Inaccuracy in Quasi-Newton Methods:
Local Improvement Theorems

by

J.E. Dennis, Jr.¹

and

Homer F. Walker²

Technical Report 83-11, March 1983.

¹Mathematical Sciences Department, Rice University, Houston, Texas 77251. Research sponsored by DOE DE-AS05-82ER13016, ARO DAAG-79-C-0124, NSF MCS81-16779. This work was supported in part by the International Business Machine Corporation, Palo Alto Scientific Center, Palo Alto, CA.

²Department of Mathematics, University of Houston, Houston, Texas 77004. Research sponsored by DOE DE-AS05-82ER13016.

Abstract

In this paper, we consider the use of bounded-deterioration quasi-Newton methods implemented in floating-point arithmetic to find solutions to $F(x) = 0$ where only inaccurate F -values are available. Our analysis is for the case where the relative error in F is less than one. We obtain theorems specifying local rates of improvement and limiting accuracies depending on the nearness to Newton's method of the basic algorithm, the accuracy of its implementation, the relative errors in the function values, the accuracy of the solutions of the linear systems for the Newton steps, and the unit-rounding errors in the addition of the Newton steps.

Key words

quasi-Newton methods, local convergence, floating-point errors, limiting accuracy.

1. Introduction

This paper is the first of several anticipated reports on the effects of inaccuracy, computational or otherwise, on quasi-Newton methods for solving nonlinear algebraic equations. In this report, we consider bounded-deterioration quasi-Newton methods and discuss the effects on their performance of inaccuracy from all sources, including not only computer arithmetic, but also the differences between the ideal problem, its mathematical model, and the computer implementation of the model. In later work, we will explore in greater depth the effects of specific sources of inaccuracy on the performance of particular least-change secant methods; in fact, we will suggest modifications of the standard versions of these methods which enable them to deal more effectively with certain types of inaccuracy. Here, however, our concern is with the general case, and our objective is to investigate rates of improvement and limiting accuracies that can be obtained near solutions. We would like eventually to provide heuristics for deciding how accurately the nonlinear residuals must be evaluated at the current point in order to improve the solution estimate. This would be very useful for certain parameter identification problems in which a system of differential equations must be integrated numerically to compute the residuals at the current parameter estimates, and the cost of the integration escalates with the required accuracy.

Urabe (1956), Collatz (1966), Lancaster (1966), and Ypma (1983) also have considered the same kinds of effects on general iterative methods. Glad and Goldstein (1977) have considered the effects of objective function inaccuracy on the unconstrained minimization problem.

They suggested an algorithm to achieve a limiting accuracy consistent with ours and with the limiting accuracy expression obtained by Ypma. Although the work was done independently, our results are especially close to Ypma's, but he follows the tradition of the earlier papers on general methods in relating each term in the iteration sequence being generated to the iterate that would be generated were there no errors. This approach may be the reason why our results, based directly on the computed sequence, are more easily applicable to methods with memory such as Broyden's method. In fact, applicability to the sparse Broyden method was our guiding light, and it will be a particular focus of subsequent work.

The problem of interest here is the following:

Problem. Given $F: \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$, find $x_* \in \Omega$ satisfying

$$F(x_*) = 0. \quad (1.1)$$

The function F can be thought of as the idealized problem being modeled on the computer, or as the problem actually on the computer, but whose values are being approximated to be used by implementations of methods which are modeled after Newton's method. Accordingly, the following is assumed for the idealized problem throughout the sequel:

The standard hypothesis. Let F be differentiable in an open convex neighborhood Ω of a point $x_* \in \mathbb{R}^n$ for which $F(x_*) = 0$ and $F'(x_*)$ is non-singular. Let $\gamma \geq 0$ and $p \in (0, 1]$ be such that for $x \in \Omega$,

$$|F'(x) - F'(x_*)| \leq \gamma |x - x_*|^p,$$

where $|\cdot|$ denotes a norm on $\mathbb{R}^n$ and its subordinate operator norm on $\mathbb{R}^{n \times n}$.

This standard hypothesis is sufficient to insure that sequences of iterates produced by Newton's method converge locally to x_* with q-order $(1 + p)$. (See, for example, Dennis and Schnabel (1983) or Ortega and Rheinboldt (1970).)

The methods of interest for solving (1.1) numerically are iterative methods of the general quasi-Newton form

$$x_{k+1} = x_k - B_k^{-1}F(x_k), \quad (1.2)$$

in which B_k is regarded as an approximation of the Jacobian matrix $F'(x_k)$. (Since our concern here is with local behavior, we consider only situations in which the full step $-B_k^{-1}F(x_k)$ would be taken for each k , if it could be computed). For generality, it is desirable to specify as little as possible about the manner in which the approximate Jacobians B_k in (1.2) are chosen; but for applicability of the results, it is useful to give some carefully chosen structure to the approximation rule. Otherwise, each application ends up requiring a contextual reproof of the general result. Therefore, we assume that there is some non-singular matrix B_* and an update function U , defined in a neighborhood $N \subseteq \mathbb{R}^n \times \mathbb{R}^{n \times n}$ of (x_*, B_*) and having values which are subsets of $\mathbb{R}^{n \times n}$, such that $B_{k+1} \in U(x_k, B_k)$ for each k provided $(x_k, B_k) \in N$.

The notion of an update function has proved to be very useful in the analysis of quasi-Newton methods (see Broyden-Dennis-More (1973) or Dennis and Walker (1981)). It provides an essentially completely general framework within which all of the popular and successful quasi-Newton methods can be described. To understand the role of B_* , one should think of it as being a distinguished approximation of $F'(x_*)$ among some set of allowable approximants. For example, in the case of

the nonlinear Jacobi iteration (see Ortega and Rheinboldt (1970)), in which each B_k is a diagonal matrix having the same diagonal as $F'(x_k)$, one might take B_* to be the diagonal matrix having the same diagonal as $F'(x_*)$. In many applications, one naturally takes $B_* = F'(x_*)$.

