $H(X) = [h_1(X) \; h_2(X) \cdots h_n(X)]^t$. Let $X^{(k)} =$ $[x_1{}^{(k)} \; x_2{}^{(k)} \ldots x_n{}^{(k)}]^t$.

The basic theorem can now be phrased as follows:

If $X^{(1)}$ is "near enough" Y and

$$X^{(k+1)} = X^{(k)} - H\left(X^{(k)}\right) \tag{4}$$

then $\lim_{k\to\infty} X^{(k)} = Y$.

Finding suitable functions d(X) is not elaborate. Since $\lambda_1 < 1$ unless n = 1, we may for d(X) use any positive constant not greater than two. In the final stages of iteration $d(X)/Sp\left(J^t(X) \; J(X)\right)$ can be replaced by a constant. Indeed, for some systems it may be practical to choose d(X) as $2 \; Sp\left(J^t(X) \; J(X)\right)/Max_R\left(Sp(J^t(X) \; J(X))\right)$ from the start, thus simplifying computation.

A cautionary note seems in order. There can exist points Y, resulting from the limit process (4), such that J(Y) is singular. In that case it may not be that $f_i(Y) = 0$ for all i. Consider the example $f_1(x_1, x_2) = x_1 + x_2 - 1$, $f_2(x_1, x_2) = x_1 + x_2 - 2$. Then

$h_1 = h_2$ and, for $d(X) = 1$, is $x_1/2 + x_2/2 - 3/4$. For the starting values $x_1^{(1)} = \alpha$ and $x_2^{(1)} = \beta$, we find $x_1^{(k)} = \alpha/2 - \beta/2 + 3/4$ and $x_2^{(k)} = -\alpha/2 + \beta/2 + 3/4$ for $k \geq 2$.

Linear Systems

We wish first to make an observation applicable to the general case. By Taylor's expansion

$$f_\ell(X) = \sum_{j=1}^{n} f_{\ell j}(\xi_\ell)(x_j - y_j)$$

where $\xi_\ell = [(y_1 + \theta\overline{x_1 - y_1}) \cdots (y_n + \theta\overline{x_n - y_n})]^t$. Now let

$$\text{(5a)} \qquad g_{ij}(X) = \{d(X)/Sp\big(J^t(X)\ J(X)\big)\} \sum_{\ell=1}^{n} f_{\ell i}(X)\ f_{\ell j}(\xi_\ell)$$

and let

$$\text{(5b)} \qquad G(X) = [g_{ij}(X)], \qquad 1 \leq i,\ j \leq n\ .$$

Then clearly $\big($see (4)$\big)$

$$\text{(6)} \qquad X - Y - H(X) = \big(I - G(X)\big)(X - Y)\ .$$

The only use we want of the above in the present section is to show that the eigenvalues μ_i of $I - G(Y)$ are given by $\mu_i = 1 - d(Y)\lambda_i$, while if (2) holds

$$\text{(7)} \qquad 1 > \mu_n, \quad \mu_{i+1} \geq \mu_i, \quad \mu_1 > -1\ .$$

To see this, it is only necessary to observe that

$$G(Y) = \{d(Y)/Sp\{J^t(Y)\ J(Y)\}\}J^t(Y)\ J(Y) \tag{8}$$

and that $d(Y) > 0$ together with $\lambda_i \geq \lambda_{i+1} > 0$ guarantee $\mu_{i+1} \geq \mu_i$ while $d(Y) < 2/\lambda_1$ means $\mu_1 > -1$ and $1 > \mu_n$.

