On Measures of Ill-Conditioning for Nonlinear Equations*

By Werner C. Rheinboldt

Abstract. Let x^* be a solution of the nonlinear equation Fx = b on a normed linear space and y^* that of a perturbed equation Gx = c. Estimates for the relativized error between x^* and y^* are derived which extend a known estimate for the corresponding matrix case. The condition number of F depends now also on the domain, and special considerations are needed to determine the existence of the solution of the perturbed equation. For differentiable F, when the domain shrinks to a point, the condition number of F is shown to reduce to that of the derivative at that point.

1. Introduction. Let

be a linear equation in
$$\mathbb{R}^n$$
 with nonsingular $A \in L(\mathbb{R}^n)$, and

$$Bx = c$$

a perturbation of (1.1) with a matrix $B \in L(\mathbb{R}^n)$ which is close to A in the sense that

Ax = b

$$||A^{-1}|| ||B - A|| < 1.$$

Then, it is well known that B is also nonsingular, and that, for $b \neq 0$, the solutions $x^* = A^{-1}b$, $y^* = B^{-1}c$ satisfy the estimate

(1.4)
$$\frac{\|x^* - y^*\|}{\|x^*\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|B - A\|/\|A\|} \left[\frac{\|B - A\|}{\|A\|} + \frac{\|b - c\|}{\|b\|} \right].$$

Here,

(1.1)

(1.5)
$$\kappa(A) = \|A\| \, \|A^{-1}\|$$

is the condition number of A under the particular norm.

This result has many uses in numerical linear algebra. It also points to the condition number of A as an indicator for the sensitivity of the solution of (1.1) under small changes of the matrix on the right-hand side.

In Section 2 below, we show that the estimate (1.4) allows a direct generalization to nonlinear equations on normed linear spaces. In particular, a condition number for nonlinear mappings is introduced which now depends also on the domain. Moreover, special considerations are needed to determine the existence of the solution of the perturbed equation in this domain. For shrinking domains the condition number decreases, and this suggests a study of the asymptotic behavior when the domain reduces to a point. This is the topic of Section 3, where also some other related results are obtained.

*This work was supported in part by the National Science Foundation under Grant GJ-35568X.

Received November 12, 1974.

AMS (MOS) subject classifications (1970). Primary 65H10; Secondary 65J05, 47H15.

2. Basic Results. Throughout the discussion, let X, Y denote real, normed linear spaces and L(X, Y) the space of bounded, linear operators from X into Y. For any mapping $F: D \subset X \longrightarrow Y$ and closed subset $C \subset D$, we introduce the greatest lower bound

(2.1a)
$$\mu(F, C) = \sup \{t \in [0, \infty); \|Fx - Fy\| \ge t \|x - y\| \quad \forall x, y \in C\}$$

and the least upper bound or Lipschitz norm

(2.1b)
$$\nu(F, C) = \inf \{t \in [0, \infty] ; \|Fx - Fy\| \le t \|x - y\| \quad \forall x, y \in C \}.$$

Note that in (2.1b) the value $t = \infty$ is specifically included.

For any finite-dimensional affine mapping $Fx = Ax - b \quad \forall x \in \mathbb{R}^n$, with nonsingular $A \in L(\mathbb{R}^n)$ and $C = \{x \in \mathbb{R}^n; \|x\| \le r\}, r > 0$, or $C = \mathbb{R}^n$, we obtain $\nu(F, C) = \|A\|$, and $\mu(F, C) = \|A^{-1}\|^{-1}$, using the induced norms on $L(\mathbb{R}^n)$.

This example suggests the definition of the condition number

(2.2)
$$\kappa(F, C) = \begin{cases} \nu(F, C)/\mu(F, C) & \text{if } 0 < \mu(F, C), \nu(F, C) < \infty, \\ \infty & \text{otherwise,} \end{cases}$$

which reduces to (1.5) in our affine case. Note that always $\kappa(F, C) \ge 1$ and that $\kappa(F, C) < \infty$ only if $\mu(F, C) > 0$, which in turn implies the injectivity of F on C.

