A CONTINUOUS ANALOGUE ANALYSIS OF NONLINEAR ITERATIVE METHODS

Paul T. Boggs* and J.E. Dennis, Jr.†

TR 74-200

March 1974

(Revised June 1975)

Department of Computer Science
Cornell University
Ithaca, New York 14850

*Department of Mathematics
Rensselaer Polytechnic Institute
Troy, New York 12181

† Research supported by National Science Foundation grant number GJ-40903.

This work is respectfully
dedicated to the memory
of
Kenneth M. Levenberg

A CONTINUOUS ANALOGUE ANALYSIS OF NONLINEAR ITERATIVE METHODS

Paul T. Boggs* and J.E. Dennis, Jr.**

Department of Mathematics
Rensselaer Polytechnic Institute
Troy, New York 12181
and
Department of Computer Science
Cornell University
Ithaca, New York 14850

Abstract:

This paper applies the asymptotic stability theory for ordinary differential equations to Gavurin's continuous analogue of several well-known nonlinear iterative methods. In particular, a general theory is developed which extends the Ortega-Rheinboldt concept of consistency to include the widely used finite difference approximations to the gradient as well as the finite difference approximations to the Jacobian in Newton's method. The theory is also shown to be applicable to the Levenberg-Marquardt and finite difference Levenberg-Marquardt methods.

---

*Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12181.

**Research supported by National Science Foundation Grant number GJ-40903.

A CONTINUOUS ANALOGUE ANALYSIS

OF NONLINEAR ITERATIVE METHODS

Paul T. Boggs and J.E. Dennis, Jr.

1. INTRODUCTION

Many practical problems in applied science give rise to one of the following finite dimensional problems.

A. Given $f:R^n \to R^1$, find $x^* \in R^n$ at which f achieves a minimum.

B. Given $F:R^n \to R^n$, find $x^* \in R^n$ for which $F(x^*) = 0$.

C. Given $F:R^n \to R^m$, $m \geq n$, find $x^* \in R^n$ at which $||F(x)||_2$ achieves a minimum.

Usually an acceptable approximation to x^* is sought by means of an iterative process of the form

$$x_{k+1} = x_k - t_k G(x_k), \quad k = 0,1,... \tag{1.1}$$

where x_0 is a given initial approximation, G is a vector valued function such that $G(x_k)$ indicates the direction to be taken by the iteration from x_k and t_k is a scalar which determines the magnitude of the step.

For example in problem A, $G(x_k)$ might be taken to be $\nabla f(x_k)$ and $t_k > 0$ might be chosen to minimize $f(x_k - tG(x_k))$ with respect to t in some interval $(0,\bar{t}]$. This is, of course, the steepest descent method. Newton's method for problem B takes $G(x_k) = F'(x_k)^{-1}F(x_k)$. In this case and in the one below, t_k is initially chosen to prevent divergence and eventually taken to be unity in order to give the maximum rate of convergence. In problem

C, one might use the Levenberg-Marquardt method which takes $G(x_k) = [\mu_k I + F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k)$ where $\mu_k \geq 0$. (If $\mu_k \equiv 0$ then the Gauss-Newton method results.) Notice that in all these examples the point x^* is a zero of G.

These basic methods all have in common a need for first derivative information. Very often such information is either not available at all or else extremely expensive. In such cases it has long been standard procedure to use the basic algorithms with difference quotients in place of derivatives. Thus, if one needs

$$(1.2) \qquad \partial_j f(x) = \frac{\partial f(x)}{\partial x_j} \quad \text{for } f: R^n \to R^1,$$

one might use in its place

$$(1.3) \qquad \delta_j f(x,h) = \frac{f(x+h_j u_j)-f(x)}{h_j}, \quad h_j \neq 0, \; j=1,\ldots,n$$

where $u_j = (\delta_{1j},\ldots,\delta_{nj})^T$ and δ_{ij} is the Kronecker δ-function. It is also possible to use a central difference quotient in place of (1.3) and we will remark on this at appropriate places in the sequel.

Theoretical and computational results surveyed in [6] indicate that if one chooses h properly, then the finite difference forms of the Newton and Levenberg-Marquardt algorithms can perform very well indeed.

In particular, it is shown that h should be chosen as a suitable multiple of $||F(x)||$ or $||F'(x)^T F(x)||$ respectively. Thus the successive values of h decrease and this in turn raises the real possibility of cancellation errors in the finite precision computation of $\delta_j f(x,h)$ which would swamp the small truncation error, $\partial_j f(x)-\delta_j f(x,h)$, associated with h.

The use of a central difference only postpones the problem for one or two steps and so it seems clear that any reasonable implementation of these ideas must admit the possibility that h is eventually held fixed in the iteration.

In using Newton's method as described above, this does not affect convergence to x*, but does, at least theoretically, slow to linear the rate of convergence. In using steepest descent however, ∇f will not in general have the same roots as its finite difference approximation which implies that convergence to x* is no longer guaranteed. The purpose of this paper, therefore, is to compute a bound between x* and the limit of a sequence $\{x_k\}$ generated by using such an iteration. The bound obtained is a natural function of h and the conditioning of the problem.

The nonlinear least squares problem is of special interest because it illustrates another way in which a parameter can enter the iteration. The Levenberg-Marquardt method [13], [14], results from taking $G(x_k) = (\mu_k I+F'(x_k)^T F'(x_k))^{-1}F'(x_k)^T F(x_k)$ where μ_k is a nonnegative scalar. If $F'(x_k)$ has less than full rank, then the implementation must either allow μ_k to eventually be held constant at some positive value or else switch to the more costly and less understood Ben-Israel iteration,

$G(x) = F'(x)^{+}F(x)$. (Here A^{+} denotes the Moore-Penrose inverse [19] of the rectangular matrix A.)

Our approach to these questions is somewhat unusual and so we outline it here. Following Gavurin [8], we define the continuous analogue to (1.1). This is the initial value problem

$$x'(s) = -G(x) \ , \quad 0 \leq s < \infty \ , \ x(0) = x_0 \ . \tag{1.4}$$

Notice that if Euler's method is applied to (1.4), then our original iteration (1.1) is recovered. The stability theory for ordinary differential equations identifies conditions on G and x_0 that imply the existence of a solution curve $x(s)$ such that $\lim_{s\to\infty} x(s) = x^*$. This provides quantative information on the directions to be used. Although the connection is well known it usually fails to say anything new about the convergence of the iterations.

On the other hand, we are not so much interested in iterations of the form (1.1) as we are in those of the form

$$x_{k+1} = x_k - t_k \tilde{G}(x_k, p) \tag{1.5}$$

where p is some parameter vector. The corresponding continuous analogue is

$$x'(s) = -\tilde{G}(x,p) \ , \quad 0 \leq s < \infty \ , \quad x(0) = x_0 \ . \tag{1.6}$$

We write (1.6) as

$$x'(s) = -G(x) + [G(x) - \tilde{G}(x,p)]$$

and view (1.6) as a perturbation, controlled by p, to (1.4). In this case, the stability theory again provides qualitative information on solutions $x_p(s)$ to (1.6). Moreover, it also provides a bound on $\overline{\lim}_{s\to\infty} ||x_p(s) - x^*||$ in terms of p, G and x_0. The application of Euler's method to (1.6) now yields a similar bound on $\lim_{k\to\infty} ||x_k-x^*||$ where x_k is a solution to (1.5).

Clearly, some conditions on $||G(x)-\mathcal{G}(x,p)||$ are necessary for such a result. In Section 2, we generalize the Ortega-Rheinboldt [17] idea of a consistent Jacobian approximation to provide a sufficient condition and we also show that the specific instances of (1.5) mentioned earlier satisfy our condition. Sections 3 and 4 are devoted to the minimal amount of stability theory for (1.6) and (1.5) necessary for the results on specific methods which we present in Section 5.