Now for the linear case let

$$f_i(X) = \sum_{j=1}^{n} a_{ij}\ x_j - b_i \tag{9}$$

and, for convenience only, consider $d(X) \equiv \delta$, δ being chosen so that it falls between zero and $2/\lambda_1$. Solution of (9) is of course solution of

$$AY = B \tag{10}$$

where $A = [a_{ij}]$ and $B = [b_1\ b_2\ \cdots\ b_n]^t$. Since $\frac{\partial}{\partial x_i} f_\ell(X) = a_{\ell i}$, it is a straightforward matter to see that

$$X - H(X) = \left[I - \{\delta/Sp(A^tA)\}A^tA\right]X + \{d/Sp(A^tA)\}A^tB. \tag{11}$$

Thus convergence of process (4) is guaranteed for the linear case (see [1], pp. 161-170) since the eigenvalues of the coefficient of X in the right member of (11) are less in magnitude than 1.

It is true that if, say, $\delta = 1$, and $\delta/Sp(A^tA)$ is used as a constant multiplier, then even relatively large inaccuracies in the computation of $Sp(A^tA)$ have no influence on the theoretical accuracy of the iteration. However, it is also known that A^tA is more ill-conditioned than A. This should be borne in mind in using the process.

The Case n = 1. For $n = 1$, the process reduces to the solution of $f(y) = 0$ by the technique

$$x^{(k+1)} = x^{(k)} - d\left(x^{(k)}\right)f\left(x^{(k)}\right)/f'\left(x^{(k)}\right) \tag{12}$$

where $d(y) = \delta$ is a positive number less than two. For the choices $d(x) = 1$, $d(x) = [f'(x)]^2/\{[f'(x)]^2 - f(x)f''(x)\}$ and related ideas see [2], p. 24 et seq.

To establish convergence for a $d(x)$ chosen in accordance with the present discussion, let R_1 be an interval in R such that, $R_1, d(x)\ f'(y + \theta\overline{x - y})/f'(x) = \delta + \varepsilon(x)$ where θ is such that $f'(y + \theta\overline{x - y})(x - y)$ is $f(x) - f(y)$ and where $|\varepsilon(x)| < \mu - |1 - \delta|$ while $0 < \mu < 1$. Let $x^{(1)}$ be in R_1. Then $|x^{(k+1)} - y|$ is $|x^{(k)} - y||1 - \delta - \varepsilon(x^{(k)})|$ which in turn does not exceed $\mu|x^{(k)} - y|$. Thus $|x^{(k+1)} - y| \leq \mu^k|x^{(1)} - y|$ and convergence occurs.

Now suppose $d(y) = 1$. Then convergence occurs as before, but following a known technique (see [3], p. 448), can be shown to be quadratic. Thus let $d'(x)$ and $f''(x)$ be continuous in R_1. Then, expanding $f(y)$ in powers of $y - x$, we are led to $x^{(k+1)} - y = x^{(k)} - y - d\left(x^{(k)}\right)\{f'\left(x^{(k)}\right)\left(x^{(k)} - y\right) - f''\left(\eta^{(k)}\right)\left(x^{(k)} - y\right)^2/2\}/f'\left(x^{(k)}\right)$. Then if $1 - d\left(x^{(k)}\right) = d(y) - d\left(x^{(k)}\right) = d'\left(\xi^{(k)}\right)\left(y - x^{(k)}\right)$, $x^{(k+1)} - y = \left(x^{(k)} - y\right)^2\{d\left(x^{(k)}\right)f''\left(\eta^{(k)}\right) - 2d'\left(\xi^{(k)}\right)f'\left(x^{(k)}\right)\}/\{2f'\left(x^{(k)}\right)\}$.