The following theorem contains the above-mentioned extension of the estimate (1.4) to nonlinear equations

$$(2.3) Fx = b,$$

$$Gx = c$$

with mappings F, G from X to Y which are close to each other in the sense that on some set C the difference mapping

$$(2.5) E: C \subset X \longrightarrow Y, \quad Ex = Fx - Gx \quad \forall x \in C,$$

has a sufficiently small Lipschitz norm.

THEOREM 2.1. Let $F: D_F \subset X \to Y$ and $G: D_G \subset X \to Y$ be given and $C \subset D_F \cap D_G$ be a closed set on which $\kappa(F, C) < \infty$ and

(2.6)
$$\nu(E, C) < \mu(F, C).$$

If solutions x^* , $y^* \in C$ of Eqs. (2.3), (2.4), respectively, exist, then they are unique in C and, with any $x^0 \in C$, $x^0 \neq x^*$, the estimate

$$(2.7) \quad \frac{\|x^* - y^*\|}{\|x^* - x^0\|} \leq \frac{\kappa(F, C)}{1 - \kappa(F, c)\nu(E, C)/\nu(F, C)} \left[\frac{\|b - c\|}{\|b - Fx^0\|} + \frac{\nu(E, C)}{\nu(F, C)} + \frac{\|Ex^0\|}{\|b - Fx^0\|} \right]$$

holds.

Proof. For any $x, y \in C$ we have

$$||Gx - Gy|| \ge ||Fx - Fy|| - ||Ex - Ey|| \ge [\mu(F, C) - \nu(E, C)]||x - y||;$$

and, hence,

WERNER C. RHEINBOLDT

(2.8)
$$\mu(G, C) \ge \mu(F, C) - \nu(E, C) > 0.$$

This proves the injectivity of G on C and hence the uniqueness of the solutions x^* , y^* . Now observe that

$$\mu(G, C) \|x^* - y^*\| \le \|Gx^* - Gy^*\| \le \|Fx^* - Gy^*\| + \|Ex^0 - Ex^*\| + \|Gx^0 - Fx^0\|$$
$$\le \|b - c\| + \nu(E, C)\|x^* - x^0\| + \|Ex^0\|,$$

and

(2.9)
$$||b - Fx^{0}|| \leq \nu(F, C)||x^{*} - x^{0}||.$$

From these estimates it follows that

$$\frac{\|x^* - y^*\|}{\|x^* - x^0\|} \leq \frac{\nu(F, C)}{\mu(G, C)} \left[\frac{\|b - c\|}{\|b - Fx^0\|} + \frac{\nu(E, C)}{\nu(F, C)} + \frac{\|Ex^0\|}{\|b - Fx^0\|} \right],$$

while (2.8) shows that

$$\frac{\nu(F, C)}{\mu(G, C)} \leq \frac{\kappa(F, C)}{1 - \kappa(F, C)\nu(E, C)/\nu(F, C)}.$$

Together this proves (2.7).

As in the matrix case, the estimate (2.7) shows that the condition number (2.2) of F represents an indicator for the sensitivity of a solution of Eq. (2.3) to small changes of F and b. In particular, let $\kappa(F, C) < \infty$, and suppose that by some numerical process we have obtained an approximation $y^* \in C$ of the exact solution $x^* \in C$ of (2.3). Then Theorem 2.1 applies and (2.7) reduces to the a posteriori estimate

(2.10)
$$\frac{\|x^* - y^*\|}{\|x^* - x^0\|} \leq \kappa(F, C) \frac{\|r\|}{\|b - Fx^0\|}, \quad r = b - Fy^*.$$

In other words, even if the norm of the residual r is small, the relativized error between x^* and y^* may be large if F has a large condition number.

In the case of the matrix equations (1.1), (1.2), with nonsingular $A \in L(\mathbb{R}^n)$ and $b \neq 0$, we may take, say, $C = \mathbb{R}^n$ and $x^0 = 0$. Then, (2.6) is exactly the condition (1.3) and (2.7) reduces to (1.4). However, note that in Theorem 2.1 the existence of the solutions x^* and y^* in C had to be assumed. In the matrix case, this follows immediately from the injectivity of the mappings.

