JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 33, No. 2, FEBRUARY 1981

Stepsize Analysis for Descent Methods

A. I. COHEN[1]

Communicated by D. Q. Mayne

Abstract. The convergence rates of descent methods with different stepsize rules are compared. Among the stepsize rules considered are: constant stepsize, exact minimization along a line, Goldstein–Armijo rules, and stepsize equal to that which yields the minimum of certain interpolatory polynomials. One of the major results shown is that the rate of convergence of descent methods with the Goldstein–Armijo stepsize rules can be made as close as desired to the rate of convergence of methods that require exact minimization along a line. Also, a descent algorithm that combines a Goldstein–Armijo stepsize rule with a secant-type step is presented. It is shown that this algorithm has a convergence rate equal to the convergence of descent methods that require exact minimization along a line and that, eventually (i.e., near the minimum), it does not require a search to determine an acceptable stepsize.

Key Words. Descent methods, rates of convergence, stepsize rules.

1. Introduction

Descent methods for minimizing real-valued functions f on R^n are methods of the form

$$x_{i+1} = x_i + \lambda_i h_i, \tag{1}$$

where h_i is a search direction such that, for all i,

$$-\langle h_i/|h_i|, f'(x_i)/|f'(x_i)|\rangle \geq \rho > 0 \tag{2}$$

and $\lambda_i > 0$ is a stepsize. A special type of descent method is the *deflected gradient method*, defined by

$$x_{i+1} = x_i - \lambda_i \Gamma_i f'(x_i), \tag{3}$$

where Γ_i is uniformly positive definite, or equivalently the eigenvalues of Γ_i lie in the interval $[g, G]$ for all i, where $g > 0$ and $G < \infty$.

[1] Senior Engineer, Systems Control, Palo Alto, California.

0022-3239/81/0200-0187$03.00/0

The purpose of this paper is to compare upper bounds on the linear convergence rates that result when different rules are used to determine the stepsize in (1) and (3). Linear convergence rate is defined as follows.

Definition 1.1. A sequence x_i *converges linearly* to a point x^* with rate at least $q \in (0, 1)$ if $\lim_{i\to\infty} x_i = x^*$ and, for some constant $c > 0$ and for all n,

$$\overline{\lim_{i\to\infty}}[|x_{i+n} - x^*|/|x_i - x^*|] \leq cq^n.$$

The following stepsize rules (in some cases, two of these rules will be combined in one algorithm) will be investigated.

(a) Set λ_i equal to a predetermined constant.

(b) Set λ_i equal to the value which yields the minimum of some polynomial (usually quadratic or cubic) that interpolates

$$\phi_i(\lambda) = f(x_i + \lambda h_i) - f(x_i)$$

and/or its derivative at one or more values of λ.

(c) Set λ_i to be the smallest value of λ such that

$$f(x_i + \lambda_i h_i) \leq f(x_i + \lambda h_i), \qquad \text{for all positive } \lambda.$$

This stepsize rule will be referred to as the *minimization rule* or *Rule (M)*.

(d) Require that λ_i satisfy a condition that is a function of the ratio

$$(f(x_i + \lambda_i h_i) - f(x_i))/\langle f'(x_i), h_i\rangle\lambda_i.$$

Stepsize rules in the class described in (d) have received much attention in the recent literature, and they will be considered in detail in this paper. One type of requirement is that λ_i satisfy

$$\eta_1 \leq [f(x_i + \lambda_i h_i) - f(x_i)]/\lambda_i\langle f'(x_i), h_i\rangle \leq \eta_2, \tag{4}$$

where $0 < \eta_1 < \eta_2 < 1$. This condition is equivalent to requiring that the decrease in f lie between the lines

$$l_1(\lambda) = \eta_1\langle f'(x_i), h_i\rangle\lambda$$

and

$$l_2(\lambda) = \eta_2\langle f'(x_i), h_i\rangle\lambda$$

(see Fig. 1).

It can be shown (Ref. 1) that conditions (2) and (4) are sufficient to ensure convergence of a subsequence of (1) to a critical point of f, if f has continuous first partial derivatives and the level set of f at x_0 is compact.

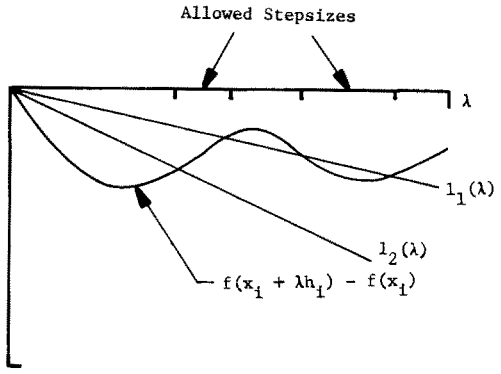

Fig. 1. Illustration of stepsize rule (4).

