MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 140
OCTOBER 1977, PAGES 962–973

Modifying Singular Values: Existence of Solutions to Systems of Nonlinear Equations Having a Possibly Singular Jacobian Matrix

By David Gay*

Abstract. We show that if a certain nondegeneracy assumption holds, it is possible to guarantee the existence of a solution to a system of nonlinear equations $f(x) = 0$ whose Jacobian matrix $J(x)$ exists but may be singular. The main idea is to modify small singular values of $J(x)$ in such a way that the modified Jacobian matrix $\hat{J}(x)$ has a continuous pseudoinverse $\hat{J}^+(x)$ and that a solution x^* of $f(x) = 0$ may be found by determining an asymptote of the solution to the initial value problem $x(0) = x_0$, $x'(t) = -\hat{J}^+(x)f(x)$. We briefly discuss practical (algorithmic) implications of this result. Although the nondegeneracy assumption may fail for many systems of interest (indeed, if the assumption holds and $J(x^*)$ is nonsingular, then x^* is unique), algorithms using $\hat{J}^+(x)$ may enjoy a larger region of convergence than those that require (an approximation to) $J^{-1}(x)$.

1. Introduction. In various settings it is necessary to solve a system of nonlinear equations. Thus, given a mapping $f\colon \mathbf{R}^n \longrightarrow \mathbf{R}^n$, it is necessary to find a point $x^* \in \mathbf{R}^n$ such that $f(x^*) = 0$. Often f is continuously differentiable, i.e., $f \in C^1(\mathbf{R}^n)$, as we shall henceforth assume.

Frequently certain features of the environment in which f arises, such as physical features, imply the existence of a solution x^*. However, it is of theoretical interest to determine conditions on f which imply the existence of a solution without employing "outside" considerations. Both constructive and nonconstructive approaches are possible. For example, degree theory represents a nonconstructive approach (see Chapter 6 of [Ortega and Rheinboldt, 1970]). Particular algorithms usually underlie constructive existence theorems. Newton's method, for instance, underlies the well-known Kantorovich theorem (see below), which can only deal with an isolated solution. In this paper we present a constructive existence theorem based on integrating a certain differential equation. Our assumptions are weaker than those in the Kantorovich theorem, and they allow situations in which a continuum of solutions x^* exists.

In the next section we introduce some notation and, for reference, state several theorems. Section 3 presents our main results, and Section 4 discusses some implications for practical algorithms.

Received April 13, 1976; revised July 26, 1976 and November 1, 1976.

AMS (MOS) subject classifications (1970). Primary 65H10; Secondary 65F20.

Key words and phrases. Continuation methods, eigenvalues, nonlinear equations, pseudoinverse, singular value decomposition.

*NBER Computer Research Center. Research supported in part by National Science Foundation Grants DCR75-10143 and MCS76-00324 to the National Bureau of Economic Research, Inc.

A number of other authors have considered integrating various differential equations in order to solve a system of nonlinear equations. See [Boggs, 1970] for a survey of such work. Fletcher [1970] has briefly considered "modifying" singular values by the use of pseudoinverses when solving general systems of nonlinear equations, while Ben-Israel [1966] has made similar use of pseudoinverses for solving nonlinear least squares problems. (See [Boggs, 1976a] for discussion of the convergence of the Ben-Israel iteration.)

2. Notation and Background. Unless otherwise stated, $\|\cdot\| = \|\cdot\|_2$ denotes the Euclidean vector norm $\|x\| = (x^Tx)^{1/2}$ or the corresponding matrix norm. $\mathbf{R}^{n\times p}$ stands for the set of real $n \times p$ matrices. $B(x, \delta)$ and $\bar{B}(x, \delta)$ denote, respectively, the open and closed balls of radius δ about $x \in \mathbf{R}^n$.

We shall make frequent use of pseudoinverses and the singular value decomposition theorem. For our present purposes, we may state the singular value decomposition theorem in the form:

(1) THEOREM. *For any $A \in \mathbf{R}^{n\times n}$ there exist orthogonal matrices U and $V \in \mathbf{R}^{n\times n}$ and scalars $\sigma_1, \dots, \sigma_n \in [0, \infty)$ such that $A = USV^T$, where $S = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ is a diagonal matrix having $\sigma_1, \dots, \sigma_n$ on the main diagonal. If the singular values $\sigma_1, \dots, \sigma_n$ are ordered so that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n \geqslant 0$, then they are unique. Moreover, if there are k distinct singular values $\sigma_{j_1}, \dots, \sigma_{j_k}$ with $j_0 = 0, j_k = n$, and $\sigma_i = \sigma_{j_l}$ for $j_{l-1} < i \leqslant j_l$, and if U and V are correspondingly partitioned as $U = [U_1U_2 \cdots U_k]$ and $V = [V_1V_2 \cdots V_k]$ with $U_l, V_l \in \mathbf{R}^{n\times(j_l-j_{l-1})}$, then the matrices $\sigma_{j_l}U_lV_l^T$ are unique, $1 \leqslant l \leqslant k$.* □

The pseudoinverse may be defined as follows. For any scalar $\sigma \in \mathbf{R}$, let

$$\sigma^+ = \begin{cases} 1/\sigma & \text{if } \sigma \neq 0, \\ 0 & \text{if } \sigma = 0. \end{cases}$$

The pseudoinverse S^+ of a diagonal matrix $S = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ is then defined by $S^+ \equiv \operatorname{diag}(\sigma_1^+, \dots, \sigma_n^+)$. Finally, if $A \in \mathbf{R}^{n\times n}$ and the notation of Theorem (1) holds, then $A^+ \equiv VS^+U^T = \Sigma_{l=1}^k \sigma_{j_l}^+V_lU_l^T$. (For more information on the singular value decomposition, see [Lawson and Hanson, 1974] or [Stewart, 1973]; for more on the pseudoinverse, see [Rao and Mitra, 1971] as well.)

