HYBRID CONJUGATE GRADIENT-STEEPEST DESCENT ALGORITHMS FOR UNCONSTRAINED MINIMIZATION

E. Messerli

Department of Electrical Engineering and Computer Sciences
Electronics Research Laboratory
University of California, Berkeley

Research sponsored partially by the National Aeronautics and Space Administration under Grant NGL-05-003-016 (Sup 6).

ABSTRACT

Three gradient-type procedures for unconstrained minimization are suggested. These procedures are hybrids between steepest descent and conjugate gradient algorithms, employing a design parameter to achieve adaptive sequence breaking. Essential convergence theory is presented in a unified fashion, and limited computational results are included to verify the efficacy of the form of the procedures. The computational results suggest that a normalized form of the Fletcher-Reeves algorithm is preferable to the original form.

I. INTRODUCTION

For the minimization of unconstrained functions, computational evidence suggests that algorithms which combine features of steepest descent and conjugate gradient algorithms may be effective. For example, it is well known that the Fletcher-Powell algorithm [1] often performs much worse than steepest descent far from a minimum. In common with most conjugate gradient methods, this behavior is usually countered by incorporating periodic steepest descent steps (sequence breaking) in the algorithm, leading to a hybrid algorithm. Many other possibilities for hybrid algorithms exist. Unfortunately, at the present time the design of good algorithms of this type is at best ad hoc, often based on heuristic results not amenable to a clear theoretical statement. We know why steepest descent works; we know it converges slowly near a minimum. We know how and why conjugate gradient methods work for quadratic functions; we know why they exhibit rapid final convergence for general functions. We do not know what overall improvement in convergence might be possible by an appropriate interleaving, or modification, of these techniques.

This paper does not purport to rectify the preceding situation. Rather, in recognition of the wide latitude with which search directions compatible with convergence may be chosen, several algorithms of comparable complexity are suggested. These algorithms employ a design parameter to produce adaptive sequence breaking. Convergence theory applicable to a wide class of algorithms is developed in conjunction with the algorithms.

Limited computational results are presented, including comparative results on steepest descent and standard conjugate gradient methods. These results confirm that non-conjugate gradient techniques can manifest

quite good behavior for functions which are not quadratic. One method tested, a normalized version of the Fletcher-Reeves conjugate gradient algorithm [2], but requiring storage of one less number, was found to converge faster than the standard Fletcher-Reeves method. Although theoretically impossible, this is apparently due to compatibility of the normalized version with the Davidon search procedure [3,1] employed to obtain a minimum (approximately) in a given search direction. The normalized form would appear to be preferable to the original form of the Fletcher-Reeves algorithm.

II. ALGORITHMS: DESCRIPTIONS AND THEORY

Let $f: R^n \to R$ be a continuously differentiable function. We are concerned here with algorithms designed to locate stationary points of f, i.e., points x^* such that $g(x^*) = 0$, where $g: R^n \to R^n$ is the gradient of f. Under additional assumptions, such as f convex, or $(\partial^2/\partial x^2)f(x^*)$ positive definite, x^* is a global minimum or a local minimum respectively. The algorithms all take the following standard form.

Basic iterative form Given x_0 arbitrary, compute the sequence $x_0, x_1, \ldots$ by the steps:

(i) If $g(x_i) = g_i \neq 0$, choose a feasible direction p_i such that $\langle g_i, p_i \rangle < 0$.

(ii) Compute x_{i+1} such that

$$f(x_{i+1}) = f(x_{i+1} + \lambda_i p_i) = \min_{\lambda \geq 0} f(x_i + \lambda p_i)$$

It follows that $f(x_{i+1}) < f(x_i)$ and

(1) $$\langle p_i, g_{i+1} \rangle = 0 \quad \text{for } i = 0, 1, 2, \ldots .$$

The various algorithms differ only in the feasible search directions chosen.

Algorithm $A_1(\beta)$ Let $\beta \in (0,1]$, let $g_0 \neq 0$, and for $i \geq 1$, let $s_i = g_i - g_{i-1}$, $p_i' = -g_{i-1} + \dfrac{\langle s_i, g_{i-1} \rangle}{||s_i||^2} s_i$. Then, set $p_0 = -g_0$, and for $i = 1, 2, \ldots$

(2) $$p_i = \begin{cases} p_i' & \text{if } ||p_i'||^2 \geq \beta ||g_i||^2 \\ -g_i & \text{otherwise.} \end{cases}$$

Proposition 1 The algorithm $A_1(\beta)$, when specialized to the quadratic function, $x \to \langle x, Qx \rangle$, Q symmetric positive-definite, satisfies

(3) $$\langle p_{i+1}', Qp_i \rangle = 0 \quad \text{for } i = 0, 1, 2, \ldots$$

Proof: Trivial; $\langle s_{i+1}, p_{i+1}' \rangle = 0$, and $s_{i+1} = \lambda_i Q p_i$, $\lambda_i > 0$.