It is perhaps worth noting at this point that other authors, including Meyer [15] and Bosarge [3], have used the continuous analogue -- numerical integration connection, but they work with a finite range of the independent variable. The advantages of our approach are set forth in Boggs [1] and provide motivation for our choice of an infinite range.

An interesting different approach to the analysis of iterative methods by differential equation techniques was given by Hurt [12]. In his paper, discrete analogues of Liapunov functions are defined and used to prove certain stability results. These results are then used, for example, to obtain regions of convergence for Newton's method and to analyze the effects of round-off error. However, the resulting bounds depend on the Liapunov functions used and thus sharp bounds are quite difficult to obtain. Also, for some iterative methods, finding any Liapunov function may require considerable ingenuity. Nevertheless, this approach does seem valuable and has been used by Boggs [2] to analyze certain algorithms in the presence of singularities. (See also Ortega [16] for a bibliography and an exposition of the basic results).

2. CONSISTENT APPROXIMATIONS

In order to present a unified theory for the various multidimensional secant methods as well as the modification of Newton's method in which the Jacobian matrix is replaced by the corresponding matrix of difference quotients, Ortega and Rheinboldt [17, p. 355] emply a very elegant formalism called a "consistent approximation" to the Jacobian. A generalization of this concept to other than Jacobians will be useful to us.

We use $\bar{P}$ to denote the closure of a set P.

Definition 2.1. Let $D \subset R^n$, $P \subset R^q$ with $0 \in \bar{P}$. Let G be a mapping of D into R^m. A function $\tilde{G}$ of the vector variables (x,p) defined on $D \times P$ is a consistent approximation rule for G on D iff

$$(2.1) \qquad ||G(x) - \tilde{G}(x,p)|| \to 0 \quad \text{as} \quad ||p|| \to 0 ,$$

uniformly on every compact subset of D. In addition, if there are positive constants α, β and ε such that for every $x \in D$ and $||p|| < \varepsilon$

$$(2.2) \qquad ||G(x) - \tilde{G}(x,p)|| \leq \beta ||p||^{\alpha} ,$$

then $\tilde{G}$ is a strongly consistent approximation rule of order α for G on D.

Clearly the property of consistency and the value of α are norm independent while β is not.

We now proceed to show that the iterations mentioned in the introduction are connected by this concept. In the lemma below we will assume that the functional f, defined from an open convex set $D \subset R^n$ into the real numbers, is continuously differentiable on D. If $x \in D$ then (1.2) is defined, but we must restrict x and h if (1.3) is to be defined. Choose $D_0 \subset D$ such that $\bar{D}_0 \subset D$ and choose $\varepsilon > 0$ such that $D \supset N(D_0, \varepsilon) \equiv \{x \in R^n: ||x-y|| < \varepsilon$ for some $y \in D_0\}$. Now for $(x,h) \in D_0 \times (0,\varepsilon]^n$, $\delta_j f(x,h)$ is well-defined by (1.3) for $1 \leq j \leq n$.

It is sometimes the case that the analyst might choose to compute some partial derivatives of f and approximate others. Hence, we define our approximation to the gradient as follows. For $(x,h) \in D_0 \times [0,\varepsilon]^n$ take $\Delta f(x,h)$ to be a column vector where for $i=1,\ldots,n$,

$$\Delta f(x,h)_i = \begin{cases} \delta_i f(x,h) & , h_i \neq 0 \\ \partial_i f(x) & , h_i = 0 \end{cases} \tag{2.3}$$

It is obvious that we intend $\Delta f(x,h)$ as an approximation to the derivative of f at x. Since this derivative is, strictly speaking, a linear functional and so more naturally represented as a row vector, while $\Delta f(x,h)$ is a column vector, we invoke the Riesz Representation Theorem once here in order to avoid frequent use of superscript T, and make the notational convention that $\nabla f(x)$ is the column vector which represents

$f'(x)$ (cf. [9, pg. 116] or Tapia [20]). There is no need to be precise when dealing with $f''(x)$.

We omit the proof of the following lemma since it is so similar to the proof given in [17, pg. 359].

Lemma 2.1. Let f satisfy the conditions given above, then Δf is a consistent approximation rule for ∇f on D_0. Furthermore, if for some K, $\alpha > 0$ and every $x,y \in D$,

$$||\nabla f(x)-\nabla f(y)|| \leq K||x-y||^{\alpha}, \tag{2.4}$$

then Δf is a strongly consistent approximation rule of order α for ∇f on D_0.

The following lemma proved in [7], is also just a restatement of a well known result. We establish our notation before the statement. Let $F = (f_1,\ldots,f_m)^T$ be a continuously differentiable mapping from D into R^m where D is an open convex subset of R^n. Choose $D_0 \subset D$ and $\varepsilon > 0$ as before and define an approximation to the Jacobian matrix for F as follows. For $(x,h) \in D_0\times[0,\varepsilon]$ set

$$\Delta F(x,h) = \begin{bmatrix} \Delta f_1(x,h)^T \\ \vdots \\ \Delta f_i(x,h)^T \\ \vdots \\ \Delta f_m(x,h)^T \end{bmatrix} \tag{2.5}$$

Lemma 2.2. Let F satisfy the conditions given above, then ΔF is a consistent approximation rule for F' on D_0. Furthermore, if for some K, $\alpha > 0$, and every $x,y \in D$,

$$||F'(x)-F'(y)|| \le K||x-y||^{\alpha} \quad , \tag{2.6}$$

then ΔF is a strongly consistent approximation rule of order α for F' on D_0 .

Let us now consider a quite different example.

<u>Theorem 2.3</u>. Let F map a bounded open convex subset D of R^n into R^m where $m \geq n$. In addition, assume that F is continuously differentiable on D and that the rank of $F'(x)$ is the same for every $x \in D$. For $\mu > 0$ and $x \in D$, define the Levenberg approximation to $F'(x)^+$ by $L(x,\mu) = (\mu I+F'(x)^T F'(x))^{-1}F'(x)^T$. Then the Levenberg approximation is strongly consistent of order 1 with respect to the parameter μ.

<u>Proof</u>: $F'(x)^+$ is a continuous function of x on D since F is continuously differentiable and $F'(x)$ has constant rank on D (cf. Theorem 4.3 of [10]). Let $F'(x) = B(x)D(x)C(x)^T$ be a singular value decomposition of $F'(x)$ with $B(x)$ and $C(x)$ unitary square matrices of order m×n and n×n respectively. Then $F'(x)^+ = C(x)D(x)^+B(x)^T$. Therefore, we drop the argument x and remembering not to confuse $D(x)$ with the domain D we write

$$\begin{aligned} E(x,\mu) &= F'^+ - (\mu I + F'^T F')^{-1}F'^T \\ &= CD^+B^T - (\mu I + CD^T DC^T)^{-1}CD^T B^T \\ &= C[D^+ - (\mu I + D^T D)^{-1}D^T]B^T \end{aligned} \tag{2.7}$$

Now let $D(x)$ equal $\begin{pmatrix} \Lambda(x) & 0 \\ 0 & 0 \end{pmatrix}$ where $\Lambda(x)$ is the r×r diagonal matrix of positive singular values $\lambda_1(x), \ldots, \lambda_r(x)$ of $F'(x)$. Thus

$$E(x,\mu) = C\begin{pmatrix} \Lambda^{-1} - (\mu I + \Lambda^2)^{-1}\Lambda & 0 \\ 0 & 0 \end{pmatrix} B^T.$$