We shall write $J(x)$ for the Jacobian matrix $f'(x)$ of f at x. Often we shall assume that $J(x)$ is locally Lipschitz continuous, i.e., that for each point $z \in \mathbf{R}^n$ there exists a constant γ and a neighborhood N of z such that

$$\|J(x) - J(y)\| \leqslant \gamma\|x - y\| \tag{2}$$

for all $x, y \in N$.

It will prove interesting below to compare our new existence theorem with the Kantorovich theorem. For ease of reference we therefore state the latter, following Ortega and Rheinboldt [1970, p. 421], as:

(3) THEOREM. *With f as above, assume* (2) *holds on a convex set $D_0 \subset \mathbf{R}^n$. Suppose for some $x_0 \in D_0$ that $\|J(x_0)^{-1}\| \leqslant \beta$ and $\alpha = \beta\gamma\eta \leqslant 1/2$, where $\eta \geqslant$*

$\|J(x_0)^{-1}f(x_0)\|$. *Let* $t^* = (\beta\gamma)^{-1}[1 - (1 - 2\alpha)^{1/2}]$ *and* $t^{**} = (\beta\gamma)^{-1}[1 + (1 - 2\alpha)^{1/2}]$ *and assume* $\bar{B}(x_0, t^*) \subset D_0$. *Then the Newton iterates*

$$x_{k+1} = x_k - J(x_k)^{-1}f(x_k) \tag{4}$$

are well defined, remain in $\bar{B}(x_0, t^*)$, *and converge to a zero* x^* *of* f *which is unique in* $B(x_0, t^{**}) \cap D_0$. *Moreover,*

$$\|x_k - x^*\| \leqslant (\beta\gamma 2^k)^{-1}(2\alpha)^{(2^k)}. \quad \square$$

We need below to be assured of the existence (and uniqueness) of solutions to certain differential equations. The following theorem (which follows easily from Theorems 1.2—the Cauchy-Peano existence theorem—and 2.2 of [Coddington and Levinson, 1955]) suffices for our purposes.

(5) THEOREM. *If* $F\colon \mathbf{R}^n \longrightarrow \mathbf{R}^n$ *is continuous, then for each* $x_0 \in \mathbf{R}^n$ *and* $t_0 \in \mathbf{R}$ *there exists a continuously differentiable function* $x\colon \mathbf{R} \longrightarrow \mathbf{R}^n$ *such that*

$$x(t_0) = x_0 \quad \textit{and} \tag{6a}$$

$$x'(t) = F(x(t)) \quad \textit{for all } t \in \mathbf{R}. \tag{6b}$$

Moreover, if F *is locally Lipschitz continuous, then the solution* $x(t)$ *of* (6) *is unique.* $\square$

3. Modifying Singular Values of $J(x)$.

The region of convergence of Newton's method (4) may often be enlarged by the introduction of appropriate damping factors $\lambda_k > 0$, in which case the iteration becomes

$$x_{k+1} = x_k - \lambda_k J(x_k)^{-1}f(x_k). \tag{7}$$

This amounts to Euler's method applied to the differential equation

$$x(0) = x_0, \tag{8a}$$

$$x'(t) = -J(x)^{-1}f(x), \tag{8b}$$

which, following Gavurin [1958], we call the "continuous analogue" of (4). This is of interest because, for fixed $t > 0$, $x_{k+1} \longrightarrow x(t)$ as $\max\{\lambda_i \mid 0 \leqslant i \leqslant k\} \longrightarrow 0$ with $\Sigma_{i=0}^k \lambda_i = t$. If $J(x)$ is singular then (4), (7), and (8) are undefined, while if $J(x)$ is nearly singular, then numerical attempts to compute (4) or (7) or to solve (8) encounter serious difficulties. We could make (4), (7), and (8) well defined by changing $J(x)^{-1}$ to $J(x)^+$, but $J(x)^+$ is discontinuous at—and unbounded near—points x where $J(x)$ changes rank. Thus it is much more appealing theoretically to modify the singular values of $J(x)$ to produce a *continuous* substitute $\hat{J}^+(x)$ for $J(x)^{-1}$. We shall do this as follows. Given $A \in \mathbf{R}^{n \times n}$ with singular value decomposition $A = USV^T = \Sigma_{l=1}^k \sigma_{j_l} U_l V_l^T$ as in Theorem (1) and $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, let $\hat{\sigma}_j$ denote the modified form of σ_j, let $\hat{S} = \operatorname{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$, and let $\hat{A} = U\hat{S}V^T = \Sigma_{j=1}^k \hat{\sigma}_{j_l} U_l V_l^T$. Although the notation suggests that $\hat{\sigma}_j$ should depend only on σ_j, in fact we shall allow $\hat{\sigma}_j$ to depend on all of $\sigma_1, \ldots, \sigma_n$. Specifically, for any $\delta \geqslant 0$ and $A' = U'S'V'^T$ with $\|A - A'\| \leqslant \delta$ we shall require the choice of $\hat{\sigma}_j$ to be such that for some tolerance