Proposition (1) indicates that if $p_{i+1} = p_{i+1}'$, then p_{i+1} and p_i are Q-conjugate. However the relation $\langle p_i, Qp_j \rangle = 0$, $i < j$, is not satisfied even if $p_k = p_k'$ for $k = i+1, \ldots, j$. The method is not a conjugate gradient method.

Proposition 2 For all i, the algorithm $A_1(\beta)$ satisfies

(4) $$-||g_i||^2 \leq \langle p_i, g_i \rangle = -||p_i||^2 \leq -\beta ||g_i||^2.$$

Proof: Relation (4) is clearly true if $p_i = -g_i$.

Suppose $p_i = p_i'$. Noting that $\langle p_i, s_i \rangle = 0$, yields

$$\langle p_i, g_i \rangle = \langle p_i, g_{i-1} \rangle = - ||g_{i-1}||^2 + \frac{\langle s_i, g_{i-1} \rangle^2}{||s_i||^2}.$$

But, by (1), $||p_i||^2 = ||g_{i-1}||^2 - \frac{\langle s_i, g_{i-1} \rangle^2}{||s_i||^2}$, hence $\langle p_i, g_i \rangle = - ||p_i||^2$, implying $||p_i||^2 \leq ||g_i||^2$, and (4) follows.

Proposition (2), along with similar relations for the following algorithms, will later be invoked to establish convergence to stationary points.

Algorithm $A_2(\beta)$ (Modified Fletcher-Reeves)

Let $\beta \in [0,1]$, let $g_0 \neq 0$, and for $i \geq 1$, let

$$p_i' = - g_i + \frac{||g_i||^2}{||p_{i-1}||^2} p_{i-1}$$

Then, set $p_0 = - g_0$, and for $i = 1, 2, \ldots$, set

$$(5) \qquad p_i = \begin{cases} \beta_i p_i' \text{ if } \beta ||p_i'|| \leqq ||g_i||, \ \beta_i \text{ arbitrary in } [\beta,1] \\ - g_i \quad \text{otherwise.} \end{cases}$$

Proposition 3 The algorithm $A_2(\beta)$, when specialized to the quadratic function, $x \to \langle x, Qx \rangle$, Q symmetric positive definite, satisfies

$$(6) \qquad \langle p_{i+1}, Q p_i \rangle = 0 \text{ if } p_i = - g_i.$$

Proposition (4) indicates that algorithm $A_2(\beta)$ takes a Q-conjugate step following a steepest descent step. It is not a conjugate gradient method.

Proposition 4 For all i, the algorithm $A_2(\beta)$ satisfies

$$(7) \qquad \begin{cases} \langle p_i, g_i \rangle \leq -\beta \, ||g_i||^2 \\ \\ \beta^2 ||p_i||^2 \leq ||g_i||^2 \end{cases}$$

Proof: If $p_i = -g_i$, (7) is satisfied. Thus, suppose $p_i = \beta_i p_i'$. Then,

$$\langle p_i, g_i \rangle = -\beta_i ||g_i||^2 \leq -\beta \, ||g_i||^2 \quad \text{by (1), (5).}$$

$$\beta^2 ||p_i||^2 = \beta^2 \beta_i^2 ||p_i'||^2 \leq \beta^2 ||p_i'||^2 \leq ||g_i||^2, \text{ by (5).}$$

Algorithm $A_3(\beta)$ (Normalized Fletcher-Reeves)

Let $\beta \in [0,1]$, let $g_0 \neq 0$, and for $i \geq 1$, let

$$\beta_i = \frac{||p_{i-1}||^2}{||p_{i-1}||^2 + ||g_i||^2} .$$

Then set $p_0 = -g_0$ and for $i = 1, 2, \ldots$ set

$$(8) \qquad p_i = \begin{cases} \beta_i \left(-g_i + \dfrac{||g_i||^2}{||p_{i-1}||^2} \, p_{i-1} \right) & \text{if } \beta_i \geq \beta \\ \\ -g_i & \text{otherwise.} \end{cases}$$

Proposition 5 The algorithm $A_3(0)$ is equivalent to the Fletcher-Reeves algorithm defined by $p_0' = -g_0$,

$$(9) \qquad p_i' = -g_i + \frac{||g_i||^2}{||g_{i-1}||^2} \, p_{i-1}' \qquad \text{for } i \geq 1$$

Proof: Equivalence of (9) and $A_3(0)$ follows if we can establish that p_i' and p_i are colinear, implying that (9) and (8) will lead to the same sequence $x_0, x_1, \ldots$. In particular we shall establish that

$$p_i = \frac{||g_i||^2}{||p_i'||^2} p_i' \quad \text{for } i = 0, 1, \ldots . \tag{10}$$

Clearly (10) holds for $i = 0$; suppose then that (10) holds for $i = 0, 1, \ldots, k$.

$$p_{k+1} = \beta_{k+1}(-g_{k+1} + \frac{||g_{k+1}||^2}{||p_k||^2} p_k) = \beta_k(-g_{k+1} + \frac{||g_{k+1}||^2}{||g_k||^2} p_k'), \text{ by (10).}$$

But, again using (10),

$$\beta_{k+1} = \frac{||p_k||^2}{||p_k||^2 + ||g_{k+1}||^2} = \frac{||g_{k+1}||^2}{||g_{k+1}||^2 + \frac{||g_{k+1}||^4}{||g_k||^4} ||p_k'||^2}$$

$$= \frac{||g_{k+1}||^2}{||p_{k+1}'||^2}, \text{ hence} \qquad p_{k+1} = \frac{||g_{k+1}||^2}{||p_{k+1}'||^2} p_{k+1}' .$$

The algorithm $A_3(0)$ is thus a conjugate gradient method, by equivalence with the Fletcher-Reeves algorithm. Proof that (9) defines a conjugate gradient method may be found in [4].