In order to bound $E(x,\mu)$, we first note that since the diagonal elements of $D(x)$ are the nonnegative square roots of the eigenvalues of $F'(x)^T F'(x)$, there is a neighborhood $N(x)$ about x such that for $y \in N(x)$, $\lambda_i(y) > \lambda_i(x)/2$ for $1 \le i \le r$. (See [18, p. 282].) Choose a finite subcover $N(x_1), N(x_2), \ldots, N(x_q)$ of the compact set $\bar{D}$. Clearly then for $x \in D$,

$$\lambda_i(x) > \min_{1\le j\le q} \min_{1\le k\le r} \lambda_k(x_j)/2 \equiv \bar{\alpha} .$$

We now return to the task of bounding $E(x,\mu)$. Since $||C(x)||_2 = ||B(x)||_2 = 1$, we need only bound $||\Lambda^{-1} - (\mu I + \Lambda^2)^{-1}\Lambda||$. But this is a nonnegative diagonal matrix so its ℓ_2 norm is its maximum diagonal element,

$$\max_{1\le i\le r} \frac{1}{\lambda_i} - \frac{\lambda_i}{\mu + \lambda_i^2} = \mu \cdot \max_{1\le i\le r} 1/\lambda_i(\mu + \lambda_i^2) \le \mu/\bar{\alpha}^3$$

which completes the proof.

In proving convergence of the Ben-Israel iteration, we will need an estimate of $||L(x,\mu)F(x) - F'(x)^+F(x)||$ which we obtain next.

Corollary 2.4. Assume the conditions of Theorem 2.3. Assume further that there is a point $x^* \in D$ such that $F'(x^*)^+F(x^*) = 0$ and

$$||F'(x)^+F(x) - F'(x^*)^+F(x^*)|| \le K\,||x - x^*||^\beta$$

for all $x \in D$. Then, for $\mu > 0$,

$$||L(x,\mu)F(x) - F'(x)^+F(x)|| \le (\mu/\bar{\alpha}^3) \cdot K\,||x - x^*||^\beta$$

for any $\bar{\alpha} > 0$, which is a uniform lower bound on the nonzero singular values of $F'(x)$, $x \in D$.

Proof: Clearly $L(x,\mu)F(x) = F'(x)^+F(x) - E(x,\mu)F(x)$. Then, from (2.7) and the fact that $D^T(x)D(x)D^+(x) = D^T(x)$, we again drop the subscripts and write

$$E(x,\mu)F(x) = C[D^+ - (\mu I + D^TD)^{-1}D^T]B^TF$$
$$= C[I - (\mu I + D^TD)^{-1}D^TD]D^+B^TF$$
$$= C[I - (\mu I + D^TD)^{-1}D^TD]C^T[F'^+F - F'(x^*)^+F(x^*)]$$

Therefore $||E(x,\mu)F(x)|| \le ||I - (\mu I + D^TD)^{-1}D^TD|| \cdot K||x - x^*||^\beta$

$$\le (\mu/\bar{\alpha}^3)K||x - x^*||^\beta$$

which completes the proof.

In [4], Brown and Dennis gave theoretical and computational justification for the use of finite differences in the Levenberg and Gauss-Newton methods. We will prove the consistency of the finite difference Levenberg approximation to $F'(x)^+$. For $\mu > 0$, $||h|| < \varepsilon$, $x \in D_0$ set

$$\tilde{L}(x,(\mu,h)) = (\mu I + \Delta F(x,h)^T \Delta F(x,h))^{-1} \Delta F(x,h)$$

but we must restrict the parameter vector $p = (\mu,h)$ somewhat. Notice that if μ goes to zero faster than h , then there is no guarantee that $\Delta F(x,h)$ has the same rank as $F'(x)$ and so it isn't clear that $\Delta F(x,h)^+ \to F'(x)^+$ as $||h|| \to 0$. Thus we define the parameter set $P = \{p=(\mu,h) \in (0,\infty) \times N(0,\varepsilon)$ and if $\{p_k\} \to 0$ as $k \to \infty$ then $||h_k||^\alpha/\mu_k^2 \to 0$ as $k \to \infty\}$. Here $P \subset R^{n+1}$ and α is from (2.6).

Theorem 2.5. Let F satisfy the hypothesis of Theorem 2.3 as well as the Lipschitz condition (2.6). Then $\tilde{L}$ as defined above on $D_0 \times P$ is a consistent approximation to $F'(x)^+$ on D_0 .

Proof: Set $\tilde{E}(x,p) = \tilde{L}(x,p)-F'(x)^+$ for $x \in D_0$, $(\mu,h) = p \in P$. If we add and subtract $L(x,\mu)$, we need only show that $||\tilde{L}(x,p)-L(x,\mu)|| \to 0$ as $||p|| \to 0$ uniformly on compact subsets of D_0, since, by Theorem 2.3, $||L(x,\mu)-F'(x)^+|| \to 0$ as $||p|| \to 0$ independent of $x \in D_0$. Now for $p = (\mu,h) \in P$,

$$(2.8)\quad \tilde{L}(x,p)-L(x,\mu) = \tilde{L}(x,p)-(\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1}F'(x)^T + (\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1}F'(x)^T-L(x,\mu) .$$

The first two terms on the right reduce to

$$(\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1}(F'(x)^T-\Delta F(x,h)^T)$$

which is bounded in norm by $C\,K||h||^\alpha/\mu$ for some C. The second two terms reduce to

$$[(\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1} - (\mu I+F'(x)^TF'(x))^{-1}]F'(x)^T .$$

Since $F'(x)$ is uniformly bounded on compact subsets of D_0, we need only consider the difference of the inverses. Each inverse is bounded in norm by $1/\mu$ and so the difference is bounded by

$$||F'(x)^TF'(x) \pm F'(x)^T\Delta F(x,h) - \Delta F(x,h)^T\Delta F(x,h)||/\mu^2$$

which is $O(||h||^\alpha/\mu^2)$ on any compact subset of D_0. If $p \to 0$ then because of the definition of P, $||h||^\alpha/\mu^2 \to 0$ and the proof is complete.

The situation is greatly simplified if $F'(x)$ has full rank for every $x \in D_0$. There is no longer any need for the restriction on P or for condition (2.6).

Theorem 2.6. Let F satisfy the hypothesis of Theorem 2.5 and assume in addition that the rank of $F'(x)$ is n for every $x \in D_0$. Let $P_0 = \{p=(\mu,h): \mu > 0, h \in E^n$, and $||h|| < \varepsilon\}$. Then $\tilde{L}$ defined on $D_0 \times P_0$ is a consistent approximation to $F'(x)^+ = [F'(x)^T F'(x)]^{-1} F'(x)^T$.

Furthermore, if D_0 is compact then there is some $\tilde{\varepsilon}$ such that if $P_1 = \{p=(\mu,h): \mu \geq 0, h \in E^n$ and $||h|| < \tilde{\varepsilon}\}$ then $\tilde{L}$ defined on $D_0 \times P_1$ is a strongly consistent approximation of order α to $F'(x)^+$ on D_0. In particular, the finite difference Gauss-Newton method is a strongly consistent, order α approximation to the Gauss-Newton method in this case.