The Case n = 2. In this instance the characteristic equation for $I - G(Y)$ is

$$\begin{gathered} \mu^2 - \mu(2 - \delta) + 1 - \delta + \delta^2 \omega^2 = 0 \\ \omega^2 = \det\left(J^t(Y)\ J(Y)\right)/\{\mathrm{Sp}\left(J^t(Y)\ J(Y)\right)\}^2,\ \delta = d(Y)\ . \end{gathered} \tag{13}$$

Following the procedure in [3], p. 450, we have asymptotic stability if and only if

$$| 2 - \delta | \leq 2 - \delta + \delta^2 \omega^2 < 2\ . \tag{14}$$

Write $\det\left(J^t(Y)\ J(Y)\right)$ as $a_{11}\ a_{22} - a_{12}{}^2$. Then a_{11} and a_{22} are positive. It follows that

$$0 < (a_{11}\ a_{22} - a_{12}{}^2)/(a_{11} + a_{22})^2 \leq a_{11}\ a_{22}/(a_{11} + a_{22})^2 \leq 1/4$$

and equality can only occur between the last two members if $a_{11} = a_{22}$. This means $\omega^2 \leq 1/4$ and $\omega^2 = 1/4$ if and only if $J(Y)$ is a scalar multiple of a unitary matrix.

Reverting to (14) it is seen that $0 < \delta < 1/\omega^2$ is necessary. Since $4 - 2\delta + \delta^2 \omega^2 \geq 0$ unless δ is between $(1 - \sqrt{1 - 4\omega^2})/\omega^2$ and $(1 + \sqrt{1 - 4\omega^2})/\omega^2$ (while this last equals or exceeds $1/\omega^2$) we have asymptotic stability if and only if $0 < \delta < (1 - \sqrt{1 - 4\omega^2})/\omega^2$. When $\delta = (1 - \sqrt{1 - 4\omega^2})/(2\omega^2)$, $I - G(Y)$ is singular. Since $(1 + \sqrt{1 - 4\omega^2})/(2\omega^2) \leq (1 - \sqrt{1 - 4\omega^2})/\omega^2$ if and only if $\omega^2 \geq 2/9$, when $\omega^2 \geq 2/9$ and $\delta = (1 + \sqrt{1 - 4\omega^2})/(2\omega^2)$, $I - G(Y)$ is again singular.

It will be shown in the next section that when $\omega^2 = 1/4$ and $\delta = 2$ we have quadratic convergence. Since $4 \leq 1/\omega^2$ always, it may be that $d(X) \equiv 2$ is in some sense a "best" choice if no detailed knowledge of the functions is at hand.

Proof of Convergence.

Consider now (see (4))

$$X^{(k+1)} - Y = X^{(k)} - Y - H(X^{(k)}) \ .$$

Using (6) this becomes

$$X^{(k+1)} - Y = (I - G(X^{(k)}))(X^{(k)} - Y) \ .$$

Let R_1 be a spherical region in R with Y as center, and small enough that the functions $f_{ij}(X)$ and $d(X)$ are continuous in R_1. As in establishing (7), let the eigenvalues of $I - G(Y)$ be μ_i, and let ε be such that $1 - \varepsilon = \max_i(\mu_i^2)$. The numbers μ_i^2 are the eigenvalues of $(1 - G(Y))^t(I - G(Y))$, and $1 - \varepsilon \geq \mu_i^2$. Let the eigenvalues of $(I - G(X))^t(I - G(X))$ be $\nu_i(X)$.

Let the characteristic polynomial of the μ_i^2 be $\sum_{i=0}^{n} a_i\mu^{2i}$ and that of the $\nu_i(X)$ be $\sum_{i=0}^{n} b_i \ \nu^i$. Because $|b_i - a_i|$ can be made as small as desired by using a proper spherical subregion of R_1, there is in R_1 a spherical subregion R_2 having Y as its center and such that, for X in R_2, $|\nu_i - \mu_i^2| < \varepsilon/2$. In particular, $\max(\nu_i) < 1 - \varepsilon/2$. By a well-known theorem (see [1], p. 65)

$(X^{(k)} - Y)^t(I - G(X)^{(k)})^t(I - G(X^{(k)}))(X^{(k)} - Y) \leq$

$\max_i(\nu_i)(X^{(k)} - Y)^t(X^{(k)} - Y) \leq (1 - \varepsilon/2)(X^{(k)} - Y)^t(X^{(k)} - Y)$.

