Numer. Math. 23, 371—385 (1975)

An Iterative Procedure for the Solution of Constrained Nonlinear Equations with Application to Optimization Problems

Steve F. McCormick*

Received July 23, 1973

Abstract. Let H_1 and H_2 denote Hilbert spaces and suppose that D is a subset of H_1. This paper establishes the local and linear convergence of a general iterative technique for finding the zeros of $G: D \to H_2$ subject to the general constraint $P(x) = x$, where $P: D \to D$. The results are then applied to several classes of problems, including those of least squares, generalized eigenvalues, and constrained optimization. Numerical results are obtained as the procedure is applied to finding the zeros of polynomials in several variables.

I. Introduction

A great deal of work has been published (cf. [1, 2, 4, 5, 8–10]) on the theoretical verification of general iterative techniques for the solution of constrained optimization problems in a Hilbert space setting. Underlying the typical convergence results for variational gradient techniques is the crucial assumption of positive definiteness of the second derivative of the objective functional. And although such theory is useful in attacking a large class of problems, this assumption often proves to be a severe one as is evidenced by the next two well-known examples.

The first is the bounded linear operator equation

$$Tx = f \tag{1}$$

where T maps the Hilbert space H into itself. With the inner product on H denoted by $\langle \cdot, \cdot \rangle$ and the induced norm by $\|\cdot\|$ we can attempt to solve (1) by optimizing either the functional

$$F(x) = \langle Tx, x \rangle - 2\langle f, x \rangle \tag{2}$$

or

$$F(x) = \|Tx - f\|^2. \tag{3}$$

The second example is the eigenvalue problem

$$Tx = \lambda x, \quad \|x\| = 1 \tag{4}$$

which may be solved by applying a gradient-restoration scheme to the "angular" measure

$$F(x) = \|\langle x, x \rangle Tx - \langle Tx, x \rangle x\|^2. \tag{5}$$

* This work was supported by NSF grant GJ 34737.

In both cases classical theory is unable to establish a very general setting in which either of these approaches is convergent although quite general settings do exist (cf. [7, 11, 3, 6]). The difficulty is the possible lack of positive definiteness of F in any of the cases. The purpose of this paper is to overcome this deficiency by first reformulating the authors previous results in order to provide a more tractable theory and then applying it to a general problem that specializes to (1) or (4). It may be observed that although the new theory presented in section 3 appears to be a mere specialization of previous work of the authors (see Section 2), it is actually a significant extension of these results in the sense that less assumptions on the underlying problem is required. That is, when the analysis of [6] is applied to a reformulation of the problem itself (see (14) below), then a much wider range of applicability is realized and the inherent deficiency of a generally incomputable step size is overcome (see (16) below).

The unconstrained form of the algorithm (see (15)–(16) below) we consider is not new and was first considered at about the same time by Altman [1] and Fridman [4]. Even in this simplified setting, however, a full development of the algorithm was not made. More specifically, in addition to addressing ourselves to questions pertaining to asymptotic rates of convergence, implementation, application, and stability, reliance here is on less restrictive assumptions which is a crucial ingredient for the special cases that will be considered. (See remark 1 below.)

II. Some Preliminaries

In this section we set up the notation and present two theorems that provide a foundation for the remaining sections. Since the results have been extracted in an appropriate form from a previous paper [6], the proofs will be omitted.

We assume throughout this paper that H represents a real (or complex) Hilbert space and let G and P denote operators mapping a non-empty subset D of H into H. The problem that concerns us in this section has the general form

$$\begin{aligned} G(x) &= 0 \\ P(x) &= x \end{aligned} \qquad x \text{ in } D. \tag{6}$$

The form of the procedure we wish to examine for the solution of (6) depends on a sequence of "step-size" functionals s_n: $D \to R$ (or C, the complex numbers) and is given by

$$x_{n+1} = P\big(x_n - s_n(x_n)\, G(x_n)\big) \qquad n = 0, 1, 2 \ldots \tag{7}$$

where the iteration requires an initial approximation, x_0, to the solution of (6). Before we can properly discuss the convergence properties of (7) it is necessary to specify some conditions on the various elements of this problem.

Throughout this section we assume the existence of an $\varepsilon > 0$ and nonempty subsets E and N of D such that the following are true:

i) $N = \{x \in H: \|x - E\| < \varepsilon\}$ where the notation

$$\|x - E\| = \inf\{\|x - u\|: u \in E\};$$

ii) Letting $P(N)=\{P(x): x\in N\}$ and $Z_N=P(N)\cap(N-E)$, then for each x in Z_N there exists a unique u_x in E such that $\|\Delta x\|=\|x-E\|$ where we denote $\Delta x=x-u_x$;

iii) E is a set of solutions of (6), that is, $G(u)=P(u)-u=0$ for all u in E;

iv) There exists a function ϱ: $R^+\to R^+$ so that $\varrho(\gamma)\to 1$ as $\gamma\to 0^+$ and

$$\|P(x)-P(u)\|\leqq\varrho(\|x-u\|)\,\|x-u\| \tag{8}$$

for all x in N and u in E;

v) There exists a $\gamma_N>0$ so that $G(x)\neq 0$ and

$$|\langle G(x),\Delta x\rangle|\geqq\gamma_N\|G(x)\|\,\|\Delta x\| \tag{9}$$

for all x in Z_N; and

vi) For some sequence of numbers $(\sigma_n: n=0, 1, 2\ldots)$ that satisfy

$$\sigma\leqq\sigma_n\leqq 2-\sigma \qquad n=0, 1, 2\ldots) \tag{10}$$

for some $\sigma>0$, the sequence $(s_n: n=0, 1, 2\ldots)$ satisfies

$$s_n(x)=\sigma_n\left(\frac{\langle\Delta x, G(x)\rangle}{\|G(x)\|^2}+\frac{o(\|\Delta x\|)}{\|G(x)\|}\right) \tag{11}$$

for all x in Z_N and $n=0, 1, 2\ldots$ [Henceforth, the notation $o(\|\Delta x\|)$ will collectively represent any quantity that satisfies

$$\lim_{\|\Delta x\|\to 0}\frac{o(\|\Delta x\|)}{\|\Delta x\|}=0$$

where the limit is taken as x ranges over the set Z_N. Note the implication of uniformity with respect to x in Z_N.]