Proof: Let D_1 be a compact subset of D_0. Since F is continuously differentiable and $F'(x)$ is of full rank on D_1, the smallest eigenvalue of $F'(x)^T F'(x)$, $\lambda(x)$ is a positive continuous function of x on the compact set D_1. Hence there is some positive $\lambda \leq \lambda(x)$ for every $x \in D_1$. From the definition of $\Delta F(x,h)$ and the hypotheses on F, for some constant C, $||\Delta F(x,h)^T \Delta F(x,h) - F'(x)^T F'(x)|| \leq C||h||^\alpha$ on D_1. There is hence an $\tilde{\varepsilon}$ such that for any $x \in D_1$, and $||h|| \leq \tilde{\varepsilon}$, $(\Delta F(x,h)^T \Delta F(x,h))^{-1}$ exists and is bounded in norm by $2/\lambda$, i.e., the smallest eigenvalue of $\Delta F(x,h)^T \Delta F(x,h)$ is bounded below by $\lambda/2$ for every $(x,h) \in D_1 \times [0,\tilde{\varepsilon}]^n$. This is a standard Banach lemma argument and need not be detailed here.

We point out here that the reason $\mu = 0$ is excluded from P_0 is that $\Delta F(x,h)$ doesn't necessarily have full rank for $||h|| < \varepsilon$. However if D_0 is compact, then D_1 in the above paragraph could be taken equal to D_0 and for $||h|| < \tilde{\varepsilon}$, $\Delta F(x,h)$ has full rank so $\tilde{L}(x,(0,h))$ is defined. Essentially, this is the only difference between the two parts of the proof.

From this point on, we argue as in the proof of Theorem 2.5 that we need only bound either side of (2.8), in this case, by $C||h||^{\alpha}$, to complete the proof. Break up the right side of (2.8) and note that for $||p||$ small enough, and some $\bar{C}$

$$||(\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1}(F'(x)^T-\Delta F(x,h)^T)|| \leq \frac{\bar{C}\, K||h||^{\alpha}}{\mu + \lambda/2} < 2\,\bar{C}\, K||h||^{\alpha} .$$

Similarly,

$$[(\mu I+\Delta F(x,h)^T\Delta F(x,h))^{-1} - (\mu I+F'(x)^TF'(x)^{-1}]F'(x)^T$$

is bounded in norm on D_1 for $||p||$ small enough by a constant times

$$\frac{||h||^{\alpha}}{(\mu+\lambda/2)(\mu+\lambda)}$$

and so is $O(||h||^{\alpha})$ and the proof is complete.

3. STABILITY ANALYSIS FOR THE CONTINUOUS ANALOGUE

This section is devoted to a development of asymptotic results for the continuous analogue initial value problems (1.4) and (1.6). Our technique will be to first view (1.6) as a perturbation of the continuous analogue initial value problem (1.4), which is in turn treated as a perturbation of an easily analyzed affine problem. Thus we write (1.6) as

$$x'(s) = -A(x) + [A(x)-G(x)] + [G(x)-\tilde{G}(x,p)] \tag{3.1}$$

$$x(0) = x_0 , \quad 0 \le s < \infty$$

where $A(x) \equiv Ax-b$ is an affine transformation.

Before proceeding with the analysis, let us illustrate (3.1) with steepest descent. The discussion in section 1 leads us to want an asymptotic analysis of

$$\begin{aligned} x'(s) &= -\Delta f(x,h),\ x(0) = x_0,\ 0 \le s < \infty . \\ &= -\nabla f(x) + [\nabla f(x)-\Delta f(x,h)] \\ &= -f''(x^*)(x-x^*) + [f''(x^*)(x-x^*)-\nabla f(x)] \\ &\qquad + [\nabla f(x)-\Delta f(x,h)] \quad . \end{aligned}$$

Now if the Hessian at x^* , $f''(x^*) = (\partial_{ij} f(x^*))$ is symmetric and positive definite, then it is well known (cf. [5, Chapter 3]) that the affine equation $x'(s) = -f''(x^*)x+f''(x^*)x^*$ is solved by $x(s)-x^* = (x_0-x^*)e^{-f''(x^*)s}$ which (we will see) has the curve $\phi(s) \equiv x^*$ as an asymptote. If x^* is a local minimizer for f then $\nabla f(x^*) = 0$ and so the second group of terms can be written as the negative of a Taylor series remainder $[\nabla f(x)-\nabla f(x^*)-f''(x^*)(x-x^*)]$. Thus in a neighborhood of x^* , the continuous analogue of steepest descent, $x'(s) = -\nabla f(x)$, is a perturbation of the affine problem. If we assume $f''(x^*)$

is the Fréchet derivative of ∇f at x^* then for any $\varepsilon > 0$ there is a $\delta > 0$ such that this perturbation, associated with the idealized iteration, is bounded in norm by $\varepsilon||x-x^*||$ for $||x-x^*|| < \delta$. An asymptotic analysis for this case of the continuous analogue (1.4) can be accomplished by standard means and its asymptotic behavior is essentially unchanged.

We have to extend the standard results in order to incorporate the third term into the stability theory for initial value problems. This term is the continuous analogue effect of using an approximation to the idealized iteration. Our extended theory can't really be expected to imply the same sort of asymptotic convergence as before, since $\nabla f(x)-\Delta f(x,h)$ depends on h instead of $x-x^*$.

The concept of consistency developed in section 2 will allow us to draw some interesting conclusions about the behavior of solutions to (1.6) as s becomes large. In the case of (1.4) under the usual hypotheses, we expect any solution $x(s)$ to have the property that for any ε there is an S such that $||x(s)-x^*|| < \varepsilon$ for $s > S$. For (1.6) this property holds generally only for ε greater than some lower bound which depends on $||p||$. For steepest descent, we will conclude roughly that

$$\overline{\lim_{s\to\infty}} ||x(s)-x^*|| \leq O(||h||^{\alpha}k_2(f''(x^*))) ,$$

where $k_2(\,\circ\,)$ is the ℓ_2 condition number of the matrix argument.

We state the following lemma because the results will be useful but we omit the proof because it is not central to this paper and it can be found elsewhere, [5].

Lemma 3.1. Let A be a real $n \times n$ matrix and let $A = C \Lambda C^{-1}$ be its Jordon canonical form. Assume that the real part of λ is positive for λ chosen to be the eigenvalue of A with minimal real part. For any real t, define

$$e^{-At} = I + \sum_{k=1}^{\infty} (-1)^k t^k A^k / k! \; .$$

Then there exist positive constants, ν, σ such that

$$||e^{-At}|| \leq \nu e^{-\sigma t} \; .$$

The values of ν and σ, depend on A and the norm as follows.

(i) If A is symmetric, then $\sigma = \lambda$ and, if the ℓ_2 norm is used, $\nu = 1$.

(ii) If the largest Jordon block corresponding to λ is 1×1, then σ is the real part of λ and, if the ℓ_2 norm is used, $\nu = k_2(C)$.

(iii) If the largest Jordon block corresponding to λ is not 1×1 then $\nu = k_2(C)/\sigma$ in the case of the ℓ_2 norm, but in any case σ is one-half the real part of λ.

Now we are ready to give the stability result we require. It is a modification of a well-known theorem which can be found in [5].