Then $(X^{(k+1)} - Y)^t(X^{(k+1)} - Y) \leq (1 - \varepsilon/2)(X^{(k)} - Y)^t(X^{(k)} - Y)$, and by recursion

$$||X^{(k+1)} - Y|| \leq (1 - \varepsilon/2)^{k/2}||X^{(1)} - Y||$$

Thus convergence occurs. Here $||X^{(k+1)} - Y||$ denotes the distance norm.

Finally it is shown that under stringent conditions the convergence is quadratic. Let J(Y) be a scalar multiple of an orthogonal matrix. Assume in addition that the first order derivatives of d(X) and the second order derivatives of $f_\ell(X)$, $1 \leq \ell \leq n$, are continuous in a neighborhood of Y. Require also that $\delta = n$. Since in this instance all the eigenvalues λ_i are equal, we have $d(Y) = \delta = n < 2/\lambda_1 = 2n$. Thus convergence occurs and $X^{(k)} - H(X^{(k)})$[see (4)] has meaning.

Write $d_i(X)$ for $\frac{\partial}{\partial x_i} d(X)$ and $f_{\ell ij}(X)$ for $\frac{\partial^2}{\partial x_i \partial x_j} f_\ell(X)$. Then

$$d(X) = \delta + \sum_{r=1}^{n} d_r(\xi)(x_r - y_r)$$

and, since $f_\ell(Y) = 0$

$$0 = f_\ell(X) + \sum_{r=1}^{n} f_{\ell r}(X)(y_r - x_r) + 1/2 \sum_{r,s=1}^{n} f_{\ell rs}(\eta)(x_r - y_r)(x_s - y_s).$$

First replace the functions $f_\ell(X)$ in X - Y - H(X) by their equivalents above. The result, after dropping the quadratic terms, is $X - Y - \{d(X)/Sp[J^t(X)\ J(X)]\}J^t(X)\ J(X)\ (X-Y)$. Now replace d(X) by $n + \sum_{r=1}^{n} d_r\ (\xi)(x_r - y_r)$. The result, after dropping the quadratic terms, is

$$X - Y - \{n/Sp\big(J^t(X)\ J(X)\big)\}J^t(X)\ J(X)\ (X - Y) \ .$$

Next in $J^t(X)\ J(X)$ replace, in the off diagonal terms, $f_{rs}(X)$ by $f_{rs}(Y) + \sum_{t=1}^{n} f_{rst}(\eta_{rs})(x_t - y_t)$. Because J(Y) is a scalar multiple of an orthogonal matrix, there result only quadratic and cubic terms in $(x_i - y_i)$ for these components. There remains to consider

(15) $$X - Y - \{1/Sp\big(J^t(X)\ J(X)\big)\}C(X)(X-Y)$$

where C(X) is a diagonal matrix whose diagonal entry $c_i(X)$ is

$$n \sum_{\ell=1}^{n} f_{\ell i}{}^2(X) \ .$$

Now rephrase (15) as $\{1/Sp\big(J^t(X)\ J(X)\big)\}\big(E(X) - C(X)\big)(X - Y)$ where E(X) is a diagonal matrix whose diagonal entry $e_i(X)$ is $SP\big(J^t(X)\ J(X)\big)$. In E(X) - C(X) replace each function $f_{ij}(X)$ by the Taylor's expansions used in arriving at (15). The result is a diagonal matrix which is made up of linear and quadratic terms in the $x_i - y_i$. This is so because J(Y) is a scalar multiple of an orthogonal matrix.

In final consequence, X - Y - H(X) is seen to contain only terms which are cubic or quadratic in the $x_i - y_i$.

Improvement in Asymptotic Rate of Convergence

It was found in practice that process (4) converged slowly. Thus it was modified in an attempt to improve the asymptotic rate of convergence.