A neighborhood, N, that satisfies condition i) for some $\varepsilon>0$ is referred to as an ε-neighborhood of E. Let $\sigma>0$ and $0<\gamma_N\leqq 1$ and define $k_\sigma=(1-\sigma(2-\sigma)\gamma_N^2)^{\frac{1}{2}}$. Note that $0\leqq k_\sigma<1$ for $\sigma<2$.

Theorem 1. Let $G, P, E, N,$ and $(s_n: n=0, 1, 2\ldots)$ satisfy conditions i) through vi) above. Then there exists an ε-neighborhood N' contained in N such that for any x_0 in $Z_{N'}$ the sequence $(x_n: n=0, 1, 2\ldots)$ given by (7) is well defined and satisfies the inequality

$$\|\Delta x_{n+1}\|\leqq(k_\sigma+\varepsilon_n)\,\|\Delta x_n\| \qquad n=0, 1, 2, \ldots \tag{12}$$

where $\varepsilon_n\to 0$ as $n\to\infty$. That is, the algorithm defined by (7) is locally and linearly convergent.

Remark 1. If $G(x)$ is the gradient of some functional $F(x)$ and if $P(x)$ is the identity operator on D, then (6) represents one formulation of the unconstrained optimization problem applied to the functional F. In this case theorem 1 provides a fairly simple, though improtant, modification of standard theory on gradient techniques. Under some differentiability assumptions on F, the usual results call for the condition that

$$\langle F''(x)h, h\rangle\geqq\gamma\langle h, h\rangle$$

for some $\gamma > 0$ and all h in H. It is particularly restrictive to require that this condition be met for all h in H. (Witness the first example of Section I.) According to Theorem 1 it is sufficient that the condition holds for all h in the set $\{\Delta x: x \text{ in } Z_N\}$. This relaxation of the positive definiteness condition becomes much more significant when constraints are included in the problem.

The next section will make use of the following definition of differentiability of the operator G.

Definition. Let E, N, and Z_N satisfy conditions i), ii), and iii) above. Then G is said to be *uniformly* (E, N, Z_N)*-differentiable* if for all x in Z_N, the Frechét derivative, $G'(x)$, exists and satisfies

$$G(x) = G'(x)\Delta x + o(\|\Delta x\|)$$

and

$$\|G'(x)\| \leqq M$$

for some $M < \infty$.

The next theorem provides a condition sufficient for such differentiability that is weaker than the usual second derivative assumption.

Theorem 2. Suppose that E, N, and Z_N satisfy conditions i), ii), and iii) and that $G'(u_x + \theta \Delta x)$ exists for all x in Z_N and $0 \leqq \theta \leqq 1$. Suppose, also, that the set $\{G'(u_x): x \text{ in } Z_N\}$ is uniformly bounded and that

$$\lim_{\varepsilon \to 0} \sup\{\|G'(u_x + \theta \Delta x) - G'(u_x)\|: x \text{ in } Z_N, \|\Delta x\| < \varepsilon, 0 \leqq \theta \leqq 1\} = 0.$$

Then G is uniformly (E, N, Z_N)-differentiable.

Remark 2. A similar differentiability assumption for $P(x)$ involves the ability to write

$$P(x) = P(u) + P'(u)(x - u) + o(\|x - u\|)$$

for x in N and u in E. It is easily seen that for P to satisfy condition iv) it is sufficient that this be true for $P(x)$ and that $\|P'(u)\| \leqq 1$ for all u in E.

III. The Problem and Procedure

Let H_1 and H_2 denote real (or complex) Hilbert spaces whose inner-product in each case is denoted by $\langle \cdot, \cdot \rangle$ and induced norm by $\|\cdot\|$. In addition, suppose that m is a positive integer and that $G_1, G_2, \ldots, G_{m-1}$ and G_m represent operators mapping a subset, D, of H_1 into H_2. Suppose also that P maps D into itself. Then the problem that we are henceforth concerned with is given by

$$\begin{aligned} G_k(x) &= 0 \\ P(x) &= x \end{aligned} \qquad k = 1, 2, \ldots, m;\ x \text{ in } D. \tag{13}$$

Such a problem can be thought of as a general form of a system of nonlinear equations with constraints. [At this point it may be of benefit to the reader to skip ahead to the discussion immediately following the corollaries of theorem 3.]

We will assume sufficient differentiability of the G_k to allow us to define the operator G on a subset of D by

$$G(x) = \sum_{k=1}^{m} G_k'^*(x) G_k(x) \tag{14}$$

where $G_k'^*$ denotes the adjoint of the Frechét derivative of G_k. The method proposed for the solution of (13) is then given formally by

$$x_{n+1} = P\left(x_n - s(x_n) G(x_n)\right) \qquad n = 0, 1, 2, \ldots \tag{15}$$

where the step-size functional s is defined on a subset of D by

$$s(x) = \frac{\sum_{k=1}^{m} \|G_k(x)\|^2}{\|G(x)\|^2}. \tag{16}$$

Notice that (15) is a gradient-projection-like scheme applied to minimizing the functional

$$F(x) = \tfrac{1}{2} \sum_{k=1}^{m} \|G_k(x)\|^2 \tag{17}$$

subject to the constraint $P(x) = x$. This results from the observation that $G(x)$ is the formal gradient of $F(x)$. It will now be shown that (15) and (16) describe a convergent process under some assumptions.