Proposition 6 For all i, the algorithm $A_3(\beta)$ satisfies

(11) $$-\|g_i\|^2 \le \langle p_i, g_i\rangle = -\|p_i\|^2 \le -\beta\|g_i\|^2$$

Proof: Relation (11) is trivial if $p_i = -g_i$, thus assume

$$p_i = \beta_i\left(-g_i + \frac{\|g_i\|^2}{\|p_{i-1}\|^2}\, p_{i-1}\right).$$ Then, $\langle p_i, g_i\rangle = -\beta_i\|g_i\|^2$

by (1)

But $\|p_i\|^2 = \beta_i^2\left(\|g_i\|^2 + \frac{\|g_i\|^4}{\|p_{i-1}\|^2}\right) = \beta_i^2\|g_i\|^2\left(\frac{\|p_{i-1}\|^2 + \|g_i\|^2}{\|p_{i-1}\|^2}\right)$

$= \beta_i\|g_i\|^2$, hence $\langle p_i, g_i\rangle = -\|p_i\|^2$, and (11) follows.

Convergence of the algorithms To prove convergence of the preceding algorithms, in an appropriate sense, we shall state a version of a general convergence theorem, essentially drawn from [5].

Let $f: R^n \to R$ be continuously differentiable, and let $x \to A(x)$ be a point-to-set mapping such that if x is not stationary, there exists $\delta = \delta(x) > 0$, and $\varepsilon = \varepsilon(x) > 0$, satisfying

(12) $$\sup_{y \in A(x')} f(y) \le f(x') - \delta \quad \text{for all } \|x'-x\| \le \varepsilon.$$

Given any x_0, set $i = 0$, and construct the sequence $x_0, x_1, \ldots$ according to the algorithm

(13)
- Step 1: If x_i stationary, stop. If x_i not stationary, choose $x_{i+1} \in A(x_i)$
- Step 2: Set $i = i+1$, and go to step 1.

Theorem 1 If the sequence x_0, x_1, ... is generated according to the algorithm (13), then either the sequence is finite and the last point is stationary, or the sequence is infinite, and any accumulation point is stationary.

Remark 1 If the sequence is infinite, accumulation points need not exist. Any assumption which ensures that the computation is carried out in a bounded set ensures that a stationary point will be found.

The proof of theorem (1) is trivial for the finite case, and a straightforward consequence of continuity for the infinite case.

We now prove convergence, in the sense of theorem (1), for the algorithms $A_1(\beta)$, $A_2(\beta)$, $A_3(\beta)$, with $\beta \in (0,1]$. We do this by constructing a set of search directions, compatible with convergence, and rich enough to include the directions specified by the preceding algorithms. Towards this end, for any $\gamma \in (0,1]$, and any $M > 0$, we define the set

$$(14)\quad P_{\gamma,M}(x) = \{p \mid ||p|| \leq M||g(x)||, \langle p, g(x)\rangle \leq -\gamma||g(x)||^2\}$$

and the point-to-set mapping $A_{\gamma,M}$ by

$$(15)\quad A_{\gamma,M}(x) = \{y \mid y = x + \lambda(x,p)p, \text{ where}$$

$$f(y) = \min_{\lambda \geq 0} f(x + \lambda p),\ p \in P_{\gamma,M}(x)\}$$

Lemma 1 Let x an arbitrary point, not stationary, let $\gamma \in (0,1]$, and $M > 0$. Then there exists $\delta = \delta(x) > 0$, and $\varepsilon = \varepsilon(x) > 0$ such that

$$(16)\quad \sup_{y \in A_{\gamma,M}(x')} f(y) \leq f(x') - \delta \quad \text{for } ||x' - x|| \leq \varepsilon$$

Theorem 2 If the sequence $x_0, x_1, \ldots$ is generated according to the algorithm of the form (13), using the mapping $A_{\gamma,M}$ (15), then either the sequence is finite and the last point is stationary, or the sequence is infinite, and any accumulation point is stationary.