Our concern here is with the effects of the inaccuracies which inevitably appear in any practical implementation of an iteration (1.2) to solve (1.1). These inaccuracies can arise from the following sources: the evaluation of an approximation to $F(x_k)$, the solution of a linear system to obtain the step $-B_k^{-1}F(x_k)$, the addition of the step to x_k to obtain x_{k+1} , and the determination of B_{k+1} from $U(x_k, B_k)$. While inaccuracy can potentially cause problems at any stage of an iteration (1.2), it will almost certainly degrade the performance of the iteration near x_* . In particular, if the iterates approach, but never equal x_* , then we will see below that inaccuracy in (1.2) causes the speed of their approach to decrease until, ultimately, further iterations can no longer be guaranteed to produce additional accuracy in the iterates. Thus, it is more appropriate to call our results local improvement theorems rather than local convergence theorems.

Our specific purpose in this paper is to investigate the local improvement properties of an iteration (1.2) in the presence of inaccuracy. In the following, we first offer a general model of a practical implementation of (1.2) which takes into account inaccuracies arising from the sources listed above. Then we analyze the performance of this model near x_* to determine rates of improvement, or convergence, and limiting accuracies that can ultimately be obtained. As they should, the results of our analysis reduce to the standard results (see Theorem

A2.1 of Dennis and Walker (1981)) for an iteration (1.2) when no inaccuracy is present. Finally, to illustrate the range of our results and to place them in some perspective, we consider the special case of stationary iteration for linear systems; for example, the analysis given here is essentially a generalization of that of Moler (1967) for LU iterative improvement.

2. Local Improvement Analysis

Our model of a practical implementation of (1.2) which takes inaccuracy into account is the following:

Iteration 2.1. Given x_0 and B_0 , do the following until satisfied or no longer feasible:

- (2.1.i) Call for $F(x_k)$ and receive $F_k = F(x_k) + N(x_k)$.
- (2.1.ii) If $k > 0$, then compute B_k which differs from an element of $U(x_{k-1}, B_{k-1})$ by M_k .
- (2.1.iii) Ask for the solution of $B_k s_k = -F_k$ and get s_k , the solution of $(B_k + \Delta_k)s_k = -F_k$.
- (2.1.iv) Try to add s_k to x_k and get $x_{k+1} = (I + \Gamma_k)(x_k + s_k)$ for a diagonal matrix Γ_k .

In (2.1.i), $N(x_k)$ is to be thought of as "noise" or inaccuracy associated with the computation and approximation of $F(x_k)$. Our assumptions on $N(x_k)$ are flexible enough for it to be regarded as the combination of noise from all the relevant sources that account for the difference between F_k and $F(x_k)$. For example, if (1.1) is of the form $G(x) - b = 0$, where G models some phenomenon and b is data, then $N(x)$ can be

regarded at the choice of the user as a combination of any or all of the following: the modeling error in G , the errors in the computer evaluation of G , the data errors in b , and the cancellation error in forming $G(x) - b$.

Our results are for the setting where F can be evaluated to at least one significant figure for any iterate from which improvement is expected on performing the next iteration. There is a considerable body of literature on stochastic approximation algorithms, which are intended for use in solving (1.1) when $N(x_k)$ is random noise, i.e., an observation on a random variable (see Young (1976)). The algorithms considered here are not at all intended to be regarded as stochastic approximation algorithms, although the case in which $N(x_k)$ is random noise is certainly not meant to be excluded. In practice, stochastic approximation algorithms are perhaps of greatest interest when $N(x_k)$ is likely to be fairly large relative to $F(x_k)$, and here we wish to study the effects of relatively small amounts of noise on quasi-Newton methods, since these methods are in common use for problems in which such noise occurs.

Similarly, M_k in (2.1.ii) is to be thought of as inaccuracy in the determination of B_k . For the purposes of this paper, one can regard it as the least troubling of the inaccuracy terms in Iteration 2.1; however, we feel that it should not be omitted altogether. There are two cases in which it seems fair to take $M_k = 0$. The first is the obvious one in which $B_k = B_{k-1}$, an option allowed in many implementations of methods of the form (1.2). The second is that in which B_k is determined in a manner independent of B_{k-1} , e.g., as an analytic or finite-difference evaluation of $F'(x_k)$. In this case, one can choose to regard

B_k as being exact with any inaccuracy incurred in its evaluation absorbed in Δ_k in (2.1.iii). Also, it is sometimes possible by increasing the constants to absorb M_k into the bounded deterioration assumptions on U . In the following, we suggest thinking of Δ_k and M_k as separate.

The effect of Δ_k in (2.1.iii) on the iteration depends on the size of $B_k^{-1}\Delta_k$, which in practice depends on the stability of the linear solution algorithm, the arithmetic in which it is implemented, and the conditioning of B_k with respect to inversion. We are not concerned here with the particular solution algorithm used, so long as it is stable, and so we hypothesize a uniform bound $|B_k^{-1}\Delta_k| \leq \mu < 1$. In practice μ is usually much smaller than 1, and one can properly regard it as such in the following. The assumption that μ is independent of k is reasonable, because for many algorithms implemented in arithmetic with unit rounding error ϵ_A , one has

$$|B_k^{-1}\Delta_k| \leq \phi(n)K(B_k)\epsilon_A, \quad (2.2)$$

independent of the right-hand side of the linear equation being solved. Here, $K(B_k) = |B_k| \cdot |B_k^{-1}|$ embodies dependence on k , but we observe in Theorem 2.5 and Remark 1 below that $\{|B_k|\}$ and $\{|B_k^{-1}|\}$ stay uniformly bounded locally. Consequently, our hypothesis on $|B_k^{-1}\Delta_k|$ seems sound. In fact, it is reasonable to think of $\mu \approx K(B_*)\epsilon_A$ in the context of this paper.

The term Γ_k in (2.1.iv) is intended to account for error incurred in adding x_k to s_k in floating-point arithmetic with unit rounding error ϵ_A , and so we assume that $|\Gamma_k| \leq \epsilon_A$.