Theorem 3. Let E denote a nonempty set of solutions of (13) and assume the existence of an ε-neighborhood, N, of E that satisfies conditions i) and ii) of section II. Suppose, also, that P satisfies condition iv). Finally, let each $G_k(x)$ be uniformly (E, N, Z_N)-differentiable and assume the existence of a $\gamma > 0$ such that

$$\sum_{k=1}^{m} \|G_k'(x) \Delta x\|^2 \geqq \gamma^2 \|\Delta x\|^2 \tag{18}$$

for all x in Z_N. Then there exists an ε-neighborhood N' contained in N such that the sequence $(x_n : n = 0, 1, 2, \ldots)$ in (15) is well defined and satisfies the inequality

$$\|\Delta x_{n+1}\| \leqq k \|\Delta x_n\| \qquad n = 0, 1, 2, \ldots \tag{19}$$

where $0 \leqq k < 1$. Thus, the algorithm (15)–(16) is locally and linearly convergent to the set, E, of solutions of (13).

Proof. The only difficulty here is to demonstrate that v) and vi) are valid which we first do for the case $m = 1$, i.e., $G(x) = G_1'^*(x) G_1(x)$. Remembering the differentiability assumption on G_1 it then follows for x in Z_N that

$$\begin{aligned} \langle \Delta x, G(x) \rangle &= \langle \Delta x, G_1'^*(x) G_1(x) \rangle \\ &= \langle G_1'(x) \Delta x, G_1(x) \rangle \\ &= \|G_1(x)\|^2 + o(\|\Delta x\|) \|G_1(x)\|. \end{aligned} \tag{20}$$

Moreover,

$$\begin{aligned} \|G_1(x)\| &= \|G_1'(x) \Delta x\| + o(\|\Delta x\|) \\ &\geqq \gamma \|\Delta x\| + o(\|\Delta x\|). \end{aligned} \tag{21}$$

Letting $M = \sup\{\|G'(x)\|\colon x \text{ in } Z_N\}$, (20) and (21) combine to yield

$$\begin{aligned} |\langle \Delta x, G(x)\rangle| &\geqq \left(\gamma\|\Delta x\| + o(\|\Delta x\|)\right)\|G_1(x)\| \\ &\geqq \left(\frac{\gamma}{M} + \frac{o(\|\Delta x\|)}{\|\Delta x\|}\right)\|\Delta x\|\,\|G_1'^{*}(x)\,G_1(x)\|. \end{aligned} \tag{22}$$

Therefore, v) is true for a proper choice of N'. To prove vi), note that by (22) we have

$$\begin{aligned} \|G(x)\| &\geqq \frac{1}{\|\Delta x\|}\langle \Delta x, G(x)\rangle \\ &\geqq \left(\gamma + \frac{o(\|\Delta x\|)}{\|\Delta x\|}\right)\|G_1(x)\|. \end{aligned} \tag{23}$$

Hence, from (20) and (23) it follows that

$$\begin{aligned} s(x) &= \frac{\|G_1(x)\|^2}{\|G(x)\|^2} \\ &= \frac{\langle \Delta x, G(x)\rangle}{\|G(x)\|^2} + \frac{o(\|\Delta x\|)\,\|G_1(x)\|}{\|G(x)\|^2} \\ &= \frac{\langle \Delta x, G(x)\rangle}{\|G(x)\|^2} + \frac{o(\|\Delta x\|)}{\|G(x)\|}. \end{aligned}$$

This establishes the assertions of the theorem in the case that $m = 1$. The proof for $m > 1$ is now a simple exercise and we omit it.

The following corollaries are immediate.

Corollary 1. If x_0 is chosen so that $\|\Delta x_0\|$ is small, then a crude estimate for the linear convergence factor

$$k = \lim_{n\to\infty} \sup \frac{\|\Delta x_{n+1}\|}{\|\Delta x_n\|} \tag{24}$$

is provided by the inequality

$$k^2 \leqq 1 - \gamma^2/\sum_k M_k^2 \tag{25}$$

where $M_k = \sup\{\|G_k'(x)\|\colon x \text{ in } Z_N\}$ $1 \leqq k \leqq m$.

Corollary 2. If each $G_k'(x)$ is sufficiently continuous at E so that $G_k'(x)\Delta x = G_k'(u_x)\Delta x + o(\|\Delta x\|)$, then (18) can be replaced by the condition that

$$\sum_k \|G_k'(u_x)\Delta x\|^2 \geqq \gamma^2\|\Delta x\|^2 \qquad x \text{ in } Z_n. \tag{26}$$

In particular, this is true when the $G_k'(x)$ are uniformly continuous on Z_N. Morever, suppose there exists an $\varepsilon > 0$ such that the set $Z_N^{\varepsilon} = \{x \text{ in } Z_N\colon \varepsilon \leqq \|\Delta x\| \leqq 2\varepsilon\}$ is nonempty and compact. (This requires H to be finite dimensional in general.) Suppose, also, that the subsets $\{(u_x, \overline{\Delta x})\colon x \text{ in } Z_N\}$ and $\{(u_x, \overline{\Delta x})\colon x \text{ in } Z_N^{\varepsilon}\}$ of $E \times H$ are equal and that the function $\phi\colon x \to u_x$ is continuous on Z_N^{ε}. Then condition (18) can be replaced by the requirement that there exist at least one k in $\{1, 2, \ldots, m\}$ for each x in Z_N such that

$$G_k'(u_x)\Delta x \neq 0. \tag{27}$$

In particular, this is true when P is the identity and N is convex and has compact closure.