Proof of lemma (1) Let $g(x) \neq 0$; then there exists $\epsilon > 0$, $\alpha > 0$ such that

$$(17)\quad \begin{cases} \alpha \leq ||g(x')|| \leq 2\alpha \quad \text{for all } ||x' - x|| \leq 2\epsilon \\ ||g(x') - g(x'')|| \leq \frac{\gamma\alpha}{4M} \quad \text{for all } ||x' - x|| + ||x'' - x|| \leq 3\epsilon \end{cases}$$

Let $\bar{\lambda} = \frac{\epsilon}{2M\alpha}$; then for all $||x - x'|| \leq \epsilon$, and for all $p \in P_{\gamma,M}(x')$, $||x' + \bar{\lambda}p - x|| \leq 2\epsilon$. Thus $f(x' + \bar{\lambda}p) = f(x') + \bar{\lambda}\langle p, g(x' + \eta\bar{\lambda}p)\rangle$, $0 \leq \eta \leq 1$, (mean-value theorem)

$$= f(x') + \bar{\lambda}\langle p, g(x')\rangle + \bar{\lambda}\langle p, g(x' + \eta\bar{\lambda}p) - g(x')\rangle$$

$$\leq f(x') - \bar{\lambda}\gamma||g(x')||^2 + \bar{\lambda}||p|| \frac{\gamma\alpha}{4M} \quad \text{by (14), (17)}$$

$$\leq f(x') - \bar{\lambda}\gamma\alpha^2 + \bar{\lambda}(M2\alpha)\frac{\gamma\alpha}{4M} \quad \text{by (14), (17)}$$

$$= f(x') - \delta, \quad \delta = \frac{1}{2}\bar{\lambda}\gamma\alpha^2.$$

Hence, for all $||x' - x|| \leq \epsilon$, and for all $p \in P_{\gamma,M}(x')$,

$f(x' + \lambda(x',p)p) \leq f(x' + \bar{\lambda}p) \leq f(x') - \delta$, i.e., (16) holds, and the proof is complete.

Convergence of the algorithms $A_1(\beta)$, $A_2(\beta)$ and $A_3(\beta)$, with $\beta \in (0,1]$, now follows directly from theorem (2), and propositions (2), (4), and (6). For $A_1(\beta)$, proposition (2) yields $p_i \in P_{\beta,1}(x_i)$; for $A_2(\beta)$, proposition

(4) yields $p_i \in P_{\beta,\frac{1}{\beta^2}}(x_i)$; for $A_3(\beta)$, proposition (6) yields $p_i \in P_{\beta,1}(x_i)$.

Finally, we note that these results include the steepest descent algorithm, $p_i = -g_i$, which is equivalent to $A_1(1)$, $A_2(1)$, or $A_3(1)$.

Remark 2 A set $P_\gamma(x)$ compatible with convergence in the same sense as $P_{\gamma,M}(x)$ is

$$(18)\quad P_\gamma(x) = \{p \mid \langle p, g(x)\rangle \leq -\gamma\, ||p||\, ||g(x)||\}$$

Note that $P_{\gamma,M}(x) \subset P_{\gamma/M}(x)$; $P_{\gamma/M}(x) \subset P_\gamma(x)$ if $M \leq 1$.

The parameter $\beta \in (0,1]$ serves as a design parameter which governs how orthogonal we allow p_i and g_i to become. If β is allowed to take the value 0, theorem (2) is no longer applicable. To prove convergence of the algorithm $A_3(0)$, which corresponds to the Fletcher-Reeves algorithm without sequence breaking, we shall prove a convergence theorem motivated by some known results for the Fletcher-Reeves method.

Theorem 3* Suppose that the sequence $x_0, x_1, \ldots$ corresponds to a sequence $p_0, p_1, \ldots$ of feasible search directions, and that $x_0, x_1, \ldots$ has a convergent subsequence $x_0', x_1', \ldots \quad x^*$. If f is a twice continuously differentiable function, and if there exists a sequence $\gamma_0, \gamma_1, \ldots$ of positive scalars such that

$$(19)\qquad \langle g_i', p_i'\rangle \leq -\gamma_i\, ||g_i'||\,||p_i'|| \qquad i = 0, 1, \ldots$$

*The form which this theorem ought to take came to the author's attention through unpublished results of G. Ribiere of I.B.M., Paris, France, and G. Zoutendijk of The University of Leiden, Netherlands, as communicated to the author by E. Polak, University of California, Berkeley.

(where the prime denotes the subsequence of interest)

$$(20) \qquad \sum_{i=0}^{\infty} \gamma_i^2 = \infty,$$

then $g(x^*) = 0$, i.e., x^* is a stationary point.

Proof: Assume that theorem (3) is false, i.e., $g(x^*) \neq 0$. Then there exists $\varepsilon > 0$, $\delta > 0$, $\gamma > 0$, such that

$$(21) \qquad ||g(x)|| \geq \delta \quad \text{for all } ||x - x^*|| \leq 2\varepsilon$$

$$(22) \qquad ||H(x)|| \leq \alpha \quad \text{for all } ||x - x^*|| \leq 2\varepsilon. \quad \left(H(x) = \frac{\partial^2 f}{\partial x^2}(x)\right)$$

Let $h_i = \frac{1}{||p_i|| \; ||s_i||} p_i$, and consider the expansion

$$f(x_i' + \lambda h_i') = f(x_i') + \lambda \langle g_i', h_i' \rangle + \frac{\lambda^2}{2} \langle h_i', H(x_i' + \eta\lambda h_i')h_i' \rangle, \; 0 \leq \eta \leq 1$$

$$\leq f(x_i') + \lambda \langle g_i', h_i' \rangle + \frac{\lambda^2}{2} ||h_i||^2 \; ||H(x_i' + \eta\lambda h_i')||$$

$$\leq f(x_i') + \lambda \langle g_i', h_i' \rangle + \frac{\lambda^2}{2} \frac{\alpha}{\delta^2} \text{ by (19), (20),}$$

providing $||x_i' - x^*|| \leq \varepsilon$ and $\lambda \leq \varepsilon\delta$ (which implies $||\eta\lambda h_i'|| = \frac{\eta\lambda}{||g_i'||}$ $\leq \frac{\lambda}{\delta} \leq \varepsilon$).