$\epsilon > 0$ and all j and k, $1 \leqslant j \leqslant n$, $1 \leqslant k \leqslant n$,

(9a) $$0 \leqslant \hat{\sigma}_j^+ \leqslant 1/\epsilon,$$

(9b) $$|\hat{\sigma}_j^+ - \hat{\sigma}_k'^+| = O(\delta + |\sigma_j - \sigma_k'|),$$

(9c) $$\sigma_j = \sigma_k \Rightarrow \hat{\sigma}_j^+ = \hat{\sigma}_k^+, \quad \text{and}$$

(9d) $$\hat{\sigma}_j^+ = O(\sigma_j).$$

At times we shall also require

(9e) $$0 \leqslant \sigma_j \sigma_j^+ \leqslant 1.$$

(10) LEMMA. *With the above notation, if* $\sigma_j \neq \sigma_k'$, *then*

(11) $$|u_j^T u_k'| \leqslant \delta / |\sigma_j - \sigma_k'|.$$

Proof. From $u_j^T A = \sigma_j v_j^T$ and $A' v_k' = \sigma_k' u_k'$ we obtain $\sigma_j v_j^T v_k' = u_j^T A v_k'$ and $\sigma_k' u_j^T u_k' = u_j^T A' v_k'$, whence

(12a) $$\sigma_k' u_j^T u_k' - \sigma_j v_j^T v_k' = u_j^T (A' - A) v_k'.$$

Similarly, since $Av_j = \sigma_j u_j$ and $u_k'^T A' = \sigma_k' v_k'^T$, we obtain

(12b) $$-\sigma_j u_j^T u_k' + \sigma_k' v_j^T v_k' = u_k'^T (A' - A) v_j.$$

Adding σ_k' times (12a) to σ_j times (12b), we have

$$(\sigma_k'^2 - \sigma_j^2) u_j^T u_k' = \sigma_k' u_j^T (A' - A) v_k' + \sigma_j u_k'^T (A' - A) v_j.$$

Since u_j, v_j, u_k', and v_k' are unit vectors and $\|A' - A\| \leqslant \delta$, we thus have $|u_j^T u_k'| \leqslant \delta(\sigma_k' + \sigma_j)/|\sigma_k'^2 - \sigma_j^2|$, whence (11) follows. □

(More generally, if $M, E \in \mathbf{C}^{n \times p}$ are complex $n \times p$ matrices and x, y are unit right singular vectors of M and $M + E$ with corresponding distinct singular values λ and $\mu \geqslant 0$ and unit left singular vectors $\widetilde{x}$ and $\widetilde{y}$, respectively, then similar reasoning shows that $(\mu^2 - \lambda^2) y^H x = \mu \widetilde{y}^H E x + \lambda y^H E^H \widetilde{x}$, whence again $|y^H x| \leqslant \|E\|/|\lambda - \mu|$.)

We may now prove that $\hat{A}^+$ is a Lipschitz continuous function of A:

(13) THEOREM. *If* (9b)–(9d) *hold, then*

(14) $$\|\hat{A}^+ - \hat{A}'^+\| = O(\|A - A'\|).$$

Proof. We shall show for any $A \in \mathbf{R}^{n \times n}$ that (14) holds whenever $\delta \equiv \|A - A'\|$ is sufficiently small, say $\delta < \delta_0(A)$, where $O(\delta)$ is independent of A. A simple compactness argument then shows that (14) holds no matter how large δ is.

It suffices to show for arbitrary j, $1 \leqslant j \leqslant n$, that $\|(\hat{A}^+ - \hat{A}'^+) u_j\| = O(\delta)$. Since $\hat{A}^+ u_j = V \hat{S}^+ U^T u_j = \hat{\sigma}_j^+ V U^T u_j$, we have

$$\|(\hat{A}^+ - \hat{A}'^+) u_j\| \leqslant \|V'(\hat{\sigma}_j^+ I - \hat{S}'^+) U'^T u_j\| + \hat{\sigma}_j^+ \|(VU^T - V'U'^T) u_j\|,$$

whence we need only show that

(15) $$\|(\hat{\sigma}_j^+ I - \hat{S}'^+) U'^T u_j\| = O(\delta) \quad \text{and}$$

$$\hat{\sigma}_j^+ \|(VU^T - V'U'^T)u_j\| = O(\delta). \tag{16}$$

To demonstrate (15), it is enough to show for each k, $1 \leqslant k \leqslant n$, that

$$|(\hat{\sigma}_j^+ - \hat{\sigma}_k'^+)u_k'^T u_j| = O(\delta). \tag{17}$$

We may assume that the singular values are arranged in decreasing order: $\sigma_1 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ and $\sigma_1' \geqslant \cdots \geqslant \sigma_n' \geqslant 0$, whence (by Theorem 6.6 of [Stewart, 1973]) $|\sigma_j - \sigma_j'| \leqslant \delta$. If $\sigma_j = \sigma_k$, then (17) follows from (9b), (9c). Otherwise we may assume $\delta < |\sigma_j - \sigma_k|/2$, whence (9b) becomes $|\hat{\sigma}_j^+ - \hat{\sigma}_k'^+| = O(|\sigma_j - \sigma_k|)$ and (17) follows from Lemma (10).