It is convenient to establish two technical lemmas before introducing our main result.

Lemma 2.2. Under the standard hypothesis, for any $x \in \Omega$,

$$|F(x) - F'(x_*)(x - x_*)| \leq \frac{\gamma |x - x_*|^{p+1}}{p+1}.$$

Proof: Let $e = x - x_*$. Then

$$\begin{aligned} |F(x) - F'(x_*)(x - x_*)| &= |F(x) - F(x_*) - F'(x_*)e| \\ &= |\{ \int_0^1 [F'(x - te) - F'(x_*)] dt \} e| \\ &\leq \gamma [\int_0^1 (1-t)^p dt] \cdot |e|^{p+1} \end{aligned}$$

and the result follows from elementary calculus.

Our second lemma is intended to shed light on the change in the error in successive approximate solutions at a single step of an iteration of the form (2.1). For convenience and clarity, we introduce a "single-step iteration" and direct the lemma toward it. Note that the constants r and r_* need not be less than 1. We will follow the lemma with a simple illustrative example.

Iteration 2.3. Suppose that x and B are given.

(2.3.i) Call for $F(x)$ and receive $F = F(x) + N(x)$.

(2.3.ii) Ask for the solution of $Bs = -F$ and get s , the solution of $(B + \Delta)s = -F$.

(2.3.iii) Try to add s to x and get $x_+ = (I + \Gamma)(x + s)$ for a diagonal matrix Γ .

Lemma 2.4. Let F satisfy the standard hypothesis and let $B_* \in R^{n \times n}$ have the property that B_*^{-1} exists and $|I - B_*^{-1}F'(x_*)| \leq r_*$. Let ϵ_A, ϵ_F , and r_A be positive constants such that $|\Gamma| \leq \epsilon_A$, $|N(x)| \leq \epsilon_F|F(x)|$, and $\epsilon_A|x_*| \leq r_A|x - x_*|$. Suppose also that B satisfies $|B - B_*| \leq \delta < 1/|B_*^{-1}|$, which implies that B is invertible, and that $|B^{-1}\Delta| \leq p < 1$. Then Iteration 2.3 is well defined and $|x_+ - x_*| \leq r|x - x_*|$, where

$$r = \frac{1 + \epsilon_A}{(1-p)(1 - |B_*^{-1}|\delta)} \{ [(1-p)|B_*^{-1}|\delta + p]|B_*^{-1}F'(x_*)| + |B_*^{-1}| \left[\frac{\gamma(1+\epsilon_F)|x-x_*|^p}{p+1} + \epsilon_F|F'(x_*)| \right] \} + (1+\epsilon_A)r_* + r_A.$$

Proof. For convenience, denote $e = x - x_*$, $e_+ = x_+ - x_*$, $\beta = |B_*^{-1}|$, and $F'_* = F'(x_*)$. Note that since

$$|B_*^{-1}(B - B_*)| \leq \beta|B - B_*| \leq \beta\delta < 1, \quad B^{-1} \text{ exists and is given by}$$

$$B^{-1} = [I + B_*^{-1}(B - B_*)]^{-1} B_*^{-1}. \quad (2.4)$$

From the Neumann series,

$$[I + B_*^{-1}(B - B_*)]^{-1} = \sum_{j=0}^{\infty} [-B_*^{-1}(B - B_*)]^j,$$

one obtains

$$B^{-1} = B_*^{-1} - [I + B_*^{-1}(B - B_*)]^{-1} B_*^{-1}(B - B_*) B_*^{-1}. \quad (2.5)$$

Similarly, since $|B^{-1}\Delta| \leq p < 1$, $(B + \Delta)^{-1}$ exists and is given by

$$(B + \Delta)^{-1} = (I + B^{-1}\Delta)^{-1} B^{-1}; \quad (2.6)$$

and one can write

$$(B + \Delta)^{-1} = B^{-1} - (I + B^{-1}\Delta)^{-1} B^{-1}\Delta B^{-1}. \quad (2.7)$$

With (2.4) and (2.5), one obtains from (2.7) that

$$\begin{aligned} (B + \Delta)^{-1} &= B_*^{-1} - [I + B_*^{-1}(B - B_*)]^{-1} B_*^{-1}(B - B_*)B_*^{-1} \\ &\quad - (I + B^{-1}\Delta)^{-1} B^{-1}\Delta[I + B_*^{-1}(B - B_*)]^{-1} B_*^{-1}. \end{aligned} \quad (2.8)$$

Also, it follows from (2.4) that

$$|B^{-1}| \leq \frac{\beta}{1 - \beta\delta}$$

and from this inequality and (2.6) that

$$|(B + \Delta)^{-1}| \leq \frac{\beta}{(1 - \mu)(1 - \beta\delta)}. \quad (2.9)$$

Since $(B + \Delta)^{-1}$ exists, Iteration 2.3 is well-defined and yields

$$x_+ = (I + \Gamma)\{x - (B + \Delta)^{-1}[F(x) + N(x)]\},$$

from which it follows that

$$\begin{aligned} e_+ &= (I + \Gamma)\{e - (B + \Delta)^{-1}[F(x) + N(x)]\} + \Gamma x_* \\ &= (I + \Gamma)\{[I - (B + \Delta)^{-1}F'_*]e - (B + \Delta)^{-1}[F(x) - F'_*e] - (B + \Delta)^{-1}N(x)\} + \Gamma x_*. \end{aligned}$$