Note that $\gamma_i \in (0,1]$ by (19); thus if $0 < \ell \leq 1$ is chosen such that $\ell\frac{\delta}{\alpha} \leq \varepsilon$, then $\lambda_i = \ell\frac{\delta^2}{\alpha}\gamma_i$ satisfies $\lambda_i \leq \varepsilon\delta$ for all $i = 0, 1, \ldots$. Hence

$$f(x_{i+1}') \leq f(x_i' + \lambda_i h_i') \leq f(x_i') - \lambda_i\gamma_i + \frac{\lambda_i^2\alpha}{2\ell\alpha^2} \text{ since } 0 < \ell \leq 1,$$

$\langle g_i', h_i' \rangle \leq -\gamma_i$ by (19). Setting $\ell' = \frac{\ell\delta^2}{\alpha} > 0$, we find that $f(x'_{i+1}) \leq f(x_i') - \ell'\gamma_i^2$ for all $||x_i' - x^*|| \leq \varepsilon$, hence (20) implies $f(x_i') \to -\infty$, which is a contradiction ($f(x_i') \to f(x^*) > -\infty$). This completes the proof.

Remark 3 It is not known whether or not the assumption that f be twice-continously differentiable can be dispensed with in theorem (3). Those conjugate gradient methods which have been proven to converge for general functions (without sequence breaking) also require at least as strong an assumption, as we see in the next theorem. (See also, the Polak-Ribiere algorithm [6]).

Theorem 4 Suppose that the sequence $x_0, x_1, \ldots$ is constructed according to the algorithm $A_3(0)$, and that $f: R^n \to R$ is twice-continuously differentiable. If $x_0, x_1, \ldots$ converges to a point x^*, then $g(x^*) = 0$.

Proof: Suppose $g(x^*) \neq 0$; then there exists $\alpha > 0$ and i^* such that $\alpha \leq ||g(x_i)|| \leq 2\alpha$ for all $i \geq i^*$. Without loss of generality, we shall assume $i^* = 0$. Now, by proposition (5), $A_3(0)$ is equivalent to the Fletcher-Reeves algorithm (9). Using (9), (1), we obtain

$$\langle p_i', g_i \rangle = -||g_i||^2 \quad \text{for all } i \tag{23}$$

$$||p_i'||^2 = ||g_i||^2 + \frac{||g_i||^2}{||g_{i-1}||^4} ||p'_{i-1}||^2 \tag{24}$$

$$= ||g_i||^2 \left(\frac{||g_i||^2}{||g_i||^2} + \ldots\ldots + \frac{||g_i||^2}{||g_0||^4} ||g_0||^2 \right)$$

$$\leq ||g_i||^2 \, 4(i+1) \qquad \text{for all } i.$$

Combining (23) and (24), we obtain

$$(25) \quad \langle p_i', g_i \rangle \leq -\gamma_i ||g_i|| \, ||p_i'||, \text{ where } \gamma_i^2 = \frac{1}{4(i+1)}$$

But $\sum_{i=0}^{\infty} \gamma_i^2 = \infty$, and hence by theorem (3), $g(x^*) = 0$ which contradicts the assumption that $g(x^*) \neq 0$. Thus, theorem (4) is true.

Remark 4 If in theorem (4), it is only assumed that the sequence x_0, x_1, ... has a subsequence converging to x^*, the proof breaks down. A convergence result for Fletcher-Reeves given in the literature [7], which is stated for a subsequence, would appear to be in error because the effect of an assumed normalization was not adequately accounted for.

The following result is intended to show that the conditions (19) and (20) of theorem (3) cannot be relaxed much. In particular, since (19) requires $0 \leq \gamma_i \leq 1$, $\sum_{i=0}^{\infty} \gamma_i^2 = \infty$ implies $\sum_{i=0}^{\infty} \gamma_i = \infty$. This latter condition is sometimes believed to be sufficient for convergence.

Proposition 7 The conditions $-\langle p_i, g_i \rangle \leq \gamma_i \, ||p_i|| \, ||g_i||$, with $\sum_{i=0}^{\infty} \gamma_i = \infty$, do not imply that the corresponding sequence x_0, x_1, ... has accumulation points which are stationary, even if the sequence is bounded, if $\sum_{i=0}^{\infty} \gamma_i = \infty$ holds for a convergent subsequence, and if f is twice continuously differentiable and strictly convex.