Theorem 3.2. Let $r \in (0,1)$ and let D be an open convex neighborhood in R^n of x^*, a zero of $G: D \to R^n$. Assume that the Fréchet derivative $G'(x^*) = A$ exists, that its eigenvalues all have positive real parts and let the constants ν and σ be defined by Lemma 3.1. Finally let $\tilde{G}$ be a consistent approximation rule for G on D. Then there is an $\varepsilon > 0$ and a neighborhood D_r of x^* such that any solution to

$$x'(s) = -\tilde{G}(x,p) \ , \ x(0) = x_0 \in D_r \ , \ ||p|| < \xi, \ 0 \leq s \ ,$$

satisfies

$$\overline{\lim_{s\to\infty}} ||x(s) - x^*|| \leq [\sup_{x\in D} ||G(x)-\tilde{G}(x,p)||] \, \nu/\sigma(1-r) .$$

Proof: From [5, p. 328], every solution of problem (3.1) satisfies

$$x(s)-x^* = e^{-As}(x(0)-x^*) + \int_0^s e^{-A(s-u)}[A(x)-G(x)] \, du$$
$$+ \int_0^s e^{-A(s-u)}[G(x)-\tilde{G}(x,p)]du$$

Define $q(p) = \sup_{x\in D}||G(x)-\tilde{G}(x,p)||$ and let $\varepsilon < \sigma r/\nu$. Choose $\delta > 0$ so that $||A(x)-G(x)|| = ||A(x)-G(x)+G(x^*)|| < \varepsilon|| x-x^*||$ for $||x-x^*|| < \delta$. By Lemma 3.1 we have

$$(3.3) \quad ||x(s)-x^*|| \leq \nu e^{-\sigma s}||x_0-x^*||+ \nu\varepsilon\int_0^s e^{-\sigma(s-u)}||x(u)-x^*||du$$
$$+ \nu\int_0^s e^{-\sigma(s-u)}q(p)du .$$

Since $\sigma > 0 \leq s$, $e^{-\sigma s} \leq 1$ so for any $s \geq 0$,

$$||x(s)-x^*|| \leq \nu||x_0-x^*|| + \nu\varepsilon[\max_{0\leq u\leq s} ||x(u)-x^*|| + \nu q(p)]/\sigma \quad ,$$

which resolves to

$$\text{(3.4)} \qquad \max_{0\leq u\leq s} ||x(u)-x^*|| \leq \nu\sigma||x_0-x^*||/(\sigma-\nu\varepsilon)$$

$$+ \nu q(p)/(\sigma-\nu\varepsilon)$$

$$\leq [\sigma\nu||x_0-x^*||+\nu q(p)]/(1-r)\sigma \quad .$$

Now choose ξ and D_r so that the right hand side of (3.4) is bounded above by δ .

It remains for us to bound $\gamma \equiv \overline{\lim_{s\to\infty}} ||x(s)-x^*||$. Choose a sequence $\{s_j\} \to \infty$ such that $||x(s_j)-x^*|| \underset{s\to\infty}{\to} \gamma$. From inequality (3.3) we obtain

$$\text{(3.5)} \qquad ||x(s_j)-x^*|| \leq ||x_0-x^*||\nu e^{-\sigma s}$$

$$+ \nu\varepsilon\int_0^{s_j/2} e^{-\sigma(s_j \; -u)}||x(u)-x^*||du$$

$$+ \nu\varepsilon\int_{s_j/2}^{s_j} e^{-\sigma(s_j-u)}||x(u)-x^*||du$$

$$+ \nu\int_0^{s_j} e^{-\sigma(s_j-u)}q(p)du \quad .$$

Let η be arbitrary in the interval $(0,\gamma)$. Then there is an integer j_η such that for every $j \geq j_\eta$, $\gamma-\eta < ||x(s_j)-x^*|| < \gamma+\eta$. Thus, for $j > j_\eta$, (3.5) implies

$$\gamma-\eta < ||x_0-x^*||\nu e^{-\sigma s_j} + (\varepsilon\delta\nu e^{-\sigma s_j/2})/\sigma$$

$$+ \varepsilon(\gamma+\eta)\;\nu/\sigma \;+ q(p)\;\nu(1-e^{-\sigma s_j})/\sigma \quad .$$

As $j \to \infty$ this becomes

$$\gamma-\eta < r(\gamma+\eta) + q(p)\nu/\sigma$$

and since η was arbitrary we get

$$\gamma \leq r\gamma + q(p)\nu/\sigma \tag{3.6}$$

from which the result follows.

We remark here that the proof of Theorem 3.2 rests on obtaining a bound on the integral representation of a solution to (3.1) and thus also demonstrates the existence of at least one solution to (3.1).

It is interesting to note at this point that the choice of D_r, ξ right after inequality (3.4) in the proof could have been less arbitrary. We could have decreased the final bound (3.6) by decreasing ξ and this would have allowed a wider choice of x_0 to achieve the bound. On the other hand, we could have restricted x_0 very sharply and allowed a wider latitude with respect to the parameter p. Furthermore, the parameter r accelerates this interdependence as it is taken nearer zero. We can paraphrase all this as follows: <u>No matter how small we choose $||p|| > 0$ and $||x_0-x^*|| > 0$, we cannot expect to come closer than $\sup_{x\varepsilon D}||G(x)-\overline{G}(x,p)||\nu/\sigma$ to x^*</u>.

4. NUMERICAL INTEGRATION OF THE CONTINUOUS ANALOGUE

In the previous section we gave a theorem which established the asymptotic behavior of the continuous analogue. Our purpose here is to first give a certain class of numerical integration procedures which mirror this behavior when applied to the affine part of problem (3.1). We then show that the effects of non-linearity and consistent approximations are just the same for these numerical solutions as for the analytic solutions.

We assume that the reader is familiar with linear multi-step methods for initial value problems and we present only the following definition which generalizes one given in [1] for a class of methods having the desired properties.

<u>Definition 4.1</u>. A linear multi-step method is <u>weakly A-stable</u> if for any matrix A which satisfies the hypothesis of Lemma 3.1, there exist positive constants $\underline{t} < \bar{t}$ for which the following is true. If the method with steplengths $t_k \in [\underline{t},\bar{t}]$, $k=0,1,\ldots$ is applied to

$$x'(s) = -A[x-x^*],\ x(0)=x_0,\ 0\leq s,\ x^* \text{ arbitrary}$$

then the sequence $\{x_k\}$ generated by the method converges to x^*.

We mentioned in section 1 that all the methods of interest here can be viewed as Euler's method applied to (1.6). Hence we will restrict our attention to Euler's method.

<u>Lemma 4.2</u>. Euler's method is weakly A-stable.

<u>Proof</u>: Let x^* be arbitrary and $e_k = x_k-x^*$. Apply Euler's method with step size t_k to $x'(t) =-A[x-x^*]$, $x(0) = x_0$ to obtain

$$e_{k+1} = e_k - t_kAe_k = [\prod_{i=0}^{k} (I-t_iA)]e_0 \quad .$$

Thus $\{||e_k||\}$ converges to zero if the spectral radius of $I-t_iA$ is uniformly bounded by some number less than 1. Let $\lambda_j = a_j + ib_j$ be any eigenvalue of A. Then $a_j > 0$ and $\mu_j = 1-ta_j - itb_j$ is the corresponding eigenvalue of $I-tA$ and $|\mu_j|^2 < 1$ for any $t \in (0,2a_j/|\lambda_j|^2)$. Thus, take $0 < \underline{t} < \bar{t} = \min_{1\leq i\leq n} 2a_i/|\lambda_i|^2$ and the proof is complete.

The following theorem is a partial extension of a result in [1]. It shows that even in the more general problem (3.1), the desired results hold for steplengths controlled by the affine term. Roughly, this is because this term dominates as s gets large.

<u>Theorem 4.3</u>. Assume the conditions of Theorem 3.2 and let $0 < \underline{t} < \bar{t} < \min\{2a/|\lambda|^2\text{: } \lambda = a+ib \text{ is an eigenvalue of } A\}$.

Then there are constants $\tilde{\xi}, \zeta > 0$ and a neighborhood $\tilde{D}_r$ of x^* such that any solution to

(4.3) $$x_{k+1} = x_k - t_k\tilde{G}(x_k,p) \ , \quad x_0 \in \tilde{D}_r$$

$$||p|| \leq \tilde{\xi} \ , \ t_k \in [\underline{t},\bar{t}] \ , \ k > 0$$

has the property that

$$\overline{\lim_{k\to\infty}} ||x_k-x^*|| \leq [\sup_{x\in D}||G(x)-\tilde{G}(x,p)||]\zeta/\sigma(1-r)$$

where σ is defined for A in Lemma 3.1.