Remark 3. Although the method (15)–(16) enjoys the stability suggested by the inclusion of σ_n in the considerations of section II, a more significant kind of stability exists and it is important for the next section to make note of it. In particular, let $W_k(x)$ represent some approximation to $G_k'^*(x)G_k(x)$ so that $W_k(x) = G_k'(x)G_k(x) + o(\|\Delta x\|)$. Then if $W_k(x)$ is used in place of $G_k'^*(x)G_k(x)$ in the iteration (15)–(16), convergence is nevertheless guaranteed and results identical to those in theorem 3 are valid. Moreover, with the assumptions of Theorem 3, suppose that $\varepsilon_k(x)\colon H_1 \to H_2$ for each x in D and $1 \leqq k \leqq m$ and that

$$\varepsilon_N = \max_k \left(\sup_{x \in Z_N} \|\varepsilon_k(x)\|\right)$$

is sufficiently small. Then convergence of the approximate iteration

$$x_{n+1} = P\left(x_n - \frac{\sum_{k=1}^{m} \|G_k(x_n)\|^2}{\|y_n\|^2} y_n\right)$$

where we let

$$y_n = \sum_{k=1}^{m} \left(G_k'^*(x_n) + \varepsilon_k^*(x_n)\right) G_k(x_n)$$

follows as in Theorem 3 with the estimate of k in corollary 1 altered accordingly. Now we have shown that x_n converges to the set E. It also follows that there exists a u in E to which the sequence (x_n) converges. [It is observed that (x_n) is a Cauchy sequence in H_1 by noting that for some k in the interval $(0, 1)$

$$\begin{aligned} \|x_{n+p} - x_n\|^2 &\leqq \sum_{j=1}^{p} \|x_{n+j} - x_{n+j-1}\|^2 \\ &\leqq \sum_{j=1}^{p} k^j \|x_{n+1} - x_n\|^2 \\ &\leqq \left(\frac{k}{1-k}\right)(\|\Delta x_{n+1}\|^2 + \|\Delta x_n\|^2).] \end{aligned}$$

The stability of (15)–(16) in this sense now implies the local linear convergence of the approximate iteration given above where we now use any subsequence (x_{l_n}) of (x_n) and define

$$y_n = \sum_{k=1}^{m} G_k'^*(x_{l_n}) G_k(x_n).$$

If we define the subsequence wisely, we appreciably reduce the need to compute the derivatives of G_k at each iteration.

Remark 4. The scope of applicability of Theorem 3 is suggested by the following examples.

a) *A simultaneous set of nonlinear equations in several unknowns.* If the problem is that of finding the zeros of $f\colon R^m \to R^l$ (or $C^m \to C^l$), then iteration (15)–(16) becomes

$$x_{n+1} = x_n - \frac{\sum_k |f_k(x_n)|^2}{\left\|\sum_k f'^*(x_n) f_k(x_n)\right\|^2} \sum_k f'^*(x_n) f_k(x_n)$$

where we use the notation $f=(f_1, f_2, \ldots, f_l)$. If E is convex and bounded, a condition sufficient for local convergence is that each f_k' is uniformly continuous on N and that, for each x in $N-E$, at least one of the numbers $f_k'^*(u_x)\Delta x$, $1 \leqq k \leqq l$, is nonzero. Indeed, with these assumptions we can guarantee the convergence of various computational modifications of the above procedure. As an example, we might modify the procedure by successively involving only the equation that causes the most trouble. More specifically, the iteration is

$$x_{n+1}=x_n-\frac{f_r(x_n)\,f_r'^*(x_n)}{\|f_r'^*(x_n)\|^2}$$

where $r=r(x_n)$ is chosen to maximize $|f_k(x_n)|$ over $k=1, 2, \ldots, l$. Another modification is defined by chosing r while rotating through the numbers $k=1, 2, \ldots, l$ to be such that $|f_k(x_n)|$ is greater than a preassigned epsilon. When no such r exists, the threshold epsilon would be reduced and the process continued until sufficient accuracy (e.g., small epsilon) is attained.

Note that in the scalar case $f: R \to R$ the iteration reduces to Newton's method and can therefore be thought of as one form of its generalization (cf. [1, 4]).

b) *Singular linear equations.* Suppose that $T: H_1 \to H_2$ is a bounded linear operator with closed range $R(T)$. Then the application of (15)–(16) to problem (1) with f in $R(T)$ yields the iteration

$$x_{n+1}=x_n-\frac{\|Tx_n-f\|^2}{\|T^*(Tx_n-f)\|^2}\,T^*(Tx_n-f).$$

Letting Q denote the null space of T, if u is a particular solution of (1) then the linear variety $E=u+Q$ is the set of all solutions of (1). It is then possible to let $N=H_1$ be an ε-neighborhood of E. Because $R(T)$ is closed it can be shown that

$$\|Th\| \geqq \gamma \|h\|$$

for some $\gamma>0$ and all h orthogonal to Q. Observing that Δx is orthogonal to Q for each x in H_1, it follows that condition (18) is satisfied and Theorem 3 applies to establish convergence of x_n to E for arbitrary x_0 in H_1.

c) *The generalized linear eigenvalue problem.* Suppose that T_1 and T_2 are bounded linear operators mapping H_1 into H_2. Then the generalized linear eigenvalue problem can be written

$$\begin{aligned} T_1 x &= \lambda T_2 x \\ \|T_2 x\| &= 1 \end{aligned} \qquad x \text{ in } H_1. \tag{28}$$