Proof: Let $f: R^n \to R$ be defined by $f(x) = ||x||^2$, let $x_1 \neq 0$, and let the p_i satisfy

(26) $-\langle p_i, g_i \rangle = \frac{r}{i} ||p_i|| \, ||g_i||$ for $i = 1, 2, \ldots$

where $r^2 = \frac{1}{8}\left(\sum_{j=1}^{\infty} \frac{1}{j^2}\right)^{-1}$. Hence $\gamma_i = \frac{r}{i}$ for $i = 1, 2, \ldots$; $\sum_{i=1}^{\infty} \gamma_i = \infty$.

But $f(x_{i+1}) = ||x_i||^2 + 2\lambda_i \langle p_i, g_i \rangle + \lambda_i^2 ||p_i||^2$, and $\lambda_i = \frac{\langle p_i, g_i \rangle}{||p_i||^2}$ yields

$$f(x_{i+1}) = ||x_i||^2 - \frac{\langle p_i, g_i \rangle^2}{||p_i||^2} = f(x_i) - \frac{r^2}{i^2} ||g_i||^2 .$$

Thus, $\lim_{i \to \infty} f(x_i) = f(x_1) - \sum_{j=1}^{\infty} \frac{r^2}{j^2} ||g_j||^2$

$$\geq f(x_1) - \left(\sum_{j=1}^{\infty} \frac{r^2}{j^2}\right) ||g_1||^2 = \frac{1}{2} f(x_1) > 0, \text{ since}$$

$||g_{i+1}||^2 = 4 \, ||x_{i+1}||^2 \leq 4 \, ||x_i||^2 = ||g_i||^2$ for all i, and $\sum_{j=1}^{\infty} \frac{r^2}{j^2} = \frac{1}{8}$.

Thus, no subsequence of $x_1, x_2, \ldots$ can converge to a stationary point. Finally, we note that, while $x_1, x_2, \ldots$ need not converge, appeal to the situation for R^2 indicates that the p_i may be chosen so that $x_1, x_2, \ldots$ zig-zags to a single accumulation point (In particular, in R^2, specify additionally that $\langle p_i, p_{i+1} \rangle < 0$ for all i. The convergence may be proven rigorously).

Since the function $||x||^2$ has about as many nice properties as one may ask for, the conditions (19) and (20) cannot be significantly relaxed.

If β is allowed to be zero in $A_2(\beta)$, the algorithm is not meaningful without a restriction on the β_i. We obtain the following results for this case.

Theorem 5 Suppose that $x_0, x_1, \ldots$ is constructed according to the algorithm $A_2(0)$, with the additional restriction that $\inf_i \beta_i = \beta^* > 0$.

If the sequence $x_0, x_1, \ldots$ converges to a point x^*, then $g(x^*) = 0$.

Proof: Suppose that $g(x^*) \neq 0$; then there exists $\alpha > 0$, $i^* \geq 1$, such that $||g(x_i)|| \geq \alpha$ for all $i \geq i^*$.

By (5), (1), we obtain

$$\langle p_i, g_i \rangle = -\beta_i ||g_i||^2, \quad ||p_i||^2 = \beta_i^2 ||g_i||^2 \left(1 + \frac{||g_i||^2}{||p_{i-1}||^2}\right)$$

for all $i \geq i^* \geq 1$.

Now $\sup_{i \geq i^*} ||p_i||^2 = \infty$ iff $\inf_{i \geq i^*} ||p_i||^2 = 0$, (recall that $\beta_i \leq 1$). But $||p_i||^2 \geq (\beta^*)^2 \alpha^2 > 0$ for all $i \geq i + 1$, hence $||p_i||$ is bounded, say by $\alpha^2 M$. Then

$$\langle p_i, g_i \rangle \leq -\beta^* ||g_i||^2 \qquad i \geq i^* \tag{27}$$

$$||p_i||^2 \leq \alpha^2 M \leq M ||g_i||^2 \qquad i \geq i^*, \tag{28}$$

i.e., $p_i \in P_{\beta^*, M}(x_i)$, $x_{i+1} \in A_{\beta^*, M}(x_i)$, for all $i \geq i^*$.

By theorem (2), $x_i \to x^*$ yields $g(x^*) = 0$, which is a contradiction. Theorem (5) thus holds.

Theorem 6 Suppose that $x_0, x_1, \ldots$ is constructed according to the algorithm $A_2(0)$, with the additional restrictions that $\sum_{i=0}^{\infty} \beta_i^2 = \infty$ and β_i/β_{i-1} bounded. If f is twice-continuously differentiable, and if the sequence $x_0, x_1, \ldots$ converges to a point x^*, then $g(x^*) = 0$.