<u>Proof</u>: Define $e_k = x_k-x^*$ and rewrite (4.3) as

(4.4) $$e_{k+1} - (I-t_kA)e_k = t_k[Ae_k-G(x_k)+G(x_k)-\tilde{G}(x_k,p)] = t_k \cdot d_k$$

which defines d_k.

Now in a manner similar to that in [11], we construct a solution to (4.4) in the form $e_k = y_k + z_k$ where y_k is a solution to the homogeneous problem ((4.4) with $d_k \equiv 0$) satisfying the initial condition and z_k is a particular solution to (4.4) with zero initial data. The reader may verify that y_k and z_k are given as follows.

Let $M_{i0} = I,\ i \geq 0$ and

$$M_{ij} = \prod_{m=0}^{j-1} (I-t_{i-1-m}A) \ , \ j > 0 \ .$$

Then

$$y_k = M_{kk}e_0 \ , \ k \geq 0$$

and

$$z_0 = 0$$

$$z_k = \sum_{i=1}^{k} M_{k,k-1}t_{i-1}d_{i-1} \ , \ k \geq 0$$

For any ε there is a $\delta > 0$ such that for $||e_j|| \leq \delta$ we have

$$||d_j|| \leq \varepsilon||e_j|| + q(p)$$

where again $q(p) = \sup_{x \varepsilon D}||G(x)-\tilde{G}(x,p)||$. We now have that if $||e_j|| < \delta$, $j = 0,\ldots,k-1$

$$||e_k|| \leq ||y_k|| + ||z_k|| \tag{4.5}$$

$$\leq ||y_k|| + \bar{t} \sum_{i=1}^{k} ||M_{k,k-1}||\cdot[\varepsilon||e_{i-1}|| + q(p)] \ .$$

Then since $||y_k||$ is monotonically decreasing and $y_0 = e_o$, it follows that

$$(4.6)\quad \max_{0\le i\le k} \|e_i\| \le \|e_0\| + \bar{t} \max_{0\le i\le k} \|e_i\| \sum_{i=1}^{k} \|M_{k,k-i}\|$$

$$+ \bar{t}q(p) \sum_{i=1}^{k} \|M_k,k-1\|$$

Now $\sum_{i-1}^{k} \|M_{k,k-i}\|$ can be bounded as follows. Let t^* maximize $\|I-tA\|$ for $t \in [\underline{t},\bar{t}]$. Then for $\bar{M} = I-t^*A$, $\|\bar{M}\| < 1$ by Lemma 4.2 and $\sum_{i=1}^{k} \|M_{k,k-1}\| \le \sum_{i=1}^{k} \|\bar{M}\|^{k-1} <$ $\sum_{i=0}^{\infty} \|\bar{M}\|^i = 1/(1-\|\bar{M}\|) \equiv \zeta$. Thus from (4.6)

$$\max_{0\le i\le k} \|e_i\| \le (\|e_0\| + \bar{t}q(p)\zeta)/(1-\bar{t}\varepsilon\zeta)$$

and for $\varepsilon = r/\bar{t}\zeta$ and $\|e_0\|$ and $q(p)$ sufficiently small, $\max_{0\le i\le k} \|e_i\| < \delta$ for all k .

We now compute $\overline{\lim_{k\to\infty}} \|x_k\| \equiv \gamma$. For every $\eta > 0$ there is an integer k_η and an infinite set of integers $N = \{i: i\ge k$ and $\|e_i\| > \gamma-\eta\}$. Also there is an integer $\hat{k}_\eta$ such that $\gamma + \eta > \|e_i\|$ for all $i > \hat{k}_\eta$. We may rewrite (4.5) as

$$\|e_i\| \le \|y_i\| + \sum_{j=1}^{i/2} \|M_{i,i-j}\|\cdot\|d_j\|$$

$$+ \sum_{j=i/2+1}^{i} \|M_{i,i-j}\|\cdot\|d_j\|$$

and thus for all $i \varepsilon N$ such that $i/2 > \hat{k}_\eta$,

$$\gamma-\eta < \|e_i\| \le \|y_i\| + \bar{t}\varepsilon\delta \sum_{j=1}^{i/2} \|M_{i,i-j}\| + \bar{t}\varepsilon(\gamma+\eta)\zeta$$

$$+ \bar{t}q(p)\zeta .$$

As $i \to \infty$, $\sum_{j=1}^{i/2} ||M_{i,i-j}|| \to 0$ and $y_i \to 0$. Therefore in the limit as $i \to \infty$,

$$\gamma-\eta \leq \bar{t}\varepsilon(\gamma+\eta)\zeta + \bar{t}q(p)\zeta . \tag{4.7}$$

But (4.7) must hold for all η and therefore

$$\gamma \leq \bar{t}q(p)\zeta/(1-\bar{t}\varepsilon\zeta)$$

and the result follows.

Note that here, as in the case of Theorem 3.2, the proof also implies the existence of at least one solution.

If A is symmetric and $\bar{t}$, instead of being maximized, is chosen so that the eigenvalues of $I-\bar{t}A$ are positive, then $\zeta = 1/\bar{t}\sigma$ where σ is the smallest eigenvalue of A. Moreover the result is then identical to Theorem 3.2, since $\nu = 1$ in this case.

In some cases, e.g. Newton's method (cf. section 5.2), the use of consistent approximations does not prevent convergence of the iteration to x^*. In these cases, the expression $[G(x)-\tilde{G}(x,p)]$ satisfies a stronger condition which enables it to be handled in the same manner as the term $[A(x-x^*)-G(x)]$. We state the following useful corollary.

Corollary 4.4. Assume the hypotheses of Theorem 4.3 and let the following condition hold: For every $\varepsilon > 0$ there exists $\delta > 0$ and a vector p such that $||G(x)-\tilde{G}(x,p)|| \leq \varepsilon||x-x^*||$ for $||x-x^*|| \leq \delta$. Then any solution of (4.3) tends to x^* as $k \to \infty$.

5. APPLICATION TO SPECIFIC METHODS

In this section, we will identify each of the methods of section 1 with Euler's linear multistep method in order to apply Theorem 4.3. The results give insight into the effect of using a consistent approximation to one of the well-known methods. Our intent is really more negative than positive, and it is to warn that for some problems, such as unconstrained minimization and nonlinear least squares, the parameter which controls the approximation must be allowed to get small in direct proportion to the accuracy required in the final answer. Our bounds will always include a constant factor C, which is a device which allows us to avoid cluttering details about the particular norms used in various parts of the bounds.

5.1 Steepest Descent.

Theorem 5.1. Let $f: D \subset E^n \to E^1$ be Fréchet differentiable on the open convex set D. Assume $x^* \in D$ has the properties that $\nabla f(x^*) = 0$ and f has a second Fréchet derivative H^* at x^* which is symmetric and positive definite with smallest eigenvalue σ. Let $r \in (0,1)$ and $\bar{t}$ be a positive number smaller than $2/||H^*||_2$. Then there exist positive constants ρ and ξ and a region D_r such that any solution of

$$(5.1) \quad x_{k+1} = x_k - t_k \Delta f(x_k,h)\ ,\ x_0 \in D_r\ ,\ ||h|| \leq \xi\ ,$$

$$t_k \in [\underline{t},\ 2/||H^*||)\ ,\ k \geq 0$$

satisfies

$$\overline{\lim_{k\to\infty}} \|x_k-x^*\| \le \sup_{x\varepsilon D}\|\Delta f(x,h)-\nabla f(x)\|\rho/\sigma(1-r) .$$

Furthermore, if there are constants K, $\alpha > 0$ such that for every $x,y \in D$

$$\|\nabla f(x)-\nabla f(y)\| \le K\|x-y\|^{\alpha}$$

then

$$\overline{\lim_{k\to\infty}} \|x_k-x^*\| \le CK\|h\|^{\alpha}\rho/\sigma(1-r) \text{ , for some } C > 0 \text{ independent of } h$$

Proof: We write (5.1) as

$$(5.2)\quad x_{k+1} = x_k - t_k H^*(x_k-x^*) + t_k[\nabla f(x^*)-\nabla f(x_k)+H^*(x_k-x^*)]$$
$$+ t_k[\nabla f(x_k)-\Delta f(x_k,h)] .$$

By identifying H^* with A , $\nabla f(x)$ with $G(x)$ and $\Delta f(x,h)$ with $\tilde{G}(x,p)$, the result follows from an application of Theorem 4.3 to problem (5.1).