If $\sigma_j = 0$, then (9d) implies $\hat{\sigma}_j^+ = 0$, whence (16) holds. Otherwise, since $u_j^T A = \sigma_j u_j^T UV^T$, we have

$$(A - A')^T u_j = (VSU^T - V'S'U'^T)u_j = V'(\sigma_j I - S')U'^T u_j + \sigma_j(VU^T - V'U'^T)u_j,$$

whence

$$\|(VU^T - V'U'^T)u_j\| \leqslant \frac{1}{\sigma_j}[\|(A - A')^T u_j\| + \|V'(\sigma_j I - S')U'^T u_j\|]$$
$$\leqslant \frac{1}{\sigma_j}[\delta + \|(\sigma_j I - S')U'^T u_j\|].$$

Lemma (10) thus implies $\|(VU^T - V'U'^T)u_j\| \leqslant O(\delta)/\sigma_j$, which, together with (9d), yields (16). □

(Note that if A and A' are symmetric, then we may substitute the eigendecomposition for the singular value one, with the result that $V = U$ and $V' = U'$, whence the left-hand side of (16) vanishes and Theorem (13) holds without (9d). This has implications for minimization problems, but we shall not pursue them here.)

Suitable choices for $\hat{\sigma}^+$ include

$$\hat{\sigma}^+ = \min\{\sigma/\epsilon^2, 1/\sigma\} = \sigma/[\sigma^2 + \max\{0, \epsilon^2 - \sigma^2\}], \tag{18}$$

$$\hat{\sigma}^+ = \sigma/[\sigma^2 + \epsilon^2/4], \quad \text{and} \tag{19}$$

$$\hat{\sigma}^+ = \sigma/[\sigma^2 + \max\{0, \epsilon^2 - \sigma_n^2\}], \tag{20}$$

where σ_n is the smallest singular value of A. Choices (19) and (20) amount to the Levenberg-Marquardt modification $\hat{A}^+ = (A^TA + \mu I)^{-1}A^T$ (see [Levenberg, 1944] and [Marquardt, 1963]) with a special choice of the modification factor μ. If $A = USV^T$, then choice (18) may be similarly expressed as $\hat{A}^+ = (A^TA + M)^{-1}A^T$, where M is the positive semidefinite matrix $V \operatorname{diag}(d_1, \ldots, d_n)V^T$, with $d_j = \max\{0, \epsilon^2 - \sigma_j^2\}$. As such, this modification bears some resemblance to the modification which Murray [1972] has proposed for the Cholesky decomposition of a symmetric matrix. Choices (18) and (20) have the virtue of producing no modification when the smallest singular value $\sigma_n \geqslant \epsilon$, while choice (19) is a bit easier to compute.

It is readily verified that choices (18) and (19) satisfy (9). As for (20), it is easily seen that (9a) and (9c)–(9e) hold. To obtain (9b), note that if A' has singular values $\sigma_1' \geqslant \sigma_2' \geqslant \cdots \geqslant \sigma_n' \geqslant 0$ with $\|A - A'\| \leqslant \delta$, and if $\mu = \max\{0, \epsilon^2 - \sigma_n^2\}$ and $\mu' = \max\{0, \epsilon^2 - \sigma_n'^2\}$, then

$$\hat{\sigma}_j^+ - \hat{\sigma}_k'^+ = \frac{\sigma_j}{\sigma_j^2 + \mu} - \frac{\sigma_k'}{\sigma_k'^2 + \mu'} = \frac{(\sigma_j\sigma_k' - \mu)(\sigma_k' - \sigma_j) + \sigma_j(\mu' - \mu)}{(\sigma_j^2 + \mu)(\sigma_k'^2 + \mu')}.$$

Since $|\sigma_k - \sigma_k'| \leqslant \delta$, we have $|\sigma_k' - \sigma_j| \leqslant \delta + |\sigma_j - \sigma_k|$. We may assume $\delta \leqslant \epsilon/2$, whence $\mu = \mu' = 0$ if $\sigma_n \geqslant 3\epsilon/2$ and $|\mu - \mu'| \leqslant |\sigma_n^2 - \sigma_n'^2| = |(\sigma_n + \sigma_n')(\sigma_n - \sigma_n')| \leqslant 4\epsilon\delta$ otherwise. Since $\mu \leqslant \epsilon^2$, $\min\{\sigma_j^2 + \mu, \sigma_k'^2 + \mu'\} \geqslant \epsilon^2$, and (9a) holds, we thus find $|\hat{\sigma}_j^+ - \hat{\sigma}_k'^+| \leqslant (5\delta + |\sigma_j - \sigma_k'|)/\epsilon^2$, which establishes (9b) for (20).

We shall devote the remainder of this section to establishing and discussing an existence theorem based on integrating the differential equation

(21a) $$x(0) = x_0,$$

(21b) $$x'(t) = -\hat{J}^+(x)f(x).$$

Theorem (13) implies that $\hat{J}^+(x)$ is well behaved for suitable choices of $\hat{\sigma}^+$: $\hat{J}^+(x)$ is continuous and is locally Lipschitz continuous whenever $J(x)$ is likewise. Thus Theorem (5) applies to (21).

Now we prove the main result of this paper. While we allow $J(x)$ to be singular, we require a certain kind of nondegeneracy: we must assume that f and J are such that

(22) $$f(x)^T J(x)\hat{J}^+(x)f(x) \geqslant \theta \|f(x)\|^2$$

for some fixed $\theta > 0$ and all relevant $x \in \mathbf{R}^n$. We shall discuss this condition in more detail below. The following theorem rests heavily upon it.