Proof: Assume $g(x^*) \neq 0$; then there exists $\alpha > 0$, $i^* \geq 1$ such that $\alpha \leq ||g(x_i)|| \leq 2\alpha$ for all $i \geq i^*$. Now $||p_i||^2 \geq (\beta_i)^2 \alpha^2$ for all $i \geq i^*$, hence $||p_i||^2 \leq \beta_i^2 4\alpha^2 \left(1 + \frac{4\alpha^2}{(\beta_{i-1})^2\alpha^2}\right) \leq 4\alpha^2 + 16\alpha^2 \left(\frac{\beta_i}{\beta_{i-1}}\right)^2 \leq M\alpha^2$

for some M. i.e., $||p_i||^2$ is bounded, and hence

$$\langle p_i, g_i \rangle = -\beta_i ||g_i||^2 \qquad i \geq i^* \tag{29}$$

$$||p_i||^2 \leq M ||g_i||^2 \qquad i \geq i^* \tag{30}$$

Thus $\langle p_i, g_i \rangle \leq -\frac{\beta_i}{\sqrt{M}} ||p_i|| \, ||g_i||$ for all $i \geq i^*$, and by theorem (3), we conclude that $g(x^*) = 0$; a contradiction. This completes the proof.

Remark 5 The assumption that β_i/β_{i-1} is bounded holds for any nonincreasing sequence of β_i. It avoids such strategies as choosing $\beta_i = 1$, i odd, $\beta_i = \frac{1}{i}$, i even, etc.

III. NUMERICAL RESULTS

Versions of the preceding algorithms were experimentally compared to steepest descent and standard conjugate gradient algorithms. For uniformity of test procedure, all comparisons were made with each method reverting to steepest descent after every n + 1 steps. For $A_2(\beta)$, β_i was chosen to be 1 for all i; the algorithm thus becomes

$$p_i = \begin{cases} -g_i + \dfrac{||g_i||^2}{||p_{i-1}||^2} p_{i-1} & i \neq k(n+1),\ k = 0, 1, \ldots \\ -g_i & \text{otherwise} \end{cases} \tag{31}$$

Note that the computational results for the preceding algorithms do not now depend on β; these results are only intended to indicate the efficacy of using search directions of the various forms.

The iterations were stopped when $||g(x)||$ was sufficiently small, $||g(x)||^2 \leq \varepsilon_{tol}$ being the particular criterion. The linear search method programmed was that originally introduced by Davidon, [3] as detailed in Fletcher and Powell [1]. Briefly, the search method consists of bracketing a (local) minimum in an interval, and using (possibly repeated) cubic interpolation to approximate the minimum.

The functions to which the algorithms were applied are

$$f_1: R^{10} \to R \text{ defined by } x \to \langle x - x^*, Q(x - x^*)\rangle \tag{32}$$

where Q is a symmetric positive definite matrix†

$$f_2(x) = \exp(f_1(x)) + (x_6^2-1)x_1 + 1 \tag{33}$$

$$f_3(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \qquad \text{(Rosenbrock, [8])}. \tag{34}$$

All the algorithms tested were coded - by the same person - in Fortran IV for execution on the CDC 6400 computer of the University of California Computing Center. Tables 1 - 3 summarize the results obtained.

† $f_1(x) = 0.1(x_1-1)^2+(x_2+0.2)^2+(x_3-0.3)^2+5(x_4+0.4)^2$
$+ 0.2(x_5-0.5)^2+.4(x_7+x_8)^2+.3(x_8-x_9)^2+1.5(x_9+x_{10})^2$
$+ (x_7+x_{10})^2+(\sum_{i=1}^{10} x_i-1.2)^2+x_6^2+.5(x_5+x_2-x_3)^2 + 4(2x_5-x_1+x_8)^2$

Remark 6 Algorithm A_1 tended to display the same characteristics as steepest descent for functions f_1 and f_2, although the convergence was more rapid. However, for function f_3, A_1 converged in fewer steps (and with higher accuracy) than all other methods tested. Although it is not clear why this occurred, the result for f_3 may be misleading because in R^2 algorithm A_1 behaves like a conjugate gradient method.

Remark 7 Algorithm A_2 performed surprisingly well. Reasonable convergence obtained for the quadratic function f_1. For f_2 and f_3 the algorithm performed as well or better than the similarly structured Fletcher-Reeves algorithm. Thus, the fact that A_2 is not a conjugate gradient method appears to be of no particular consequence when dealing with nonquadratic functions, providing the search directions are of comparable form to those selected by the Fletcher-Reeves algorithm. Methods to determine the best way to use the freedom to choose the β_i would no doubt lead to better versions of the algorithm A_2.

Remark 8 The algorithm A_3, the normalized Fletcher-Reeves algorithm, converged faster than the Fletcher-Reeves algorithm. This is theoretically impossible by Proposition (5). In practice it is due to the use of an approximate procedure to locate the minimum in a given search direction. Apparently, the normalized version is more compatible with the Davidon search technique, which utilizes $||p_i||$ in obtaining an initial estimate of the step length λ_i. Very likely, the relation $||p_i|| \leq ||g_i||$ for the normalized version (Proposition (6)), as compared to $||p_i'|| > ||g_i||$, $i \neq 0$, for (9), is responsible for the improvement. For this reason, and because the parameter β_i (8) (which measures the orthogonality of p_i

and g_i) should be monitored during any computation; the normalized version of the Fletcher-Reeves algorithm appears to be preferable to the original version. Note that the normalized version does not require $||g_{i-1}||^2$ for the computation of p_i (8).