From what we will learn in the next section, there are at least two formulations of (28) that allow for the effective application of iteration (15)–(16). Both depend on letting

$$G(x)=\langle T_2 x, T_2 x\rangle T_1 x-\langle T_2 x, T_1 x\rangle T_2 x \qquad x \text{ in } D;$$

and differ by the definitions

$$P(x)=\frac{x}{\|T_2 x\|} \qquad x \text{ in } D; \quad T_2 x \neq 0;$$

and

$$\psi(x)=\|T_2 x\|^2-1 \qquad x \text{ in } D.$$

The first formulation is gotten by choosing $m=1$ and $G_1(x)=G(x)$ and using the definition of P given above. The second arises from letting P be the identity operator on D and defining G_1: $D \to H_2 \oplus R$ (or $H_2 \oplus C$) by

$$G_1(x) = \big(G(x), \psi(x)\big) \qquad x \text{ in } D.$$

Writing

$$f(x) = \langle T_1 x, T_1 x\rangle \langle T_2 x, T_2 x\rangle - |\langle T_2 x, T_1 x\rangle|^2$$

and

$$g(x) = \langle T_1 x, T_1 x\rangle T_2^* T_2 x + \langle T_2 x, T_2 x\rangle T_1^* x T_1 x - \langle T_2 x, T_1 x\rangle (T_2^* T_1 x + T_1^* x T_2 x)$$

and using a simplification, the procedures are, respectively,

$$x_{n+1} = P\left(x_n - \frac{f(x_n)\, g(x_n)}{\|g(x_n)\|^2}\right)$$

and

$$x_{n+1} = x_n - \frac{\alpha g(x_n) + |\psi(x_n)|^2}{\|\alpha g(x_n) + \psi(x_n)\psi'^*(x_n)\|^2} \left(\alpha g(x_n) + \psi(x_n)\psi'^*(x_n)\right)$$

where $\alpha = \langle T_2 x_n, T_2 x_n\rangle$. It is not too difficult now to show that the assumptions needed for convergence are satisfied is this instance when E represents a set of eigenvectors belonging to an eigenvalue λ of (28) that are properly normalized, if E does not intersect the null space of T_2, and provided for any u in E there exists no h in $E^{\perp}$ that satisfies $(T_1 - \lambda T_2)h = T_2 u$. (The reader is referred to [3, 11] for related work.)

When the operators T_1 and T_2 are nonlinear the assumptions required for convergence do not translate quite as nicely as they do for example c above, as is to be expected. The conditions do have a similar appearance, however. Since the nonlinear case is a basis for constrained optimization, we devote the next section to this example.

IV. Minimization with Equality Constraints

The form of the constrained optimization problem we wish to consider involves two real functionals, F and ψ, defined on a subset, D, of H. With F the *objective* and ψ the *constraint* functional, the problem is that of finding and element u of the constraint set $C = \{x \text{ in } D: \psi(x) = 0\}$ that satisfies

$$F(u) = \min_{x \in C} F(x). \tag{29}$$

[We note here that a minimization problem with several equality constraints $\psi_1(x) = \psi_2(x) = \cdots = \psi_m(x) = 0$ is of this form if we take $\psi(x) = \sum_k \psi_k^2(x)$.] To pose (29) in the proper formulation we define G: $D \to H$ where possible by

$$G(x) = \langle \nabla\psi(x), \nabla\psi(x)\rangle \nabla F(x) - \langle \nabla F(x), \nabla\psi(x)\rangle \nabla\psi(x). \tag{30}$$

[Note the relationship that G shares with the Lagrangian of the pair (F, ψ).] It is well-known (cf. [1]) that solutions of (29) are precisely the zeros of G in the set C under certain conditions. Thus, if P: $D \to H$ so that $C = \{x \in D: P(x) = x\}$, then (29) is equivalent to problem (13) in this case. It is therefore with (29) in mind that we devote this section to the following problem formulation.

Let g_1 and g_2 denote mappings from a subset, D, of H_1 into H_2. Suppose that $P: D \to D$ and define $G: D \to H_2$ by

$$G(x) = \langle g_2(x), g_2(x)\rangle g_1(x) - \langle g_2(x), g_1(x)\rangle g_2(x). \tag{31}$$

The problem we consider in this section is given by

$$\begin{aligned} G(x) &= 0 \\ P(x) &= x \end{aligned} \qquad x \text{ in } D \tag{32}$$

and the method by

$$x_{n+1} = P\left(x_n - \frac{\|G(x_n)\|^2}{\|G'^*(x_n)\,G(x_n)\|^2} G'^*(x_n)\,G(x_n)\right). \tag{33}$$

The next theorem establishes the convergence of iteration (33) to solution of (32).

Theorem 4. Let E be a nonempty subset of solutions of (32) exhibiting an ε-neighborhood, N, satisfying conditions i) and ii) of Section II. Suppose that P satisfies condition iv) and that g_1 and g_2 are uniformly (E, N, Z_N)-differentiable and have uniformly continuous Frechét derivatives on Z_N. Assume that g_2 is nonzero on E. Then for each x in Z_N define the scalar λ_x by

$$\lambda_x = \frac{\langle g_2(u_x), g_1(u_x)\rangle}{\langle g_2(u_x), g_2(u_x)\rangle} \tag{34}$$

and the operator $L_x: H_1 \to H_2$ by

$$L_x = g_1'(u_x) - \lambda_x g_2'(u_x). \tag{35}$$

Finally, let $M = \sup\{\|G'(u_x)\|: x \text{ in } Z_N\}$ and suppose for some $\gamma > 0$ and all x in Z_N that

$$\|\langle g_2(u_x), g_2(u_x)\rangle L_x \Delta x - \langle g_2(u_x), L_x \Delta x\rangle g_2(u_x)\| \geqq \gamma \|\Delta x\|. \tag{36}$$

Then there exists an ε-neighborhood N' contained in N such that for any x_0 in Z_N, the sequence $(x_n: n = 0, 1, 2, \ldots)$ defined by (33) satisfies the inequality

$$\|\Delta x_{n+1}\| \leqq k_n \|\Delta x_n\| \qquad n = 0, 1, 2, \ldots \tag{37}$$

where $k_n \to (1 - \gamma^2/M^2)^{\frac{1}{2}}$ as $n \to \infty$.