Substituting (2.8) in this expression gives

$$\begin{aligned} e_+ &= (I + \Gamma)\{(I - B_*^{-1}F'_*) + \{[I + B_*^{-1}(B - B_*)]^{-1} B_*^{-1}(B - B_*) + \\ &\quad (I + B^{-1}\Delta)^{-1} B^{-1}\Delta[I + B_*^{-1}(B - B_*)]^{-1}\} B_*^{-1}F'_*\} e - \\ &\quad (I + \Gamma)(B + \Delta)^{-1}\{[F(x) - F'_*e] + N(x)\} + \Gamma x_*; \end{aligned}$$

and using (2.9) and Lemma 2.2, one obtains

$$\begin{aligned}
|e_+| &\leq (1 + \epsilon_A) \{ r_* + [\frac{\beta\delta}{1 - \beta\delta} + \frac{\mu}{(1-\mu)(1 - \beta\delta)}] |B_*^{-1} F'_*| \} |e| \\
&\quad + (1 + \epsilon_A) \frac{\beta}{(1 - \mu)(1 - \beta\delta)} \{ \frac{\gamma|e|^P}{p+1} + \epsilon_F (\frac{\gamma|e|^P}{p+1} + |F'_*|) \} |e| + r_A |e| \\
&= \left\{ \frac{(1 + \epsilon_A)}{(1 - \mu)(1 - \beta\delta)} \left[\{ (1-\mu)\beta\delta + \mu \} |B_*^{-1} F'_*| + \beta \left\{ \frac{\gamma(1 + \epsilon_F)|e|^P}{p+1} + \epsilon_F |F'_*| \right\} \right] \right. \\
&\quad \left. + (1 + \epsilon_A) r_* + r_A \right\} |e|.
\end{aligned}$$

We are used to thinking of relative errors, so ϵ_F is easily interpreted, and r_A is just the ratio of the best possible relative error ϵ_A in representing x_* to the relative error in the current approximator x .

When (2.2) is applied to a function whose values are more inaccurate than one supposes, or when the naive user asks for excessive accuracy in x , typical behavior is for the iteration sequence to thrash around and neither converge nor diverge. Often, this behavior is misdiagnosed as failure to provide a sufficiently good initial guess. The preceding lemma suggests an alternate explanation which we will illustrate with a simple example.

Consider $F(x) = x^2 - 2$, and suppose that the subroutine to evaluate F is implemented in 3-digit arithmetic, by which we mean that x is rounded to 3 decimal digits, x^2 is likewise rounded, and 2 is subtracted. In order to isolate the influence of function inaccuracy on local behavior, we will use Newton's method and arithmetic of sufficient accuracy so that x_+ , defined with the given inaccurate F , is correctly computed to 6 decimal digits. This causes r , the rate of improvement in the above lemma, to be simplified in each step of the example.

At $x_0 = 1.$, we computed $F_0 = -1.0$, $\epsilon_F = 0$, $r_A \approx 2 \times 10^{-7}$, which gives no indication of problems. Indeed, there were none since $x_1 = 1.5$, $|e_1| \approx .061$, was a good step. We found $r_A \approx 8.2 \times 10^{-5}$, $F_1 = .25$, $\epsilon_F = 0$, again no indication of problems. And we made a good step, to $x_2 = 1.416667$, so $|e_2| \approx .0017$. This time $r_A \approx 2.9 \times 10^{-3}$, $F_2 = .02$, $\epsilon_F \approx 1.9 \approx r$, which predicts $|e_3| \approx .0032$. If we only knew it, we should stop here since $r > 1$. As predicted, the error did increase with $x_3 = 1.409600$, $|e_3| \approx .0033$. But this time, $r_A \approx 1.5 \times 10^{-3}$, $F_3 = -.01$, $\epsilon_F \approx .23$. Notice that we have now been thrown out far enough from x_* so that the cancellation error in $F(x)$ is not as bad, and $r \approx \epsilon_F \approx .23$. This predicts $|e_4| \approx .23 |e_3| \approx .00076$, and in fact, $|e_4| \approx .00075$. Now we are too close again, and $\epsilon_F \approx 2.3 \approx r$. This behavior continued through the eight iterations we made with x_5 , x_6 moving away until F_6 had a relative error $\epsilon_F \approx .23$, which was sufficient to make x_7 move back toward x_* , but then F_7 had a relative error $\epsilon_F \approx 2.4$ and x_8 moved off again.

Theorem 2.5 below is our main result. In it, we assume that there is a norm $\|\cdot\|$ on $\mathbb{R}^{n \times n}$ in which a bounded deterioration inequality (2.10) holds for the update function U . This assumption is not a serious restriction, because the update functions appropriate for essentially all of the popular and successful quasi-Newton methods exhibit such bounded deterioration (see Broyden-Dennis-More (1973)). The theorem asserts that the iterates x_k will improve as approximations to x_* at a q -linear rate until one of three conditions is violated. These conditions can be categorized as representing respectively: the inaccuracies of the problem (condition (ii)), the inaccuracies in the basic algorithm

(condition (iii)), and the ratio of the accuracy of the current iterate to the unit rounding error of the arithmetic of the implementation (condition (i)). One interpretation of the theorem is that suggested by Lemma 2.4 and the example above: If the relative errors in x_k and F_k are not too great, and if B_k is a sufficiently good approximate Jacobian, then one can expect the step to x_{k+1} to be an improvement. An alternate to this one-step interpretation is to think of choosing a good enough basic algorithm, implementing it with sufficient accuracy, and supplying it with accurate enough function values, so that improvement is obtained at a fast enough rate r until condition (i) with $r_A = \frac{\epsilon_A}{\text{rel}x}$ is the first to fail, where $\text{rel}x$ is the user supplied convergence condition on $\frac{|x_k - x_*|}{|x_*|}$.

In Theorem 2.5, we let η denote a constant for which $|A| \leq \eta \|A\|$ for every $A \in \mathbb{R}^{n \times n}$.