CONCLUSIONS

The results of this paper confirm that algorithms which retain some of the features of conjugate gradient methods, and some of the features of steepest descent, can be quite effective. The global behavior of these methods is extremely difficult to predict. In fact, this behavior depends on a complex interplay between the choice of search directions and the choice of search procedure used. Perhaps when we can better characterize this interplay - or learn how to proceed without this convenient decomposition - the flexibility and power of computers can be more effectively utilized.

Even when comparing relatively simple classes of algorithms, there are questions yet to be answered. How should we best utilize information on orthogonality between the gradient and the search direction? How many function-gradient evaluations per iteration should we expect (or tolerate) for a given choice of search direction-search procedure? The answer to questions such as these would appear to be essential to the comparison of algorithms within a framework that includes not only iterative quality, but the costs of computer implementation.

Algorithm	Steps required	Function-gradient evaluations	Computation time (sec.)†	Final value
A_1	48	98	.448	2.0×10^{-6}
A_2	38	78	----	3.9×10^{-5}
A_3	11	24	.108	8.5×10^{-9}
Steepest Descent	$\geq$99	$\geq$200	$\geq$.82	2.1×10^{-3}
Fletcher-Reeves	11	24	----	1.4×10^{-10}
Fletcher-Powell	10	22	----	1.4×10^{-28}
Modified Fletcher-Powell*	10	23	----	1.3×10^{-27}

TABLE 1 Function f_1: starting point $x_0 = (.5, .5, \ldots, .5)$, $f_1(x_0) = 260$, stopping criterion $\varepsilon_{tol} = 10^{-5}$

---

†The computation times include the times for two redundant function-graident evaluations per iteration, which are not included in the count of function-gradient evaluations.

*The modified Fletcher-Powell method is described in [4]; it uses $H_{i+1} = H_i - \frac{H_i s_{i+1} s_{i+1}^T H_i}{\langle s_{i+1}, H_i s_{i+1} \rangle}$, where $s_{i+1} = g_{i+1} - g_i$. The results for this algorithm, as well as for steepest descent, Fletcher-Reeves and Fletcher-Powell are drawn from a report by Nuytten [9].

Algorithm	Steps required	Function-gradient evaluations	Computation time (sec.)	Final value
A_1	95	420	1.4	.88
A_2	55	216	.74	.88
A_3	54	208	.74	.88
Steepest Descent	309	1672	4.86	.88
Fletcher-Reeves	59	221	.78	.88
Fletcher-Powell	35	109	2.63	.88
Modified Fletcher-Powell	65	206	3.11	.88

TABLE 2 Function f_2: starting point $x_0 = (-.5, -.5, \ldots, -.5)$, $f_2(x_0) = 2.4 \times 10^{20}$, stopping criterion $\varepsilon_{tol} = 10^{-5}$.

Algorithm	Steps required	Function-gradient evaluations	Computation time (sec.)	Final value
A_1	23	53	0.16	1.8×10^{-12}
A_2	31	71	0.21	2.4×10^{-10}
A_3	26	60	0.18	2.9×10^{-10}
Steepest Descent	134	290	0.81	---------
Fletcher-Reeves	31	71	0.21	2.1×10^{-10}
Fletcher-Powell	32	72	0.25	1.2×10^{-10}
Modified Fletcher-Powell	30	63	0.22	3.1×10^{-8}

TABLE 3 Function f_3: starting point $x_0 = (-1.2, 1.0)$, $f_3(x_0) = 240$, stopping criterion $\varepsilon_{tol} = 10^{-6}$

ACKNOWLEDGEMENT

The author would like to thank Mr. Christian Nuytten for programming the algorithms and obtaining the computational results presented in this paper.

REFERENCES

1. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," The Computer Journal, Vol. 6, 1963, p. 163.

2. R. Fletcher and C. m. Reeves, "Function Minimization by Conjugate Gradients," The Computer Journal, Vol. 7, No. 2, 1964, pp. 149-154.

3. W. C. Davidon, "Variable Metric Methods for Minimization," A.E.C. Research and Development Report ANL 5990 (Rev)., 1959.

4. G. E. Myers, "Properties of the Conjugate Gradient and Davidon Methods," Journal of Optimization Theory and Applications," Vol. 2, No. 4, (July 1968) pp. 209-220.

5. E. Polak, "On the Convergence of Optimization Algorithms," Revue Francais, d'Informatique et de Recherche Operationelle, Serie Rough, No. 16, 1969, pp. 17-34.

6. E. Polak and G. Ribiere, "Note sur la Convergence de Methods de Directions Conjugees," Revue Francais d'Informatique et de Recherche Operationelle, No. 16, 1969.

7. L. S. Lasdon, S. K. Mitter, and A. D. Waren, "The Conjugate Gradient Method for Optimal Control Problems," IEEE Trans., Vol. AC-12, No. 2, April 1967, pp. 132-138.

8. D. J. Wilde, Optimum Seeking Methods, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

9. C. Nuytten, Unconstained Minimization Problems by Conjugate Gradient Methods, Masters Plan II Report, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, June, 1969.