Proof. By the assumptions on g_1 and g_2 it follows that $G'(x)$ exists for each x in Z_N and is given by

$$\begin{aligned} G'(x)\Delta x = {} & \langle g_2(x), g_2(x)\rangle g_1'(x)\Delta x - \langle g_2(x), g_1(x)\rangle g_2'(x)\Delta x \\ & + \langle g_2'(x)\Delta x, g_2(x)\rangle g_1(x) - \langle g_2'(x)\Delta x, g_1(x)\rangle g_2(x) \\ & + \langle g_2(x), g_2'(x)\Delta x\rangle g_1(x) - \langle g_2(x), g_1'(x)\Delta x\rangle g_2(x). \end{aligned} \tag{38}$$

It is easy to show that G is uniformly (E, N, Z_N)-differentiable, by using (38) and the differentiability of g_1 and g_2. The fact that $G'(x)$ is uniformly continuous also follows from (38) and the uniform continuity of $g_1'(x)$ and $g_2'(x)$. To show that (26) is true we use the fact that $G(u_x) = 0$ and, hence,

$$\begin{aligned} G'(u_x)\Delta x = {} & \langle g_2(u_x), g_2(u_x)\rangle g_1'(u_x)\Delta x - \langle g_2(u_x), g_1(u_x)\rangle \\ & \cdot g_2'(u_x)\Delta x + \langle g_2(u_x), g_2'(u_x)\Delta x\rangle g_1(u_x) \\ & - \langle g_2(u_x), g_1'(u_x)\Delta x\rangle g_2(u_x) \\ = {} & \langle g_2(u_x), g_2(u_x)\rangle L_x \Delta x - \langle g_2(u_x), L_x \Delta x\rangle g_2(u_x). \end{aligned} \tag{39}$$

(26) is a consequence of (39) and the theorem is proved.

The following corollary concerns itself with a simplification of iteration (33) that enjoys the same properties of convergence. The simplification amounts to eliminating a small vector quantity that is unnecessary to compute, at least asymptotically.

Corollary 3. With the assumptions of Theorem 4, define the scalar functional f on D and the vector function g on D by

$$\begin{aligned} f(x) &= \langle g_1(x), g_1(x)\rangle \langle g_2(x), g_2(x)\rangle - |\langle g_2(x), g_1(x)\rangle|^2 \\ g(x) &= \langle g_2(x), g_2(x)\rangle g_1'^*(x) g_1(x) + \langle g_1(x), g_1(x)\rangle g_2'^*(x) g_2(x) \\ &\quad - \langle g_2(x), g_1(x)\rangle \left(g_2'^*(x) g_1(x) + g_1'^*(x) g_2(x)\right). \end{aligned} \tag{40}$$

Then provided $\|\Delta x_0\|$ is small, the sequence $(x_n : n = 0, 1, 2, \ldots)$ given by

$$x_{n+1} = P\left(x_n - \frac{f(x_n)\, g(x_n)}{\|g(x_n)\|^2}\right) \qquad n = 0, 1, 2, \ldots \tag{41}$$

is well-defined and satisfies inequality (37).

Proof. The proof rests on the observation that (41) is the result of the elimination of the term

$$q(x)\Delta x = \langle g_2'(x)\Delta x, g_2(x)\rangle g_1(x) - \langle g_2'(x)\Delta x, g_1(x)\rangle g_2(x)$$

when computing $G'(x)\Delta x$ as in (38). Since $q(u_x)\Delta x = 0$, it is easy to show by the assumptions on G that $q^*(x)G(x)$ is $o(\|\Delta x\|)$. The theorem is now a consequence of Remark 3.

Remark 5. The above corollary is readily interpreted as an approach to solving problem (29) with one important gap in the application. More specifically, if we define

$$\begin{aligned} g_1(x) &= \nabla F(x) \\ g_2(x) &= \nabla \psi(x) \end{aligned} \qquad x \text{ in } D, \tag{42}$$

then we are still left with making an acceptable choice for $P(x)$. Although a more appropriate selection might be made for a specific form of (29) like, say, the eingenvalue problem (see example c of the last section), in general a very reasonable choice for $P(x)$ is given by

$$P(x) = x - \frac{\psi(x)\nabla\psi(x)}{\|\nabla\psi(x)\|^2} \qquad x \text{ in } D. \tag{43}$$

If $\psi(x)$ exhibits the differentiable property of Remark 2, it can easily be seen that P satisfies condition iv) of Section II. Under these assumptions it then follows that Eqs. (40)–(43) define a locally convergent procedure for the solution of optimization problem (29).

So far the approach that has been taken for the solution of (29) has maintained the viewpoint that the problem is a constrained one and certain difficulties in analysis of convergence arise in this way. For this reason it is important to attempt to transform (29) into an unconstrained problem as we now do. It will be

seen that more can be said using this formulation without affecting the convergence properties of the iteration. In particuar, if a solution set of (29) has an ε-neighborhood with compact closure, then it will be possible to replace (36) with a condition that is much easier to verify.