Theorem 2.5. Let F satisfy the standard hypothesis and let $B_* \in \mathbb{R}^{n \times n}$ have the property that B_*^{-1} exists and $\|I - B_*^{-1}F'(x_*)\| \leq r_* < 1$. Let U be defined in a neighborhood $N = N_1 \times N_2$ of (x_*, B_*) where $N_1 \subset \Omega$ and N_2 contains only nonsingular matrices. Assume that there are nonnegative constants α_1, α_2 such that for each $(x, B) \in N$ and $x_+ = x - B^{-1}F(x)$, every $B_+ \in U(x, B)$ satisfies

$$\|B_+ - B_*\| \leq [1 + \alpha_1 \sigma(x, x_+)^p] \|B - B_*\| + \alpha_2 \sigma(x, x_+)^p \quad (2.10)$$

for $\sigma(x, x_+) = \max\{|x - x_*|, |x_+ - x_*|\}$. Let $\epsilon_A, \mu, \epsilon_F$, and r_A be non-negative constants such that $\mu < 1$ and

$$1 > r_0 = \frac{(1 + \epsilon_A)}{(1 - \mu)} [\epsilon_F |B_*^{-1}| \cdot |F'(x_*)| + \mu |B_*^{-1} F'(x_*)|] + (1 + \epsilon_A) r_* + r_A.$$

Then for any r such that $r_0 < r < 1$ and any ϵ_r, δ_r such that

(a) $(x, B) \in N$ whenever $|x - x_*| \leq \epsilon_r$ and $|B - B_*| \leq \delta_r$,

(b) $\delta_r < 1/|B_*^{-1}|$,

(c)

$$r \geq r_0 + \frac{(1 + \epsilon_A) |B_*^{-1}|}{(1 - \mu)(1 - |B_*^{-1}| \delta_r)} \left[\frac{(1 + \epsilon_F)}{p + 1} \gamma \epsilon_r^p + \right.$$

$$\left. + (|B_*^{-1} F'(x_*)| + \epsilon_F |B_*^{-1}| \cdot |F'(x_*)|) \delta_r \right],$$

the following is true:

If $|x_0 - x_*| \leq \epsilon_r$ and $|B_0 - B_*| \leq \delta_r$, then any iteration of the form (2.1) with $|B_k^{-1} \Delta_k| \leq \mu$ and $|\Gamma_k| \leq \epsilon_A$ is well defined and satisfies

$$|x_{k+1} - x_*| \leq r |x_k - x_*| \text{ and } |B_k - B_*| \leq \delta_r \quad (2.11)$$

until the first value of $k \geq 0$ for which one of the following fails to hold:

(i) $\epsilon_A |x_*| \leq r_A |x_k - x_*|$;

(ii) $|N(x_k)| \leq \epsilon_F |F(x_k)|$;

(iii) if $k > 0$, then

$$||B_0 - B_*|| + \sum_{i=1}^k ||M_i|| + \frac{\alpha_2}{1-r^p} |x_0 - x_*|^p \leq \frac{\delta_r}{r\eta},$$

where

$$\tau = \prod_{i=0}^{\infty} [1 + \alpha_1 |x_0 - x_*|^{p_i}].$$

Remark 1: In light of (2.11) and the hypotheses of the theorem, one sees that not only $|B_k|$ but also $|B_k^{-1}|$, $|B_k + \Delta_k|$, and $|(B_k + \Delta_k)^{-1}|$ are bounded independent of k as long as (i), (ii), and (iii) hold.

Remark 2: Suppose that for each k , $F(x_k)$ and B_k are determined exactly (so that $N(x_k) = 0$ and $M_k = 0$) and the step s_k , which may be determined inexactly, is added exactly to x_k (so that $\Gamma_k = 0$). Then one can take $\epsilon_F = \epsilon_A = 0$ and obtain from the conclusion of the theorem that $\{x_k\}$ converges to x_* q -linearly with rate constant at most r , provided

$$||B_0 - B_*|| + \frac{\alpha_2}{1-r^p} |x_0 - x_*|^p \leq \frac{\sigma_r}{\tau n}.$$

Proof. Suppose that the hypotheses of the theorem hold and that $|x_0 - x_*| \leq \epsilon_r$ and $|B_0 - B_*| \leq \sigma_r$ for some r such that $r_0 < r < 1$ and some positive ϵ_r , σ_r satisfying (a), (b), and (c). We prove the theorem by induction on k .

If (i) and (ii) hold for $k = 0$ (and if $|B_0^{-1}\Delta_0| \leq p$ and $|\Gamma_0| \leq \epsilon_A$), then an iteration of the form (2.1) is well-defined for $k = 0$ and $|x_1 - x_*| \leq r|x_0 - x_*|$ by Lemma 2.4.

As an inductive hypothesis, suppose that for some $j > 0$, an iteration of the form (2.1) is well-defined and satisfies (2.11) for $k = 0, \dots, j-1$. One has that

$$\sigma(x_{k+1}, x_k) = |x_k - x_*| \leq r^k |x_0 - x_*| \quad (2.12)$$

for $k = 0, \dots, j-1$. If B_j is determined as in (2.1.ii), then (2.12)

and repeated applications of (2.10) yield

$$\begin{aligned} \|B_j - B_*\| &\leq \tau_{jj} \|B_0 - B_*\| \\ &+ \sum_{k=0}^{j-1} \tau_{jk} [\|M_{j-k}\| + \alpha_2 r^{p(j-k-1)} |x_0 - x_*|^p], \end{aligned} \quad (2.13)$$

where

$$\tau_{jk} = \begin{cases} 1, & \text{if } k = 0 \\ \prod_{i=1}^k [1 + \alpha_1 r^{p(j-i)} |x_0 - x_*|^p], & \text{if } 1 \leq k \leq j. \end{cases}$$

It follows from (2.13) that

$$\|B_j - B_*\| \leq \tau_j \{ \|B_0 - B_*\| + \sum_{k=1}^j \|M_k\| + \frac{\alpha_2}{1-r^p} |x_0 - x_*|^p \}.$$

Consequently, if (iii) is satisfied for $k=j$, then $\|B_j - B_*\| \leq \epsilon_r$ and B_j^{-1} exists. If (i) and (ii) are also satisfied for $k = j$ (and if $\|B_j^{-1} \Delta_j\| \leq \mu$ and $\|\Gamma_j\| \leq \epsilon_A$), then Lemma 2.4 implies that x_{j+1} is well-defined by (2.1.iii) and (2.1.iv) and satisfies $|x_{j+1} - x_*| \leq r |x_j - x_*|$. This completes the induction, and the theorem is proved.