More generally, under the assumptions on F of Theorem 2.1, the restriction $F_C = F|C$ of F to C has an inverse $F_C^{-1}: FC \to X$ with

$$\|F_C^{-1}u - F_C^{-1}v\| \leq \frac{1}{\mu(F, C)} \|u - v\| \quad \forall u, v \in FC.$$

Hence, if

$$(2.11) EC + c \subset FC,$$

then H: $C \to X$, $Hx = F_C^{-1}(Ex + c) \quad \forall x \in C$, is well defined and satisfies

$$||Hx - Hy|| \leq \frac{\nu(E, C)}{\mu(F, C)} ||x - y|| \quad \forall x, y \in C$$

as well as $HC \subset C$. Therefore, by the contraction theorem, H has a unique fixed point in C which evidently is a solution of Gx = c.

The condition (2.11) certainly holds when FC = Y. In the special case C = X, we then obtain the following result:

COROLLARY 2.2. Let $F: X \to Y$ satisfy $\kappa(F, X) < \infty$ and FX = Y, and consider any $G: X \to Y$ such that (2.6) holds with C = X. Then, for any $b, c \in Y$, Eqs. (2.3), (2.4) have unique solutions which satisfy the estimate (2.7).

In particular, consider any linear operator $A \in L(X, Y)$ for which the inverse $A^{-1} \in L(Y, X)$ exists. Then, for any $b \in X$, the affine mapping $Fx = Ax - b \quad \forall x \in X$ satisfies $\kappa(F, X) < \infty$, and Corollary 2.2 provides the well-known extension of the estimate (1.4) to linear equations in infinite-dimensional normed linear spaces.

As a simple example of a mapping for which the condition number is readily estimated, consider $F: C \subset X \longrightarrow X$, $Fx = x - Hx \quad \forall x \in C$, where $H: C \longrightarrow X$ is a contraction with constant $\alpha \in (0, 1)$. Then, we have

$$||Fx - Fy|| = ||(x - y) - (Hx - Hy)|| \begin{cases} \leq (1 + \alpha)||x - y|| \\ \geq (1 - \alpha)||x, y|| \end{cases} \quad \forall x, y \in C,$$

and hence, $\kappa(F, C) \leq (1 + \alpha)/(1 - \alpha)$. Thus, if $x^* \in C$ is a fixed point of H and $y^* \in C$ any approximation of x^* , then, provided the choice $x^0 = 0$ is possible, (2.10) takes the form

$$\frac{\|x^* - y^*\|}{\|x^*\|} \le \frac{1 + \alpha}{1 - \alpha} \frac{\|Hy^* - y^*\|}{\|H0\|}$$

As another example, consider the two-point boundary value problem

$$(2.12) -u''(t) = f(t, u(t)), 0 < t < 1, u(0) = u(1) = 0,$$

with real, continuous f and $f_u(t, u) \ge \gamma > -\pi^2$ on $[0, 1] \times R^1$. More specifically, let X be the subspace of the Sobolev space $W_2^1[0, 1]$ consisting of all $u \in W_2^1[0, 1]$ with u(0) = u(1) = 0. Then there exists a unique mapping $F: X \longrightarrow X$ such that, for any $u \in X$,

$$(Fu, v) = B(u, v) \equiv \int_0^1 [u'(t)v'(t) - f(t, u(t))v(t)] dt \quad \forall v \in X,$$

where (\cdot, \cdot) is the inner product of $W_2^1[0, 1]$. The solutions of Fu = 0 are the (generalized) solutions of (2.12).

The mapping F is uniformly monotone,

 $(Fu - Fv, u - v) \ge \alpha ||u - v||_X^2 \quad \forall u, v \in X,$

as well as Lipschitz continuous on any ball $C_r = \{x \in X; \|x\|_X \leq r\}, r > 0$,

$$\|Fu - Fv\|_X \leq \beta_r \|u - v\|_X \quad \forall u, v \in C_r$$

(see, e.g., [3]). Thus, we have $\kappa(F, C_r) \leq \beta_r / \alpha$.