Using the operator given by (31) we shall now attempt to solve

$$\begin{matrix} G(x)=0 \\ \psi(x)=0 \end{matrix} \quad x \text{ in } D \tag{44}$$

by applying the procedure described in Section III to the operator $G_1: D \to H_2 \oplus R$ (or $H_2 \oplus C$) defined by

$$G_1(x) = \big(G(x), \psi(x)\big).$$

The inner-product on the cross-product space is the natural one and the method thus has the form

$$\begin{aligned} x_{n+1} = x_n - \Big(&\frac{\|G(x_n)\|^2 + |\psi(x_n)|^2}{\|G'^*(x_n)G(x_n) + \psi(x_n)\psi'^*(x_n)\|^2}\Big) \\ &\cdot \big(G'^*(x_n)G(x_n) + \psi(x_n)\psi'^*(x_n)\big) \qquad n=0,1,2,\ldots \end{aligned} \tag{45}$$

We shall close this section with a theorem concerning the convergence of (45) and a corollary that deals with a simplification of this procedure.

Theorem 5. Let E be a nonempty subset of solutions of (44) on which g_2 is nonzero and which exhibits an ε-neighborhood, N, satisfying conditions i) and ii) of Section II. Note that $P=I$, the identify operator on D, and that $Z_N = N - E$. Suppose that g_1, g_2, and ψ are uniformly (E, N, Z_N)-differentiable and have uniformly continuous Frechét derivatives on Z_N. Let

$$M = \sup\{\big(\|G'(u_x)\|^2 + \|\psi'(u_x)\|^2\big)^{\frac{1}{2}} : x \text{ in } N - E\}$$

and define λ_x by (34) and L_x by (35). Finally, assume the existence of a $\gamma > 0$ such that

$$\|\langle g_2(u_x), g_2(u_x)\rangle L_x \Delta x - \langle g_2(u_x), L_x \Delta x\rangle g_2(u_x)\|^2 + |\psi'(x)\Delta x|^2 \geqq \gamma^2 \|\Delta x\|^2 \tag{46}$$

for all x in N. Then there exists an ε-neighborhood, N', contained in N such that the sequence $(x_n : n=0, 1, 2, \ldots)$ defined by (46) converges according to inequality (37) for any x_0 in N'. Moreover, assume that H_1 is finite dimensional and E is convex and bounded and for each x in N define E_x as the set of unit vectors in H_1 orthogonal to $\psi'^*(u_x)$. Then condition (46) is equivalent to the assumption that the problem

$$L_x h = \alpha g_2(u_x) \qquad \alpha \text{ scalar}; \; x \in N; \; h \in E_x. \tag{47}$$

has no solution.

Proof. Clearly $G_1(x)$ is uniformly (E, N, Z_N)-differentiable and its derivative is given by

$$G_1'(x)h = \big(G'(x)h, \psi'(x)h\big) \qquad x \text{ in } N; \; h \text{ in } H_1.$$

With the natural innerproduct it follows that

$$G_1'^*(x)G_1(x) = G'^*(x)G(x) + \psi(x)\psi'^*(x) \qquad x \text{ in } N.$$

The first conclusion of the theorem is then a consequence of applying Theorem 3 with $m=1$ and G_1 and P as above. The last conclusion follows from the compactness of the closure of N and the continuity of $\psi'(u_x)\overline{\Delta x}$ and

$$\langle g_2(u_x), g_2(u_x)\rangle L_x\overline{\Delta x} - \langle g_2(u_x), L_x\overline{\Delta x}\rangle g_2(u_x)$$

for x in $N-E$. The theorem is proved.

Corollary 4. Let the conditions of Theorem 5 be satisfied and let f and g be defined by (40). Consider the iteration

$$x_{n+1} = x_n - \frac{\alpha f(x_n) + |\psi(x_n)|^2}{\|\alpha g(x_n) + \psi(x_n)\psi'^*(x_n)\|^2}\left(\alpha g(x_n) + \psi(x_n)\psi'^*(x_n)\right) \tag{48}$$

where we have denote the quantity $\alpha = \langle g_2(x_2), g_2(x_n)\rangle$, $n=0, 1, 2, \ldots$. Then the sequence $(x_n: n=0, 1, 2, \ldots)$ that it generates is well defined and is convergent according to (37) provided $\|\Delta x_0\|$ is sufficiently small.

Proof. The proof is similar to that given in Corollary 3 and will be omitted.

V. Computational Results

Experimentation with the technique and its various modifications presented in this paper was accomplished in FORTRAN on a DECsystem-10 which uses a 36-bit word-length. The results given in this section represent that portion of the work that dealt with the problem of finding the roots of polynomials in the n variables $x=(x_1, x_2, \ldots, x_n)$. With m the degree they have the form

$$f(x) = \sum_{i_1+i_2+\cdots+i_n\leq m} a_{i_1 i_2 \ldots i_n} x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n}. \tag{49}$$

The following specific examples were used in the investigation:

$$f_1(x) = x_1 x_2 x_3 - 2x_2^2 x_4^2 + 4x_1^3 x_2^3 x_3^3 x_4^3 x_5^3 - 24 x_1 x_2 x_3 x_4 x_5 + 24,$$

$$f_2(x) = (x_1^2 - 4x_2^2)(3x_1 - x_2),$$

$$f_3(x) = [4x_1(x_1^2 - x_2) + 0.02(x_1 - 1)]^2 + [2(x_1^2 - x_2)]^2,$$

$$f_4(x) = x_1^2 + 4x_2^2 + 9x_3^2 + 16x_4^2,$$

$$f_5(x) = 3x_1 x_2^2 x_3^6 - 56 x_1^3 x_2^2 x_3^2 + x_1 x_2.$$

Note that in these examples we have, respectively, $n=5, 2, 2, 4$, and 3 and $m=15$, 3, 6, 2, and 9.