(23) THEOREM. *If $f \in C^1(\mathbf{R}^n)$ and* (9a)–(9d) *and* (22) *hold, then for each $x_0 \in \mathbf{R}^n$ there exists a solution $x(t)$ to* (21). *Such a solution has an asymptote $x^* = \lim_{t\to\infty} x(t)$ with $f(x^*) = 0$. Moreover, the following bound holds*:

(24) $$\|x(t) - x^*\| \leqslant [\|f(x_0)\|/(\theta\epsilon)]\, e^{-\theta t}.$$

Proof. Fix x_0. As already remarked, the existence of $x(t)$ follows easily from Theorems (13) and (5).

Note that $\lim_{t\to\infty} f(x(t)) = 0$. Indeed, let $\phi(t) = \|f(x(t))\|^2$. Then $\phi'(t) = -2f^T J\hat{J}^+ f$, so (22) implies $\phi'(t) \leqslant -2\theta\|f(x(t))\|^2 = -2\theta\phi(t)$. Hence $\psi(t) \equiv \ln \phi(t)$ has $\psi'(t) \leqslant -2\theta$, so (for $t \geqslant 0$)

$$\psi(t) = \psi(0) + \int_0^t \psi'(\tau)\,d\tau \leqslant \psi(0) - 2\theta t$$

and

$$\|f(x(t))\|^2 = \phi(t) = e^{\psi(t)} \leqslant \|f(x_0)\|^2 e^{-2\theta t}.$$

Now we show that $\lim_{t\to\infty} x(t) = x^*$ exists and (24) holds. It suffices to show that

$$\|x(t_1) - x(t_2)\| \leqslant [\|f(x_0)\|/(\theta\epsilon)]\,|e^{-\theta t_2} - e^{-\theta t_1}|,$$

which follows from (9a), since

$$\|x'(t)\| = \|\hat{J}^+ f(x(t))\| \leqslant \|f(x(t))\|/\epsilon \leqslant (\|f(x_0)\|/\epsilon)e^{-\theta t},$$

whence

$$\|x(t_1) - x(t_2)\| = \left\| \int_{t_1}^{t_2} x'(\tau)\,d\tau \right\| \leqslant \left| \int_{t_1}^{t_2} \|x'(\tau)\|\,d\tau \right| \leqslant \frac{\|f(x_0)\|}{\epsilon} \left| \int_{t_1}^{t_2} e^{-\theta\tau}\,d\tau \right|.$$

Thus the sequence $x(t_1), x(t_2), x(t_3), \ldots$ is a Cauchy sequence for any choice of $t_1, t_2, \ldots$ with $\lim_{i\to\infty} t_i = +\infty$, whence $x^* = \lim_{t\to\infty} x(t)$ exists. By the continuity of f, $f(x^*) = \lim_{t\to\infty} f(x(t)) = 0$. □

It complicates the proof only slightly if Theorem (23) is restated in "semilocal" form; we state this form as a corollary:

(25) COROLLARY. *Suppose $f \in C^1(D)$, where $D \subset \mathbf{R}^n$, and assume that* (22) *holds on D. If $x_0 \in D$ is such that $\bar{B}(x_0, \|f(x_0)\|/(\theta\epsilon)) \subset D$, then the conclusion of Theorem* (23) *holds, $x^* \in D$, and $x(t) \in D$ for all $t \in$* $[0, \infty)$. □

While Theorems (3) and (23) are both existence theorems, they differ in a significant way. Whereas the nondegeneracy assumptions of (3) imply that $J(x)$ is nonsingular at each Newton iterate x_k, the corresponding assumption (22) of (23) allows $J(x)$ to be singular everywhere (as we shall see presently). This weaker nondegeneracy assumption is made at the cost of one of the prime conclusions of (3): the uniqueness of x^*. For example, if f: $\mathbf{R}^2 \longrightarrow \mathbf{R}^2$ is the linear mapping $f(x) = \left(\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right)x$, then (22) holds with $\theta = 1$ and x^* can be any point in the set $\{0\} \times \mathbf{R}$.

Note that (22) implies

$$(26) \qquad \|J(x)\hat{J}^+(x)f(x)\| \geqslant \theta\,\|f(x)\|.$$

On the other hand, if (9e) holds, then (26) implies $f^TJ\hat{J}^+f(x) \geqslant \theta^2\,\|f\|^2$. To see this, let $J = J(x)$ have singular value decomposition USV^T, whence

$$f^TJ\hat{J}^+f = (U^Tf)^TS\hat{S}^+(U^Tf) \geqslant (U^Tf)^T(S\hat{S}^+)^2(U^Tf) = \|J\hat{J}^+f\|^2 \geqslant \theta^2\,\|f\|^2.$$

Thus (22) and (26) are qualitatively the same, and we could have assumed (26). We have chosen (22) since it yields sharper bounds.