3. Stationary iteration for linear systems

We now consider the special case of (1.1) in which the system to be solved is linear, i.e., $F(x) = Ax - b$ for $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The iteration of interest is a stationary iteration

$$x_{k+1} = x_k - B^{-1} F(x_k),$$

in which $B \in \mathbb{R}^{n \times n}$ is a convenient invertible approximation of $A = F'(x_k)$. Of course, this iteration is just (1.2) with $B_k = B$ for

each k . Classical examples of such an iteration are Jacobi iteration, Gauss-Seidel iteration, and iterative refinement. In theory, the convergence properties of this iteration are determined by how well B approximates A . For example, iterative refinement in its ideal form uses $B = A$ and therefore produces the exact solution in one step if no inaccuracy enters the iteration. In practice, of course, inaccuracy not only affects the rate of improvement of the iterates but also the limiting accuracy that can be obtained.

We discuss below the implications of the results of the preceding section for an iteration of this type. Our practical model is that obtained from Iteration 2.1 in the case of interest, i.e., the following:

Iteration 3.1. Given x_0 , do the following until satisfied:

- (3.1.i) Call for $F(x_k)$ and receive $F_k = F(x_k) + N(x_k)$.
- (3.1.ii) Ask for the solution of $Bs_k = -F_k$ and get s_k , the solution of $(B + \Delta_k)s_k = -F_k$.
- (3.1.iii) Try to add s_k to x_k and get $x_{k+1} = (I + \Gamma_k)(x_k + s_k)$ for a diagonal matrix Γ_k .

The development here parallels and extends to a large extent the treatment of iterative refinement by Moler (1967). In particular, Iteration 3.1 reduces to his model when $B = A$; and our results share many of the same essential features in this case.

Our objective is to determine rates of improvement and limits of accuracy that can be obtained by Iteration 3.1 near $x_* = A^{-1}b$. Our plan is first to apply Theorem 2.5 in a straightforward way and then to

obtain more refined conclusions by making more specific assumptions about the context in which the iteration is carried out and looking more closely at the bounds μ and ϵ_F . One has quite a bit of flexibility in applying Theorem 2.5, most of it centering around the choice of B_* . For example, B_* can be taken to be $F'(x_*) = A$ (even if this matrix can not be observed), or the representation of A in the computer, if this differs from A , or the matrix B used to define the iteration matrix. In fact, it might be of interest to take B_* to be the matrix that would define an ideal iteration matrix and let B be the product of its computed factors. We leave the full exploration of these possibilities to the reader and take $B_* = B$ below, a choice which gives the sharpest results in the cases which we believe will most often be of interest.

We apply Theorem 2.5 with $F'(x_*) = A$, $B_* = B$, and $\gamma = \sigma_r = 0$. If $|I - B^{-1}A| = r_* < 1$ and if ϵ_A , μ , ϵ_F , and r_A are non-negative constants such that $\mu < 1$ and

$$1 > r_0 = \frac{(1 + \epsilon_A)}{(1 - \mu)} [\epsilon_F |B^{-1}| |A| + \mu |B^{-1}A|] + (1 + \epsilon_A)r_* + r_A, \quad (3.2)$$

then Theorem 2.5 immediately implies that for any x_0 , Iteration 3.1, with $|B^{-1}\Delta_k| \leq \mu$ and $|\Gamma_k| \leq \epsilon_A$, is well-defined and the iterates $\{x_k\}$ satisfy $|x_{k+1} - x_*| \leq r_0 |x_k - x_*|$ as long as both $|N(x_k)| \leq \epsilon_F |F(x_k)|$ and $\epsilon_A |x_*| \leq r_A |x_k - x_*|$.

We note in passing that if one makes the alternative choices $F'(x_*) = B_* = A$, $\gamma = 0 = r_*$, and $\sigma_r = |B - A|$, then for appropriate ϵ_A , μ , ϵ_F , r_A , and σ_r , Theorem 2.5 yields the linear convergence rate constant

$$\frac{(1 + \epsilon_A)}{(1 - \mu)(1 - |A^{-1}| \sigma_r)} [(1 - \mu) |A^{-1}| \sigma_r + \mu + \epsilon_F K(A)] + r_A.$$

But this rate constant is at least as great as r_0 given by (3.2), since

$$|B^{-1}A| \leq \frac{1}{1 - |A^{-1}|\epsilon_r}, \quad |B^{-1}||A| \leq \frac{K(A)}{1 - |A^{-1}|\epsilon_r}, \quad \text{and} \quad |I - B^{-1}A| \leq \frac{|A^{-1}|\epsilon_r}{1 - |A^{-1}|\epsilon_r}.$$

To draw more refined conclusions, let us consider more closely the two terms in brackets in (3.2), which are associated with the inaccuracy in evaluating F and the inaccuracy in solving linear systems with coefficient matrix A . Below, we obtain more detailed results by determining a useful expression for μ and then deriving a reasonable value of ϵ_F such that $|N(x_k)| \leq \epsilon_F |F(x_k)|$ whenever $\epsilon_A |x_*| \leq r_A |x_k - x_*|$. We will then be able to determine how much improvement, if any, can be guaranteed by an iteration of (3.1) when there is a relative error of at least $\frac{\epsilon_A}{r_A}$ in x_k . As is customary in discussions of iterative refinement, we assume that all computation except that involved in the computation of F is done in some "working precision" arithmetic, while an approximate value of F is determined in some other (perhaps higher) precision arithmetic, and then rounded to working precision. In the following, we take $\epsilon_1 = \epsilon_A$ to be the unit rounding error in the working precision arithmetic and denote by ϵ_2 the unit rounding error in the arithmetic used in the evaluation of F . For convenience, we also assume that $|\cdot|$ is the l_∞ vector norm, i.e., that $|v| = \max_i |v_i|$ for $v = (v_1, \dots, v_n)^T \in \mathbb{R}^n$, although the results below hold without substantive change for any vector norm.