If $M \subset X$ is a subspace with basis $\varphi_1, \ldots, \varphi_m \in X$, then the Ritz-Galerkin approximation of Fu = 0 has the form

$$F_M x = A_M x + H_M x = 0 \quad \forall x \in M,$$

$$A_M = (B(\varphi_i, \varphi_j), i, j = 1, \dots, m), \quad H_M x = \left(\int_0^1 f\left(t, \sum_{j=1}^n x_j \varphi_j(t)\right) \varphi_i(t) dt, i = 1, \dots, m \right).$$

Interestingly, if, say, B-splines over a partition of [0, 1] of norm h are used as basis functions, then under appropriate conditions it follows, [1], that

$$\gamma_1 h^{-2} x^T x \ge x^T A_M x \ge \gamma_2 x^T x \quad \forall x \in M, \ \gamma_1 \ge \gamma_2 > 0,$$

while H_M inherits from F the uniform monotonicity and the Lipschitz continuity on any ball C_r of M. Hence, we obtain $\kappa(F_M, C_r) = \text{const} \cdot h^{-2}$; that is, the conditioning of the approximating systems may deteriorate with increasing dimension.

3. Some Variations of the Basic Results. Unlike in the matrix case, the quantities β , ν , κ in the estimate (2.7) depend also on the chosen set C. Evidently, we have

(3.1)
$$\mu(F, C_1) \ge \mu(F, C_2), \quad \nu(F, C_1) \le \nu(F, C_2) \text{ if } C_1 \subseteq C_2,$$

and, hence,

(3.2)
$$\kappa(F, C_1) \leq \kappa(F, C_2) \quad \text{if } C_1 \subset C_2$$

This suggests consideration of the asymptotic behavior of the condition number when C shrinks to a point.

More generally, for any $F: D \subset X \longrightarrow Y$, closed set $C \subset D$, and point $z \in C$, we introduce the localized bounds

(3.3)
$$\mu^{0}(F, C, z) = \sup\{t \in [0, \infty); \|Fx - Fz\| \ge t \|x - z\| \quad \forall x \in C\}, \\ \nu^{0}(F, C, z) = \inf\{t \in [0, \infty]; \|Fx - Fz\| \le t \|x - z\| \quad \forall x \in C\}$$

and the corresponding localized condition number

(3.4)
$$\kappa^{0}(F, C, z) = \begin{cases} \nu^{0}(F, C, z)/\mu^{0}(F, C, z) & \text{if } 0 < \mu^{0}(F, C, z), \ \nu^{0}(F, C, z) < \infty, \\ \infty, & \text{otherwise.} \end{cases}$$

Then the following localized version of Theorem 2.1 is valid.

THEOREM 3.1. Suppose that, for given $F: D_F \subset X \to Y$, Eq. (2.3) has a solution x^* in the closed set $C \subset D_F$ and that $\kappa^0(F, C, x^*) < \infty$. Let $G: D_G \subset X \to Y$, with $C \subset D_G$, be any mapping for which the difference (2.5) satisfies

(3.5)
$$\nu^0(E, C, x^*) < \mu^0(F, C, x^*).$$

Then, for any solution $y^* \in C$ of (2.4) and any $x^0 \in C$, $x^0 \neq x^*$, the estimate

$$(3.6) \frac{\|x^* - y^*\|}{\|x^* - x^0\|} \leq \frac{\kappa^0(F, C, x^*)}{1 - \kappa^0(F, C, x^*) \frac{\nu^0(E, C, x^*)}{\nu^0(F, C, x^*)}} \left[\frac{\|b - c\|}{\|b - Fx^0\|} + \frac{\nu^0(E, C, x^*)}{\nu^0(F, C, x^*)} + \frac{\|Ex^0\|}{\|b - Fx^0\|} \right]$$

holds.

The proof is analogous to that of Theorem 2.1 and is omitted.

The localized quantities (3.3) are closely related to the derivative of F at z, if it exists.