Let us see what (22) means if $f(x) = Ax - b$ is affine. We may assume that b lies in the column space of A, for otherwise at $x = A^+b$ we would have $\hat{J}^+(x)f(x) = 0$ with $f(x) \neq 0$, whence (22) could not hold. By the change of variables $y = x - A^+b$ we may thus arrange that $b = 0$, and hence $f(x) = Ax$. Let $A = USV^T$ be a singular value decomposition of A, with the singular values σ_j ordered so that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_\nu > 0 = \sigma_{\nu+1} = \cdots = \sigma_n$. If $\hat{\sigma}$ is given by (18), then $\sigma_j\hat{\sigma}_j^+ \geqslant \min\{1, \sigma_\nu^2/\epsilon^2\}$ for $j \leqslant \nu$, so if $g = (g_1, \ldots, g_n)^T = U^Tf(x)$, then $g_j = 0$ for $j > \nu$, and

$$f^TJ\hat{J}^+f(x) = f^TUS\hat{S}^+U^Tf = g^TS\hat{S}^+g \geqslant \|g\|^2\min\{1, \sigma_\nu^2/\epsilon^2\}.$$

Since $\|f(x)\| = \|g\|$ in this case, we thus see that (22) holds with $\theta = \min\{1, \sigma_\nu^2/\epsilon^2\}$.

Assumption (22) implies that only zeroes x^* of f can be critical points of the least squares function $\Phi(x) = \|f(x)\|^2$. But it implies more than this, at least when $J(x)$ is locally Lipschitz continuous, which we henceforth assume. In this case the zeroes of f form a connected set, and if $J(x^*)$ is nonsingular for some zero x^*, then this set consists exactly of x^*, i.e. f has a unique zero. Indeed, from Theorems (13) and (5) we see that the solution $x(t)$ of (21) and hence $x^* = \lim_{t\to\infty} x(t)$ are uniquely determined by $x_0 = x(0)$. Thus we may define X: $\mathbf{R}^n \longrightarrow \mathbf{R}^n$ by

$$(27) \qquad X(x_0) = x^*.$$

Note that $f \circ X \equiv 0$ and $X(x^*) = x^*$ for any zero x^* of f. Therefore, $X(\mathbf{R}^n) = f^{-1}(0)$, i.e. the range of X is the set of zeroes of f. The above claims about this set now follow from

(28) THEOREM. *The mapping X defined by* (27) *is continuous.*

Proof. Let $y_0 \in \mathbf{R}^n$ and $\zeta > 0$ be given: we must demonstrate the existence of $\delta > 0$ such that $X(B(y_0, \delta)) \subset B(X(y_0), \zeta)$. Let $y(t)$ solve $y'(t) = -\hat{J}^+(y)f(y)$ with $y(0) = y_0$. Using (24) and (13), it is easy to show that there are constants Γ and K such that if $x_0 \in B(y_0, 1)$ and $x(t)$ solves (21), then $\|x(t) - x^*\| \leqslant Ke^{-\theta t}$ and $\|x(t) - y(t)\| \leqslant \|x_0 - y_0\|e^{\Gamma t}$ for all $t \in [0, \infty)$ (with $x^* = X(x_0)$). Let t^* be large enough that $Ke^{-\theta t^*} < \zeta/4$, and let $\delta = \min\{1, \zeta e^{-\Gamma t^*}/2\} > 0$. Setting $y^* = X(y_0)$, we then find for $\|x_0 - y_0\| < \delta$ that

$$\begin{aligned}\|X(x_0) - X(y_0)\| &\leqslant \|x(t^*) - x^*\| + \|x(t^*) - y(t^*)\| + \|y(t^*) - y^*\| \\ &\leqslant 2Ke^{-\theta t^*} + \|x_0 - y_0\|e^{\Gamma t^*} \\ &< \zeta/2 + \zeta/2 = \zeta. \quad \square\end{aligned}$$

4. Practical Implications. Theorem (28) implies that if $f^{-1}(0)$ has at least two components (in particular, if f has at least two isolated zeroes), then (22) cannot hold. (Note that the existence of θ such that (22) holds does not depend on which value of $\epsilon > 0$ has been chosen, though the value of θ does, of course, depend on ϵ.) Thus, we may expect (22) to hold globally only for a small class of problems. However, it appears very likely that (22) would often hold in a region D (as in Corollary (25)) containing points x where $J(x)$ is singular or nearly so and thus that methods using $\hat{J}^+(x)$ instead of $J^{-1}(x)$ would enjoy a larger region of convergence.

Boggs [1971] advocates the use of A-stable integration techniques for numerically solving (8). His arguments suggest that weakly A-stable integration techinques (see [Boggs and Dennis, 1974]) would be appropriate for attacking (21) directly: such techniques aim to determine the asymptote x^* quickly without spending excessive time to compute $x(t)$ accurately. In practice, Boggs [1976b] has experienced numerical difficulties when $J(x)$ becomes singular or nearly so. Intended numerical experiments will hopefully indicate how much these problems can be alleviated by using $\hat{J}^+(x)$ in place of $J^{-1}(x)$.

The damped Newton's method (7) results when (8) is numerically integrated by

Euler's method with kth stepsize λ_k. By considering (21) in place of (7), we obtain a modified damped Newton's method $x_{k+1} = x_k - \lambda_k \hat{J}^+(x_k)f(x_k)$. While a proper choice of the damping factor λ_k surely makes this more robust than the undamped method

$$x_{k+1} = x_k - \hat{J}^+(x_k)f(x_k), \tag{29}$$

it is possible for (29) to state a theorem similar to (3) (but without the uniqueness assertion), as the following crude example illustrates.