To determine a useful expression for μ , we suppose either that one of the standard factorization techniques is applied to B in arithmetic with unit rounding error ϵ_1 , or that B is already given in factored

form. In the first case, (2.2) holds with $B_k = B$ and $\epsilon_A = \epsilon_1$, and so one need take μ no greater than $\phi(n)K(B)\epsilon_1$. With $\|\cdot\|$ the l_∞ norm, $\phi(n)$ rarely exceeds n^2 in practice and is often of order unity (Wilkinson (1965)). At any rate, one should be able to take μ far less than one unless A is very ill-conditioned. For convenience, we follow Moler(1967) and simply write

$$\mu = \rho K(B)\epsilon_1 \quad (3.3)$$

for an appropriate value of ρ , e.g., for the smallest ρ such that $\|B^{-1}\Delta_k\| \leq \rho K(B)\epsilon_1$ for all k . In the second case, one should have $\|B^{-1}\Delta_k\| \leq \rho K(B)\epsilon_1$ with a relatively small value of ρ , since Δ_k only accounts for the small errors in solving the factor systems. Thus we express μ by (3.3) in either case.

To determine ϵ_F , we take the novel point of view that F_k is determined in (3.1.i) by performing an inaccurate evaluation of a function $\tilde{F}(x) = \tilde{A}x - \tilde{b}$ defined by $\tilde{A} \in \mathbb{R}^{n \times n}$ and $\tilde{b} \in \mathbb{R}^n$ which are regarded as approximations of A and b , respectively. In practice, $\tilde{A}$ and $\tilde{b}$ arise as inaccurate representations within a computer of some ideal A and b associated with a problem of interest. For example, $\tilde{A}$ might be the floating-point representations of measured values of some independent variables A that produced observations whose floating-point representations are $\tilde{b}$ of some dependent variable b . This point of view is taken not only for the sake of generality but also to illustrate another point which we feel is very important: the treatment of inaccuracy given here allows consideration of the inaccuracy which arises from the formulation of a problem itself, i.e., from the difference between the problem which one actually tries to solve and the problem which one ideally would like

to solve. Of course, it is always possible to take $\tilde{A} = A$ and $\tilde{b} = b$, which is the traditional way. Proceeding from this point of view, we denote the inaccuracy in the evaluation of $\tilde{F}(x_k)$ by $\tilde{N}(x_k)$ and have $F_k = \tilde{F}(x_k) + \tilde{N}(x_k) = F(x_k) + N(x_k)$, where

$$N(x_k) = \tilde{F}(x_k) - F(x_k) + \tilde{N}(x_k). \quad (3.4)$$

Under the supposition that $\epsilon_1 |x_*| \leq r_A |x_k - x_*|$, we derive reasonable relative bounds on $|\tilde{F}(x_k) - F(x_k)|$ and $|\tilde{N}(x_k)|$. In order to arrive at relatively simple inequalities which embody essential qualitative features, a few somewhat generous substitutions are made below. Also, it is assumed from here on that either $b \neq 0$ or $\tilde{b} = b$ in order that the error in $\tilde{b}$ as an approximation of b be expressible in relative terms. First, one has

$$|\tilde{F}(x_k) - F(x_k)| \leq |\tilde{A} - A| |x_k| + |\tilde{b} - b|; \quad (3.5)$$

and, since $|x_k - x_*| \leq |A|^{-1} |F(x_k)|$ and $|b| \leq |A| |x_*| \leq \frac{r_A}{\epsilon_1} K(A) |F(x_k)|$,

$$|\tilde{A} - A| |x_k| + |\tilde{b} - b| \leq (1 + \frac{r_A}{\epsilon_1}) [|\tilde{A} - A| |A|^{-1}| + K(A) |\tilde{b} - b| |b|^+] |F(x_k)|, \quad (3.6)$$

where the superscript "+" denotes pseudo-inverse, i.e., $|b|^+ = 0$ if $b = 0$ and $|b|^+ = |b|^{-1}$ otherwise. It follows from (3.6) that

$$|\tilde{F}(x_k) - F(x_k)| \leq (1 + \frac{r_A}{\epsilon_1}) [|\tilde{A} - A| |A|^{-1}| + K(A) |\tilde{b} - b| |b|^+] |F(x_k)|. \quad (3.7)$$

Now, to bound $\tilde{N}(x_k)$, we make from here on the assumptions of Moler (1967) that $n\epsilon_2 \leq 0.1$ and $1.06(1 + \epsilon_1) \leq 1.2$ and obtain from the inequality preceding his (7), p. 318, that

$$|\tilde{N}(x_k)| \leq \epsilon_1 |\tilde{F}(x_k)| + 1.2 \epsilon_2 |\tilde{b}| + 1.2(n+2)\epsilon_2 |\tilde{A}| |x_k|. \quad (3.8)$$

Using the triangle inequality, (3.5), and (3.6), one sees from (3.8) that

$$\begin{aligned}
 |\tilde{N}(x_k)| &\leq \epsilon_1 |F(x_k)| + 1.2\epsilon_2 |b| + 1.2(n+2)\epsilon_2 |A| |x_k| \\
 &\quad + [\epsilon_1 + 1.2(n+2)\epsilon_2] (|\tilde{A} - A| |x_k| + |\tilde{b} - b|) \\
 &\leq \{[\epsilon_1 + 1.2(n+3)(1 + r_A/\epsilon_1)\epsilon_2 K(A)] \\
 &\quad + [\epsilon_1 + 1.2(n+2)\epsilon_2](1 + r_A/\epsilon_1)[|\tilde{A} - A| |A|^{-1} + K(A)|\tilde{b} - b| |b|^+]\} |F(x_k)|.
 \end{aligned} \tag{3.9}$$

From (3.4), (3.7), and (3.9), one obtains the following:

If $\epsilon_1 |x_*| \leq r_A |x_k - x_*|$, then $|N(x_k)| \leq \epsilon_F |F(x_k)|$, where

$$\begin{aligned}
 \epsilon_F &= [\epsilon_1 + 1.2(n+3)(1 + r_A/\epsilon_1)\epsilon_2 K(A)] \\
 &\quad + [1 + \epsilon_1 + 1.2(n+2)\epsilon_2](1 + r_A/\epsilon_1)[|\tilde{A} - A| |A|^{-1} \\
 &\quad + K(A)|\tilde{b} - b| |b|^+].
 \end{aligned} \tag{3.10}$$

To proceed toward a detailed statement of the convergence properties of Iteration 3.1, we define for convenience

$$\sigma = \frac{1 + \epsilon_1}{1 - \rho\epsilon_1 K(B)}; \tag{3.11}$$

$$\tau = [1 + \epsilon_1 + 1.2(n+2)\epsilon_2][|\tilde{A} - A| |A|^{-1} + K(A)|\tilde{b} - b| |b|^+]; \tag{3.12}$$

$$\begin{aligned}
 C_1 &= \sigma\{[\epsilon_1 + 1.2(n+3)\epsilon_2 K(A) + \tau]|B|^{-1} ||A| + \rho K(B)\epsilon_1 |B|^{-1} A|\} \\
 &\quad + (1 + \epsilon_1)|I - B|^{-1} A|;
 \end{aligned} \tag{3.13}$$

$$C_2 = 1 + \frac{\sigma}{\epsilon_1}[1.2(n+3)\epsilon_2 K(A) + \tau]|B|^{-1} ||A|. \tag{3.14}$$

With μ and ϵ_F given by (3.3) and (3.10), respectively, (3.2) becomes

$$1 > r_0 = C_1 + C_2 r_A. \quad (3.15)$$

Note that in order for (3.15) to be satisfied, it is necessary that $C_1 < 1$ and $r_A < r_A^* = (1 - C_1)/C_2$.

Our results for Iteration 3.1 in the above context are stated in the following theorem, the proof of which is immediate from the preceding discussion.

Theorem 3.2 Let ρ be such that $\rho \epsilon_1 K(B) < 1$ and let σ , τ , C_1 , and C_2 be defined by (3.11)-(3.14). For any x_0 , the iterates $\{x_k\}$ produced by Iteration 3.1, with $|B^{-1}A_k| \leq \rho \epsilon_1 K(B)$ and $|\Gamma_k| \leq \epsilon_1$, are well-defined and satisfy the following: If $C_1 < 1$ and $r_A < r_A^* = (1 - C_1)/C_2$ and if $\epsilon_1 |x_*| \leq r_A |x_k - x_*|$ for some k , then $|x_{k+1} - x_*| \leq (C_1 + C_2 r_A) |x_k - x_*|$.

We conclude with some remarks about Theorem 3.2. Under the assumption of the theorem, if $C_1 < 1$ and $r_A < r_A^* = (1 - C_1)/C_2$, then eventually a relative bound

$$\frac{|x_k - x_*|}{|x_*|} < \epsilon_1 / r_A$$

will be achieved for some k (provided $x_* \neq 0$). Note in particular that the infimum of such relative bounds is ϵ_1 / r_A^* and that the q -linear convergence rate constant $(C_1 + C_2 r_A)$ approaches 1 as r_A approaches r_A^* . Also, C_1 and C_2 both decrease as ϵ_2 is decreased; therefore, greater rates of improvement and greater limiting accuracies can be obtained with smaller values of ϵ_2 . Of course, the practical worth of taking ϵ_2 to be anything other than either ϵ_1 or the double precision choice $\epsilon_2 = \epsilon_1^2$ is questionable in most situations because of the cost of doing so. However, it is not difficult to imagine circumstances (especially

in the case of iterative refinement as considered by Moler, in which $A = B = \tilde{A}$ and $b = \tilde{b}$) such that a very small choice of ϵ_2 results in much faster convergence and much greater limiting accuracy being assured by Theorem 3.2 than the choices $\epsilon_2 = \epsilon_1$ and $\epsilon_2 = \epsilon_1^2$.

4. Acknowledgements.

We wish to thank Richard Byrd for goading us into this research. Since we obtained these results in the summer of 1982, we have had the benefit of discussing them with too many colleagues to name. Such discussions, especially with Trond Steihaug and Richard Tapia, were a great help in formulating the results to better communicate their implications.

References.

- C.G.Broyden, J.E.Dennis Jr., and J.J.More', "On the local and superlinear convergence of quasi-Newton methods", J.Inst.Maths.Applics. 34 (1973) 223-245.
- L.Collatz, Functional Analysis and Numerical Analysis, (Academic Press, New York, 1966).
- J.E.Dennis Jr. and R.B.Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, (Prentice-Hall, Englewood Cliffs, 1983).
- J.E.Dennis Jr. and H.F.Walker, "Convergence theorems for least-change secant update methods", SIAM J.Numer.Anal., 18 (1981) 949-987.
- T.Glad and A.Goldstein, "Optimization of functions whose values are subject to small errors", B.I.T., 17 (1977) 160-169.
- P.Lancaster, "Error analysis for the Newton-Raphson method", Numer.Math., 9 (1966) 55-68.
- C.B.Moler, "Iterative refinement in floating-point", J.A.C.M., 14 (1967) 316-321.
- J.M.Ortega and W.C.Rheinboldt, Iterative Solution of Nonlinear Equations In Several Variables, (Academic Press, New York, 1970).

- M. Urabe, "Error estimation in numerical solution of equations by iterations process", J. Sci. Hiroshima Univ., 19, (1956) 479-489.
- J. B. Wilkinson, The Algebraic Eigenvalue Problem, (Oxford Univ. Press, Oxford, 1965).
- P. Young, "Optimization in the presence of noise--a guided tour", in: L. C. W. Dixon, ed., Optimization in Action, (Academic Press, London, 1976).
- T. J. Ypma, "The effect of rounding errors on Newton-like methods", Inst. Maths. Applics. J. Numer. Anal. (to appear).