THEOREM 3.2. Suppose that the continuous mapping $F: D \subset X \to Y$ has a (Fréchet) derivative $F'(z) \in L(X, Y)$ at $z \in int(D)$ for which the inverse $F'(z)^{-1} \in L(Y, X)$ exists. Then for any sufficiently small $\epsilon > 0$ there is a $\delta > 0$ such that $C = \{x \in X; ||x - z|| \le \delta\} \subset D$ and

(3.7)
$$|\nu^{0}(F, C, z) - ||F'(z)|| | \le \epsilon, \quad |\mu^{0}(F, C, z) - ||F'(z)^{-1}||^{-1}| \le \epsilon.$$

Proof. Let $\epsilon > 0$, $\epsilon < \|F'(z)^{-1}\|^{-1}$, be given and $\delta > 0$ such that $C \subset D$ and

$$||Fx - Fz - F'(z)(x - z)|| \le \epsilon ||x - z|| \quad \forall x \in C$$

Then, for any $x \in C$ it follows that

$$||Fx - Fz|| = ||F'(z)(x - z) - (Fx - Fz - F'(z)(x - z))|| \begin{cases} \leq (||F'(z)|| + \epsilon)||x - z||, \\ \geq (||F'(z)^{-1}||^{-1} - \epsilon)||x - z||, \end{cases}$$

whence

(3.8)
$$\nu^{0}(F, C, z) \leq ||F'(z)|| + \epsilon, \quad \mu^{0}(F, C, z) \geq ||F'(z)^{-1}|| - \epsilon > 0.$$

Conversely, for any $h \in X$, we have $x = z + th \in C$ for some t > 0, and thus,

$$t||F'(z)h|| = ||(Fx - Fz) - (Fx - Fz - F'(z)(x - z))|| \begin{cases} \leq t(\nu^0(F, C, z) + \epsilon)||h||, \\ \geq t(\mu^0(F, C, z) - \epsilon)||h||. \end{cases}$$

This shows that

 $||F'(z)|| \leq \nu^0(F, C, z) + \epsilon, \quad ||F'(z)^{-1}||^{-1} \geq \mu^0(F, C, z) - \epsilon,$

which, together with (3.8), gives (3.7).

For the basic quantities (2.1) the same result holds under the stronger assumption that F is continuously F-differentiable on D. The details of the result should be self-evident.

From (3.7) we obtain after some computation that

$$-\frac{\epsilon}{\|F'(z)^{-1}\|^{-1}+\epsilon} \leq \frac{\kappa^0(F, C, z) - \kappa(F'(z))}{1 + \kappa(F'(z))} \leq \frac{\epsilon}{\|F'(z)^{-1}\|^{-1}-\epsilon},$$

where $\kappa(F'(z)) = ||F'(z)|| ||F'(z)^{-1}||$ is the condition number of the derivative on X.

These results show that asymptotically near z the conditioning of the nonlinear mapping F and its derivative F'(z) are the same. This has various consequences. For example, if Newton's method is used for the solution of Eq. (2.3) then at the kth step the linear equation

(3.9)
$$F'(x^k)y = Fx^k, \quad k = 0, 1, ...,$$

has to be solved. If F is badly conditioned near the solution x^* of (2.3) then the same will be true for the derivatives in (3.9) once x^k comes closer to x^* . In other words, we may expect the solution of (3.9) to be susceptible to errors, and, as a result, the iteration to slow down or even fail. This is a frequently observed phenomenon.

We end the discussion with a comment about the conditions (2.6) or (3.5) which

are not always easily verifiable. These conditions are only used to guarantee that the lower bounds $\nu(G, C)$ or $\nu^0(G, C, x^*)$ are greater than zero. Actually, we may derive an estimate for $||x^* - y^*||$ without these assumptions.

THEOREM 3.3. Let $F: D_F \subset X \to Y$ be such that $\kappa(F, C) < \infty$ on some closed set $C \subset D_F$ and that (2.3) has a solution $x^* \in C$. Then, for any $G: D_G \subset X \to Y$, $C \subset D_G$, for which (2.4) has a solution $y^* \in C$, we have

(3.10)
$$\frac{\|x^* - y^*\|}{\|x^* - x^0\|} \le \kappa(F, C) \left[\frac{\|b - c\|}{\|b - Fx^0\|} + \frac{\|G - F\|_C}{\|b - Fx^0\|} \right]$$

Here $x^0 \in C$, $x^0 \neq x^*$, is any point, and $||G - F||_C = \sup\{||Fx - Gx|| \ \forall x \in C\}$. The proof follows directly from

 $\mu(F, C) \|x^* - y^*\| \le \|Fx^* - Fy^*\| \le \|b - c\| + \|Gy^* - Fy^*\| \le \|b - c\| + \|G - F\|_C$ and (2.9).