Corollary 5.2. Let the hypotheses of Theorem 5.1 hold and assume that for some constants L, $\overline{\alpha}$, and every $x,y \in D$, f is second differentiable and

$$\|f''(x)-f''(y)\| \le L\|x-y\|^{\overline{\alpha}} \text{ , } \overline{\alpha} \in (0,1] .$$

Then, any solution $\{x_k\}$ to the difference equation (5.1) satisfies

$$(5.3)\quad \overline{\lim_{k\to\infty}} \|x_k-x^*\| \le \frac{C\rho\|h\|}{(1-r)} [k_2(H^*)+\frac{L}{\sigma}\sup_{x\in D}\|x-x^*\|^{\overline{\alpha}}] .$$

Proof: The result follows easily from the well-known fact that a first order Lipschitz constant for a function in a convex region can be taken as a bound on the norm of the derivative of the function in the same region. Thus, in the conclusion of Theorem 5.1 we can take $\alpha = 1$ and

$$K = \sup_{x\varepsilon D}||f''(x)||_2 = ||H^*||_2 + \sup_{x\varepsilon D}||f''(x) - H^*||_2$$
$$\leq ||H^*||_2 + \sup_{x\varepsilon D} L||x-x^*||_2^{\overline{\alpha}} \quad .$$

Of course, $||H^*||_2/\sigma = k_2(H^*)$.

The main point of this paper lies in expressions like (5.3). If you use a fixed step size finite difference gradient to try to minimize even a very smooth function $(\overline{\alpha} = 1)$, then the distance by which you should expect to miss the minimizing point is directly proportional to the magnitude of the perturbation used in the gradient approximation. Furthermore, the constant of proportionality is composed of two parts; one depends on the conditioning of the quadratic Taylor approximation to f near the minimum and the other depends jointly on how much f actually deviates from that quadratic model ($L = 0$ iff f is quadratic) and how far one starts from the minimizing point.

Now, clearly it is not always impossible to find x^* using fixed size finite difference gradient, but rather the point is that neither is it necessarily always possible. Dennis [6] considered the strictly convex quadratic function $f(x) = x^TH^*x$ where

$$H^* = \begin{pmatrix} 2.6 & -2.4 \\ -2.4 & 2.5 \end{pmatrix} .$$

The iteration $x_{k+1} = x_k - t_k \Delta f(x_k, 10^{-3}x_0)$, $x_0 = (100,105)^T$ was carried out in double precision (APL) on Cornell University's IBM 360/65. In this computation, t_k was computed from a formula for the exact minimizer for $f(x_k - t\Delta f(x_k, 10^{-3}x_0))$. The choice of h or $p = 10^{-3}x_0$ was made because it is often used in practice.

Let θ_k be the angle between $\Delta f(x_k,h)$ and $\nabla f(x_k)$. Initially, the approximate gradient was excellent with $\cos\theta_0 = .99998$. After 42 iterations, the progress towards the minimum at the origin was excellent with $\cos\theta_{42} \doteq .75$ and $x_{42} = (.34678, .386$ After 50 iterations, the method had fallen into the trap predicted by the theory with $x_{50} = (.00083, .00490)^T$, $\cos\theta_{50} \doteq -4 \times 10^{-7}$ and convergence to six decimal places apparent. If we had switched to a central difference at this point, it would probably have been possible to decrease the function a bit further.

5.2 Newton's Method.

In this case, it is well known that it is even possible to implement the finite difference analogue in such a way as to preserve the second order convergence of the original method. See [6]. We include this result for completeness and because it allows a greater latitude of step size than previous results.

<u>Theorem 5.2</u>. Let $F: D \subset R^n \to R^n$ have a continuous Fréchet derivative $F'(x)$ on the open convex set D_0. Assume $x^* \in D$ is such that $F(x^*) = 0$ and that $F'(x^*)^{-1}$ exists. Then there exist a positive constant ξ and a region D_r such that if $\underline{t} \in (0,2)$ then any solution of

$$(5.3) \quad x_{k+1} = x_k - t_k \Delta F(x_k,h)^{-1} F(x_k), \; x_0 \in D_r, \; ||h|| \leq \xi,$$

$$t_k \in [\underline{t},2), \; k \geq 0$$

satisfies $x_k \to x^*$.

<u>Proof</u>: We write (5.3) as

$$x_{k+1}-x^* = x_k - t_k(x_k-x^*) + t_k F'(x_k)^{-1}[F(x_k)-F(x^*)+F'(x_k)^{-1}(x_k-x^*)]$$
$$+ t_k[F'(x_k)^{-1}F(x_k) - \Delta F(x_k,h)^{-1}F(x_k)] .$$

We again identify I with A, $F'(x)^{-1}F(x)$ with $G(x)$ and $F(x,h)^{-1}F(x)$ with $\tilde{G}(x,p)$. Now, however,

$$||F'(x)^{-1}F(x) - \Delta F(x,h)^{-1}F(x)|| \leq ||F'(x)^{-1} - \Delta F(x,h)^{-1}||\cdot||F(x)||$$
$$\leq ||F'(x)^{-1}||\cdot||F'(x)-\Delta F(x,h)||\cdot||\Delta F(x,h)^{-1}|| \cdot ||F(x)-F(x^*)||$$

Since $F'(x^*)^{-1}$ exists we may assume without loss of generality that $F(x)^{-1}$ exists and is uniformly bounded on D. Similarly, by the consistency assumption, we may as well assume $\Delta F(x,h)^{-1}$ exists and is uniformly bounded on D for $||h||$ sufficiently small. The consistency assumption also implies that there is a function $q(h)$ such that $||F'(x)-\Delta F(x,h)|| \leq q(h)$ and $q(h) \to 0$ as $||h|| \to 0$ uniformly for $x \in D$. Thus there is a constant

$\eta \geq 0$ such that

$$||F'(x)^{-1}F(x)-\Delta F(x,h)^{-1}F(x)|| \leq \eta q(h)||x-x^*|| \quad .$$

Now for any $\varepsilon > 0$ there is a h such that

$$||F'(x)^{-1}F(x)-\Delta F(x,h)^{-1}F(x)|| \leq \varepsilon||x-x^*||$$

and the result now follows immediately from Corollary 4.4.

5.3 Levenberg-Marquardt/Gauss-Newton.

In section 2 we proved that the finite difference Levenberg-Marquardt iteration could be viewed as a consistent approximation to the Ben-Israel iteration which is the same as the Gauss-Newton iteration if $F'(x)$ has full column rank. The following theorem gives the appropriate convergence result.