(30) THEOREM. *Suppose $f\colon D_0 \longrightarrow \mathbf{R}^n$ is continuously differentiable and that* (2) *holds for $x, y \in D_0 \subset \mathbf{R}^n$. Suppose further that $x_0 \in D_0$ and $\theta \in [0, 1]$ are such that*

$$\alpha \equiv \frac{\gamma}{2\epsilon^2}\|f(x_0)\| + \sqrt{1-\theta^2} < 1, \tag{31}$$

$\bar{B}(x_0, t^*) \subset D_0$, and (26) *holds for $x \in B(x_0, t^*)$, where*

$$t^* = \frac{\|f(x_0)\|}{(1-\alpha)\epsilon}. \tag{32}$$

If (9a), (9e) *hold, then the iterates x_k generated by* (29) *are well defined, remain in D_0, and converge at least Q-linearly to a zero $x^* \in D_0$ of f. Moreover,*

$$\|x_k - x^*\| \leqslant t^*\alpha^k. \tag{33}$$

Proof. Below we show that

$$\|f(x) - J(x)\hat{J}^+(x)f(x)\| \leqslant \sqrt{1-\theta^2}\|f(x)\| \tag{34}$$

for those $x \in D_0$ of interest. Since (9a) implies

$$\|x_{k+1} - x_k\| \leqslant \|\hat{J}^+(x_k)f(x_k)\| \leqslant \|f(x_k)\|/\epsilon, \tag{35}$$

we thus obtain the estimate

$$\begin{aligned} \|f(x_{k+1})\| &\leqslant \|f(x_k) - J(x_k)\hat{J}^+(x_k)f(x_k)\| + \frac{\gamma}{2}\|x_{k+1} - x_k\|^2 \\ &\leqslant \sqrt{1-\theta^2}\|f(x_k)\| + \frac{\gamma}{2\epsilon^2}\|f(x_k)\|^2 \\ &\leqslant \left(\sqrt{1-\theta^2} + \frac{\gamma\|f(x_k)\|}{2\epsilon^2}\right)\|f(x_k)\|. \end{aligned} \tag{36}$$

Using (31), we find by induction on k that

$$\|f(x_{k+1})\| \leqslant \alpha\|f(x_k)\|, \tag{37a}$$

whence

$$\|f(x_k)\| \leqslant \|f(x_0)\|\alpha^k. \tag{37b}$$

Combining this with (35), we find

$$\|x_k - x_0\| \leqslant \frac{1-\alpha^k}{1-\alpha}\frac{\|f(x_0)\|}{\epsilon},$$

whence $x_k \in B(x_0, t^*)$ for all k; moreover, we see that x^* exists and (33) holds. Let $c = \max\{ \|f'(x)\| : x \in \bar{B}(x_0, t^*)\}$. Since $x_k, x^* \in \bar{B}(x_0, t^*)$, we have $\|f(x_k)\| \leqslant c\|x_k - x^*\|$. Together with (35) and (37a), this implies

$$\|x_{k+1} - x^*\| \leqslant \sum_{j=k+1}^{\infty} \|x_j - x_{j+1}\| \leqslant \frac{1}{\epsilon} \sum_{j=k+1}^{\infty} \|f(x_j)\|$$

$$\leqslant \frac{\|f(x_k)\|}{\epsilon} \sum_{j=1}^{\infty} \alpha^j \leqslant \left(\frac{c\alpha}{\epsilon(1-\alpha)}\right) \|x_k - x^*\|,$$

which establishes the Q-linearity of the convergence. Now it only remains to establish (34).

Without loss of generality $J(x) = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Writing $f = f(x) = (f_1, \ldots, f_n)^T$, we see from (26) that $\|J\hat{J}^+ f\| = T\|f\|$ for some $T \in [\theta, 1]$ and hence

$$\|J\hat{J}^+ f\|^2 = \sum_{j=1}^{n} (\sigma_j \hat{\sigma}_j^+ f_j)^2 = T^2 \sum_{j=1}^{n} f_j^2,$$

whence

$$\|f - J\hat{J}^+ f\|^2 = \sum_{j=1}^{n} (1 - \sigma_j \hat{\sigma}_j^+)^2 f_j^2 = (1 + T^2)\|f\|^2 - 2 \sum_{j=1}^{n} (\sigma_j \hat{\sigma}_j^+) f_j^2. \tag{38}$$

From (9e) we obtain

$$\sum_{j=1}^{n} (\sigma_j \hat{\sigma}_j^+) f_j^2 \geqslant \sum_{j=1}^{n} (\sigma_j \hat{\sigma}_j^+)^2 f_j^2 = T^2 \|f\|^2,$$

which with (38) implies $\|f - J\hat{J}^+ f\|^2 \leqslant (1 - T^2)\|f\|^2 \leqslant (1 - \theta^2)\|f\|^2$, whence (34) follows. □

As can be seen from (36), the bounds (33) and (37) are not optimal, and a value smaller than (32) would suffice for t^*. However, a better factor β_k than α^k based on (36) would still satisfy $\beta_k > (1 - \theta^2)^{k/2}$.

As (34) suggests, even if f is linear the iterates generated by (29) may converge only Q-linearly to x^*. The speed of convergence depends strongly on ϵ: in the linear case, for instance, the iterates converge in one step if ϵ is no larger than the smallest nonzero singular value $\|J^+\|^+$ and $\hat{\sigma}^+$ is computed by (18) or (20). Moreover, the factor $1/(\theta\epsilon)$ which appears in (24) may change with ϵ: in the linear case, if (18) is used, ϵ_1 and ϵ_2 are two choices for ϵ, θ_1 and θ_2 are the corresponding largest possible choices for θ in (22), and $\epsilon_2 > \epsilon_1 \geqslant \|J^+\|^+$, then $\theta_2 = (\epsilon_1/\epsilon_2)^2\theta_1$ and $1/(\theta_2\epsilon_2) = [1/(\theta_1\epsilon_1)](\epsilon_2/\epsilon_1)$. From this standpoint, the tolerance ϵ should be chosen as small as possible. In practice, the accuracy to which f is computed implies a lower bound on ϵ. Moreover, the smaller ϵ is, the closer the search direction $-\hat{J}^+ f(x)$ can come to orthogonality with the gradient $2J^T f(x)$ of $\phi(x) = \|f(x)\|^2$; this phenomenon can severely hamper the numerical solution of $f(x^*) = 0$, so ϵ should not be too small. The intended numerical experiments should indicate how crucial the choice of ϵ is.