In each case two initial guesses were used, the components of one of which being chosen randomly in an appropriately selected interval. The iteration was stopped when ε_n, the absolute value of the polynomial, became smaller than 10^{-6} even though in several instances convergence was much more rapid past this point. The following table describes the results which, in this environment, are felt to be fairly representative of the computational performance of the procedure:

Polynomial	Initial vector	ε_0	ε_1	ε_2	ε_3
f_1	(1.00, 1.00, 1.00, 1.00, 1.00)	3×10^{2}	5×10^{-1}	6×10^{-2}	2×10^{-3}
	(−0.435, 0.379, 0.176, −0.934, 0.482)	5×10^{3}	2×10^{3}	5×10^{2}	1×10^{2}
f_2	(0.195, 0.732)	3×10^{-1}	2×10^{-2}	6×10^{-5}	1×10^{-8}
	(10.0, 1.00)	3×10^{3}	7×10^{2}	7×10^{1}	5×10^{-1}
f_3	(−1.20, 1.00)	5×10^{0}	2×10^{0}	4×10^{-1}	1×10^{-1}
	(0.358, −0.217)	7×10^{-1}	2×10^{-1}	6×10^{-2}	2×10^{-2}
f_4	(1.00, 1.00, 1.00, 1.00)	3×10^{1}	9×10^{1}	3×10^{0}	1×10^{0}
	(−0.610, 0.465, 0.217, −0.385)	4×10^{0}	1×10^{0}	5×10^{-1}	3×10^{-1}
f_5	(−0.142, 0.677, 0.751)	9×10^{-2}	2×10^{-2}	2×10^{-4}	9×10^{-11}
	(0.736, −0.661, 0.573)	4×10^{0}	1×10^{0}	5×10^{-1}	2×10^{-1}

Remark 6. The effectiveness of the technique in finding zeros of a single polynomial is appearent in the above table. Note the almost quadratic convergence characteristics for some of the examples. The performance of the algorithm as it applies to systems of nonlinear equations is a topic of current investigation and is, however, enjoying mixed success. While the method is computationally superior to other existing techniques in some cases (e.g., the problem on page 305 of [9]), it proves to be too slow in others. On the other hand, it is important to note that the modification suggested in example a of section IV compares favorably in terms of the number of iterations with the full procedure and is therefore considerably more efficient. Further advantage is expected to be gained by the use of a difference approximation to the derivatives that need be computed during the iteration. Finally, it is anticipated that the low computational costs and simplicity of the algorithms will be underscored by the evidence of their susceptibility to straight-foreward acceleration techniques. These observations are highly suggestive of the potential for the development of a generally effective process based on the schemes presented here.

ε_4	ε_5	ε_6	ε_7	ε_8	Error at Convergence
7×10^{-7}	—	—	—	—	$\varepsilon_4 = 7 \times 10^{-7}$
4×10^{1}	1×10^{1}	2×10^{0}	2×10^{-1}	1×10^{-3}	$\varepsilon_9 = 7 \times 10^{-7}$
—	—	—	—	—	$\varepsilon_3 = 1 \times 10^{-8}$
3×10^{-5}	0	—	—	—	$\varepsilon_5 = 0$
3×10^{-2}	7×10^{-3}	1×10^{-3}	2×10^{-4}	3×10^{-6}	$\varepsilon_9 = 0$
4×10^{-3}	1×10^{-3}	2×10^{-4}	3×10^{-5}	9×10^{-7}	$\varepsilon f = 9 \times 10^{-7}$
4×10^{-1}	2×10^{-1}	1×10^{-1}	1×10^{-1}	6×10^{-1}	$\varepsilon_{34} = 7 \times 10^{-7}$
2×10^{-1}	2×10^{-1}	1×10^{-1}	1×10^{-1}	5×10^{-2}	$\varepsilon_{35} = 6 \times 10^{-7}$
—	—	—	—	—	$\varepsilon_3 = 9 \times 10^{-11}$
7×10^{-2}	2×10^{-2}	4×10^{-3}	2×10^{-4}	9×10^{-7}	$\varepsilon_8 = 9 \times 10^{-7}$

References

1. Altman, M.: Connection between gradient methods and Newton's method for functionals. Bull. Acad. Polon. Sci. **9**(12), 877–880 (1961)
2. Blum, E. K.: Numerical analysis and computation: Theory and practice. Reading, Mass.: Addison-Wesley 1972
3. Blum, E. K., Rodrique, G. H.: Solution of eigenvalue problems and least squares problems in Hilbert space by a gradient method. J. Comput. Sys. Sci. **8**, 220–238 (1974)
4. Fridman, V. M.: An iteration process with minimum errors for a nonlinear operator equation. Dokl. Akad. Nauk SSR **139** (61), 1063–1066
5. Goldstein, A. A.: Constructive real analysis. New York: Harper and Row 1967
6. McCormick, S. F.: A general approach to one-step iterative methods with application to eigenvalue problems. J. Comput. Syst. Sci. **6**, 354–372 (1972)
7. McCormick, S. F., Rodrique, G. H.: A uniform approach to gradient methods for linear operator equations. To appear in J.M.A.A.
8. Ortega, J. M., Rheinboldt, W. C.: Iterative solution of non-linear equations in several variables. New York: Academic Press 1970
9. Polak, E.: Computational methods in optimization. New York: Academic Press 1971
10. Rheinboldt, W. C.: A unified convergence theory for a class of iterative processes. SIAM J. Numer. Anal. **5**, 42–63 (1968)
11. Rodrique, G. H.: A gradient method for the matrix eigenvalue problem. This Journal **22**, 1–16 (1973)

Dr. Steve McCormick
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523
U.S.A.