Theorem 5.3. Let $F\colon\ D \subset R^n \to R^m$, $m \geq n$ have a continuous Fréchet first derivative at each point x of the open convex set D and let the rank be a constant independent of x . In addition let $x^* \in D$ have the properties that it is a zero of $F'(x)^TF(x)$ and that this function has a positive definite Fréchet derivative at x^* with largest eigenvalue λ_1 . Let the smallest eigenvalue of $F'(x^*)^TF'(x^*)$ be λ_2 .

Then for any $r \in (0,1)$ and any $\mu > 0$, there exist constants σ, ξ, ρ and a neighborhood D_r of x^* such that any solution of

$$x_{k+1} = x_k - t_k(\mu I+\Delta F(x_k,h)^T\Delta F(x_k,h))^{-1}\Delta F(x_k,h)^TF(x_k)$$

$$x_0 \in D_r \ , \ ||h|| \leq \xi \ , \ t_k \in [\underline{t},\ 2/\lambda] \quad \underline{t} > 0$$

where λ is the largest eigenvalue of $(\mu I+F'(x^*)^T F'(x^*))^{-1}\cdot [F'(x^*)^T F'(x^*) + \sum_{i=1}^{m} f_i(x^*) f_i''(x^*)]$ and $2/\lambda \leq 2(\mu+\lambda_2)/\lambda_1$, satisfies

$$\overline{\lim_{k\to\infty}} ||x_k - x^*|| \leq [\sup_{x\in D_r} ||F'(x)-\Delta F(x,h||]\ \rho/\sigma(1-r) \quad .$$

Furthermore, if for some K, $\alpha > 0$ and every $x, y \in D$,

$$||F'(x) - F'(y)|| \leq K||x-y||^{\alpha}$$

then for some $C > 0$, independent of h,

$$\overline{\lim_{k\to\infty}} ||x_k - x^*|| \leq C\cdot K||h||^{\alpha}\ \rho/\sigma(1-r) \quad .$$

Proof: The proof follows simply from Theorem 4.3 once we identify G with $(\mu I+F'(x)^T F'(x))^{-1} F'(x)^T F(x)$, A with $G'(x^*) = (\mu I+F'(x^*)^T F'(x^*))^{-1}[(F'(x^*)^T F'(x^*) + \sum_{i=1}^{m} f_i(x^*) f_i''(x^*)]$ and $\tilde{G}(x,p)$ with $\tilde{L}(x,(\mu,h))$. Note that if the rank of $F'(x)$ is not n, then μ may not be taken to be 0.

In the full rank case, we may take $\mu = 0$ and obtain the convergence of the Gauss-Newton iteration. One interesting sidelight is the following corollary.

Corollary 5.4. Under the hypotheses of Theorem 5.3 with $F'(x)$ of rank n, if the largest eigenvalue of $[F'(x^*)^T F'(x^*)]^{-1} [F'(x^*)^T F'(x^*) + \sum_{i=1}^{m} f_i(x^*) f_i''(x^*)] < 2$ then the Gauss-Newton (with $t_k = 1$) iteration is locally convergent to x^*.

Proof: Take $p = 0$ and notice that $\bar{t} \geq 1$.

The reader will find a similar but less general condition in [4], namely that $K||F(x^*)|| < \lambda_1$. The condition is, of course, to be expected since it is the same as the spectral radius of $G'(x^*)$ less than 1 and so it would be predicted by Ostrowski's Theorem [17].

We complete this section by giving a convergence proof for the Ben-Israel iteration in the rank deficient case. To do this, however, we need a somewhat stronger condition on F. This is necessary to compensate for the fact that at x^* the A matrix ($G'(x^*)$ for $G(x) = F'(x)^+F(x)$) is singular, which means that the asymptotic character of the solutions is determined by the perturbations. An alternate approach, to handle this and other singularities, is given in [2].

Theorem 5.5: Assume the conditions of Theorem 5.3 except that the rank of $F'(x)$ may be less than n. Assume the conditions of Corollary 2.4 with $\beta \geq 1$. Then for any $r \in (0,1)$ and for any $\mu > 0$ there is a neighborhood D_r such that any solution of

$$x_{k+1} = x_k - t_k F'(x_k)^+ F(x_k)$$

$$x_0 \in D_r,\ t_k \in [\underline{t},\ 2\mu/\lambda_i]\ ,\ \underline{t} > 0, \lambda,\ \text{as in Theorem 5.3}$$

tends to x^* as $k \to \infty$.

Proof: Identify $\tilde{G}(x,\mu)$ with $(\mu I + F'(x)^T F'(x))^{-1} F'(x)^T F(x)$, A with $G_x(x^*,\mu)$ and $G(x)$ with $F'(x)^+F'(x)$. Then apply Corollaries 2.4 and 4.4.

Note that to use an arbitrarily large steplength we must, by Corollaries 2.4 and 4.4, start correspondingly close to x^*.

1. Boggs, P.T. (1971) "The solution of nonlinear operator equations by A-stable integration techniques", SIAM J. Numer. Anal. 8, pp. 767-785.

2. ___________ (1975) "The convergence of the Ben-Israel iteration for nonlinear least squares problems (to appear in Math. Comp.).

3. Bosarge, W.E., Jr., (1968) "Infinite dimensional iterative methods and applications", IBM Publications 230-2347, Houston.

4. Brown, K.M., and J.E. Dennis, Jr., (1970), "Derivative-free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares approximation," Numer. Math., 18, pp. 289-297.

5. Coddington, E.A., and N. Levinson, (1955) Theory of Ordinary Differential Equations, McGraw-Hill, New York.

6. Dennis, J.E., Jr. (1971), "Algorithms for nonlinear problems which use discrete approximations to derivatives," Proc. ACM 1971 Nat'l Conference, Chicago.

7. _______________ (1970) "On the convergence of Newton-like methods", in Numerical Methods for Nonlinear Algebraic Equations, edited by Philip Rabinowitz, Gordon-Breach, New York.

8. Gavurin, M.K., (1958) "Nonlinear functional equations and continuous analogs", Isv. Vyss. Ucebn. Zaved. Mathematika, 6 pp. 18-31; English Transl., Tech. Rep. 68-70, University of Maryland, College Park, 1968.

9. Goldstein, A.A., (1967), Constructive Real Analysis, Harper & Row, New York.

10. Golub, G.H., and V. Pereyra (1972), "The differentiation of pseudoinverses and nonlinear least square problems whose variables separate", SIAM J. Numer. Anal. 10, pp. 413-432.

11. Henrici, P., (1962), Discrete Variable Methods in Ordinary Differential Equations, John Wiley, New York.

12. Hurt, J., (1967), "Some stability theorems for ordinary difference equations," SIAM J. Numer. Anal. 4, pp. 582-596.

13. Levenberg, K. (1944), "A method for the solution of certain nonlinear problems in least squares", Quart. Appl. Math. 2, pp. 164-168.

14. Marquardt, D.W., (1963), "An algorithm for least squares estimation of nonlinear parameters", SIAM J. Appl. Math., 11, pp. 431-441.

15. Meyer, Gunter H., (1968), "On solving nonlinear equations with a one-parameter operator imbedding", SIAM J. Numer Anal. 5, pp. 739-752.

16. Ortega, J.M. (1973), "Stability of difference equations and convergence of iterative processes", SIAM J. Numer. Anal. 10, pp. 268-282.

17. ___________, and W.C. Reinboldt, (1970), Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York.

18. Ostrowski, A.M., (1966), Solution of Equations and Systems of Equations. Academic Press, New York.

19. Rao, Radhakrishna C., and Sujit Kumar Mitra (1971), Generalized Inverse of Matrices and Its Applications, John Wiley, New York.

20. Tapia, R.A., (1971), "The differentiation and integration of nonlinear operators", in Nonlinear Functional Analysis and Applications (Louis, B. Rall, ed.), Academic Press, New York.