Choices (18)–(20) for $\hat{\sigma}^+$ all behave similarly for $\sigma << \epsilon$ or $\sigma >> \epsilon$: the relative difference between these choices remains bounded. Computationally, we should, therefore, not expect major differences between the performances to which they lead. Since the small singular values contribute little to $f^TJ\hat{J}^+f/\|f\|^2$, we should expect the same to be true of any other choice of $\hat{\sigma}^+$ which satisfies (9) along with $\sigma\hat{\sigma}^+ \longrightarrow 1$ as $\sigma \longrightarrow +\infty$.

Once $J(x)$ and $f(x)$ are known, $\hat{J}^+(x)f(x)$ can be computed with $\hat{\sigma}^+$ given by (18) or (20) in $(4/3)n^3 + O(n^2)$ multiplications (and a similar number of additions), as opposed to $(1/3)n^3 + O(n^2)$ multiplications for computing $J(x)^{-1}f(x)$ by Gaussian elimination (assuming that $J(x)$ is nonsingular without special structure); when (19) is used, $\hat{J}^+(x)f(x)$ may be computed in $(2/3)n^3 + O(n^2)$ multiplications; thus $\hat{J}^+$ may be introduced with only a minor increase in the cost of an iteration. Golub and Reinsch [1970] show how the singular value decomposition of a matrix may be efficiently and accurately computed. The above operation count for (18) assumes that the factors U and V of the singular value decomposition USV^T of $J = J(x)$ are not explicitly computed, but rather that U^Tf is accumulated and V is maintained in factored form.

Acknowledgement. The author gratefully acknowledges helpful comments by John E. Dennis, Jr. on a draft of this paper.

National Bureau of Economic Research, Inc.
Computer Research Center for Economics and Management Science
575 Technology Square
Cambridge, Massachusetts 02139

A. BEN-ISRAEL (1966), "A Newton-Raphson method for the solution of systems of equations," *J. Math. Anal. Appl.*, v. 15, pp. 243–252. MR **34** #5273.

P. T. BOGGS (1970), *The Solution of Nonlinear Operator Equations by A-Stable Integration Techniques*, Doctoral thesis, Cornell University, Ithaca, New York; Report TR70–72, Computer Science Dept., Cornell University, Ithaca, N. Y.

P. T. BOGGS (1971), "The solution of nonlinear systems of equations by *A*-stable integration techniques," *SIAM J. Numer. Anal.*, v. 8, pp. 767–785. MR **45** #6179.

P. T. BOGGS (1976a), "The convergence of the Ben-Israel iteration for nonlinear least squares problems," *Math. Comp.*, v. 30, pp. 512–522.

P. T. BOGGS (1976b), Private communication.

P. T. BOGGS & J. E. DENNIS, JR. (1976), "A stability analysis for perturbed nonlinear iterative methods," *Math. Comp.*, v. 30, pp. 199–215. MR **52** #16007.

E. A. CODDINGTON & N. LEVINSON (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York. MR **16**, 1022.

R. FLETCHER (1970), "Generalized inverses for nonlinear equations and optimization," *Numerical Methods of Nonlinear Algebraic Equations* (P. RABINOWITZ, Editor), Gordon and Breach, London and New York, pp. 75–85. MR **49** #8328.

M. K. GAVURIN (1958), "Nonlinear functional equations and continuous analogues of iteration methods," *Izv. Vysš. Učebn. Zaved. Matematika*, no. 5 (6), pp. 18–31; English transl., Report 68–70, Computer Science Center, University of Maryland, College Park, Md. MR **25** #1380.

G. H. GOLUB & C. REINSCH (1970), "Singular value decomposition and least squares solutions," *Numer. Math.*, v. 14, pp. 403–420; Also: contribution I/10 of *Handbook for Automatic Computation*. Vol. II, *Linear Algebra* (J. H. WILKINSON & C. REINSCH, Editors), Springer-Verlag, Berlin and New York, 1971.

C. L. LAWSON & R. J. HANSON (1974), *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N. J. MR **51** #2270.

K. LEVENBERG (1944), "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, v. 2, pp. 164–168. MR **6**, 52.

D. W. MARQUARDT (1963), "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, v. 11, pp. 431–441. MR **27** #3040.

W. MURRAY (1972), "A numerically stable modified Newton method based on Cholesky factorization," §4.9 (pp. 64–68) of *Numerical Methods for Unconstrained Optimization* (W. MURRAY, Editor), Academic Press, New York.

J. M. ORTEGA & W. C. RHEINBOLDT (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York and London. MR **42** #8686.

C. R. RAO & S. K. MITRA (1971), *Generalized Inverse of Matrices and Its Applications*, Wiley, New York. MR **49** #2780.

G. W. STEWART (1973), *Introduction to Matrix Computations*, Academic Press, New York.