Clearly, the result also holds with $\kappa^0(F, C, x^*)$ in place of $\kappa(F, C)$. The estimate (3.10) is certainly weaker than (2.7) but reduces to the latter one for G = F.

In the setting of Theorem 3.3, degree theory may be used to guarantee the existence of the solution of the perturbed equation. We show this for the case of mappings on \mathbb{R}^n ; under suitable compactness assumptions on the operators, the result may also be extended to Banach spaces.

THEOREM 3.4. Suppose that $F: D \subset \mathbb{R}^n \to \mathbb{R}^n$ satisfies $\kappa(F, C) < \infty$ on some closed, bounded set $C \subset D$ with nonempty interior and that $\deg(F, C, b) \neq 0$ for some $b \notin F(\partial C)$. Then, (2.3) has a solution $x^* \in int(C)$ which is unique in C. Moreover, there is a $\gamma > 0$, depending only on the norm, such that for any continuous $G: D_G \subset \mathbb{R}^n \to \mathbb{R}^n$, $C \subset D_G$, with

$$(3.11) ||G - F||_C < \delta = \gamma r \mu(F, C), r = \inf\{||x - x^*|| \quad \forall x \in \partial C\}$$

Eq. (2.4) has a solution $y^* \in C$. For these solutions, (3.10) holds.

Proof. By Kronecker's theorem (see, e.g., [2, p. 161]), (2.3) has a solution $x^* \in int(C)$ which, because of $\mu(F, C) > 0$, is unique in C. Then

$$||Fx - b|| \ge \mu(F, C)||x - x^*|| \ge \alpha = r\mu(F, C) > 0 \quad \forall x \in \partial C;$$

and, hence, there is a $\gamma > 0$ which depends only on the norm (e.g., $\gamma = 1/7$ for the Euclidean norm [2]) such that (3.11) implies deg(G, C, b) = deg(F, C, b). This ensures the existence of a (not necessarily unique) solution $y^* \in C$ of (2.4); and hence, by Theorem 3.3, the validity of (3.10).

Once again the result also holds with the localized quantities κ^0 and μ^0 in place of κ and μ . As a simple example involving the latter form of the theorem, consider the polynomials

$$f(t) = \prod_{k=1}^{20} (t-k), \quad g(t) = f(t) - \epsilon t^{19} \quad \forall t \in \mathbb{R}^1,$$

on $C = [19 - \eta, 19 + \eta]$ with sufficiently small $\eta > 0$. Then, deg(f, C, 0) = -1 and

 $||f - g||_C \le 3 \cdot 10^{24} \epsilon$, if at least $\eta < 0.1$. Moreover, we have approximately $\mu^0(f, C, x^*) \doteq |f'(19)| = 18! \doteq 6.4 \times 10^{15}$. Hence, for $\epsilon = 2^{-23}$, the localized form of (3.11) is violated, and indeed the computational results of Wilkinson [4] show that the roots of f at t = 18 and 19 turn into a pair of conjugate complex roots of g. On the other hand, for $\epsilon = 2^{-55}$, we see that g has a solution $y^* \in C$; and the localized form of (3.10) with $x^0 = 18.9$, and $f(x^0) \doteq 5 \times 10^{14}$ gives approximately $|y^* - 19| \le 2 \times 10^{-8}$, while the computational results in [4] show that $|y^* - 19| \doteq .9 \times 10^{-8}$.

Computer Science Center University of Maryland College Park, Maryland 20742

1. G. J. FIX & K. LARSEN, "On the convergence of SOR iterations for finite element approximations to elliptic boundary value problems," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 536-547. MR 45 #2935.

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

3. R. S. VARGA, Functional Analysis and Approximation Theory in Numerical Analysis, Conference Board of the Mathematical Sciences Regional Conference Series in Appl. Math., no. 3, SIAM, Philadelphia, Pa., 1971. MR 46 #9602.

4. J. WILKINSON, Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N. J., 1963. MR 28 #4661.