We start by defining

$$\bar{G}(X) \equiv \frac{d(X)}{Sp(J^t(X)\ J(X))}\ J^t(X)\ J(X)$$

where J(X) and Sp again denote the Jacobian matrix and spur respectively. The new process may now be stated as

$$(16) \qquad X^{(k+1)} = X^{(k)} - 2H(X^{(k)}) + \bar{G}(X^{(k)})\ H(X^{(k)}) \ .$$

Then clearly

$$(17) \quad X - Y - 2H(X) + \bar{G}(X)\ H(X) = \left[I - 2G(X) + \bar{G}(X)\ G(X)\right]\ (X - Y)$$

where G(X) is defined by (5).

For the linear case, $\bar{G}(X) = G(X)$ and (17) becomes

$$(18) \qquad X - Y - 2H(X) + \bar{G}(X)\ H(X) = \left[I - G(X)\right]^2\ (X - Y) \ .$$

As before let $d(X) \equiv \delta$ where δ is constant and consider a solution of (10). We have from (18)

$$(19) \qquad X - Y - 2H(X) + G(X)\ H(X) = (I - \frac{\delta}{Sp(A^tA)}\ A^tA)^2\ X$$

$$+ (2I - \frac{\delta}{Sp(A^tA)}\ A^tA)\ \frac{\delta}{Sp(A^tA)}\ A^tB$$

Define M to be $(I - \frac{\delta}{Sp(A^tA)}\ A^tA)$ which is the coefficient of X in the right member of (11). Then M^2 is the coefficient of X in the right member of (19). If the eigenvalues of M are μ_i, the eigenvalues of

M^2 are μ_i^2. Since $|\mu_i| < 1$ (see (7)), then $\mu_i^2 < 1$ and we are guaranteed that process (16) converges for the linear case.

The asymptotic rate of convergence of M (see [4], p. 67) is defined as

$$R\infty\ (M) = -\ln\ (\rho(M)) \tag{20}$$

where $\rho(M)$ is the magnitude of the largest eigenvalue of M. Denote $\rho(M)$ by μ, then $\rho(M^2)$ is μ^2 and $R\infty\ (M^2) = 2\ R\infty\ (M)$. Thus the asymptotic rate of convergence of process (16) is twice that of process (4) for the linear case.

For the non-linear case, we consider (6) and (17). Let β_i and γ_i denote the eigenvalues of $I - G(X)$ and $I - 2G(X) + \overline{G}(X)\ G(X)$ respectively and as before let μ_i denote the eigenvalues of $I - G(Y)$. Let each eigenvalue be so indexed that β_i is approximated by μ_i and γ_i by μ_i^2.

Now let $\sum a_i\mu^i$, $\sum b_i\beta^i$ and $\sum c_i\gamma^i$ represent the characteristic polynomials whose zeros are μ_i, β_i and γ_i respectively. The coefficient b_i is a continuous function of the elements of G(X) and $|b_i - a_i|$ can be made as small as desired by choosing a proper subregion R_3 in R. There is a subregion in R_3 having Y as its center such that $|\beta_i - \mu_i| < \varepsilon$ provided X is in this region.

In a similar manner it can be shown that $|\gamma_i - \mu_i^2| < \varepsilon$ by similarly choosing a proper subregion of R_3.

Now β_i and γ_i can be made arbitrarily close to μ_i and μ_i^2 respectively by choosing the proper subregion of R. With the condition that $\mu_i^2 < \mu_i$ it follows that the asymptotic rate of convergence for process (16) is double that of process (4).

References

1. Bodewig, E. Matrix Calculus. North-Holland Publishing Co., Amsterdam, 1959.

2. Traub, J. F. Iterative Methods for the Solution of Equations. Prentice-Hall, Inc., Englewood Cliffs, 1964.

3. Hildebrand, F. B. Introduction to Numerical Analysis. McGraw-Hill Book Co., Inc., New York, 1956.

4. Varga, R. S. Matrix Iterative Analysis. Prentice-Hall, Inc., Englewood Cliffs, 1962.