A stepsize rule in the form of (4), due to Goldstein (Ref. 2), is the following.

Rule (G). Pick λ_i to satisfy (4), with $\eta_1 = \alpha$, $\eta_2 = 1 - \alpha$, where $\alpha \in (0, \frac{1}{2})$.

Another stepsize rule, suggested by Armijo (Ref. 3) is the following.

Rule (A). Let $\lambda_i = \beta^j d$, where $\beta \in (0, 1)$, $d > 0$, and j is the first nonnegative integer such that, for $\alpha \in (0, 1)$,

$$[f(x_i + \lambda_i h_i) - f(x_i)]/\lambda_i \langle f'(x_i), h_i \rangle \geq \alpha, \tag{5}$$

or equivalently such that the decrease in f lies below the line

$$l(\lambda) = \alpha \langle f'(x_i), h_i \rangle \lambda.$$

Note that Rule (A) implies either

$$\lambda_i = d, \qquad \text{if} \quad j = 0,$$

or

$$[f(x_i + (\lambda_i/\beta) h_i) - f(x_i)]/(\lambda_i/\beta) \langle f'(x_i), h_i \rangle < \alpha. \tag{6}$$

A variant of Rule (A) is the following.

Rule (A'). Let $d > 0$ and $\alpha \in (0, 1)$. If

$$[f(x_i + d h_i) - f(x_i)]/d \langle f'(x_i), h_i \rangle \leq \alpha, \tag{7}$$

pick λ_i according to Rule (A). If not, let

$$\lambda_i = d/\beta^j,$$

where $\beta \in (0, 1)$ and j is the first nonnegative integer such that (6) is satisfied. Therefore, for this stepsize rule, λ_i satisfies both (5) and (6).[2]

Known Results. The following results in the rate of convergence of descent methods are known. If f'' exists and has eigenvalues between $m > 0$ and $M < \infty$, then these results hold.

(i) The descent method (1) converges linearly if λ_i is chosen according to either Rule (M) or Rule (G); see Ref. 5, pp. 242–246.

(ii) Descent methods of the form (3), with λ_i chosen according to Rule (M), converge linearly; see Ref. 5, pp. 248–249.

(iii) The descent method (1) with $h_i = -f(x_i)$ and $\lambda_i \in [\delta, 2/M - \delta]$, where $\delta \in (0, 1/M)$, converges linearly. The best rate estimate on this method occurs when

$$\lambda_i = 2/(M + m), \qquad \text{for all } i.$$

The resultant rate is

$$q = (M - m)/(M + m);$$

see Refs. 6–8.

(iv) If f is quadratic, then (3) with Rule (M) converges with rate at least

$$q = (R - r)/(R + r),$$

where $R < \infty$ and $r > 0$ are, respectively, bounds on the maximum and minimum eigenvalues of $\Gamma_i f''(x)$; see Refs. 8–10.

Remark 1.1. The ratio $(R - r)/(R + r)$ is sometimes referred to as the Kantorovitch ratio. Given the assumptions, this is the tightest bound on convergence that is known. It has been shown by Akaike (Ref. 11) that, for Γ_i equal the identity matrix, barring certain degenerate starting points, the convergence rate is exactly $(M - m)/(M + m)$ (note that, since $\Gamma_i = I$, $M = R$, $m = r$).

Lemma 1.1. *Kantorovitch Lemma.* The proof of (iv) uses the Kantorovitch lemma which states (Ref. 10) that, for a positive-definite matrix A,

$$\langle s, s\rangle^2/\langle s, As\rangle\langle s, A^{-1}s\rangle \geq 4a\mathscr{A}/(a + \mathscr{A})^2, \tag{8}$$

where a and $\mathscr{A}$ are respectively the smallest and largest eigenvalues of A. Bauer and Householder (Ref. 12) extended the Kantorovitch result to show

[2] Wolfe (Ref. 4) gives some other stepsize rules that differ from (4). However, as he shows, they are closely related.

that, for all nonzero vectors u and v and positive-definite matrices A where

$$|\langle u, v\rangle| \geq |u| \cdot |v|\rho, \tag{9}$$

for $\rho \in [0, 1]$, then

$$\frac{\langle u, u\rangle\langle v, v\rangle}{\langle u, Au\rangle\langle v, A^{-1}v\rangle} \geq \frac{4x}{[(x+1)+(x-1)\sqrt{(1-\rho^2)}]^2}, \tag{10}$$

where $x = \mathscr{A}/a$, the ratio of the largest and smallest eigenvalues of A. Clearly, if

$$s = u = v \qquad \text{and} \qquad \rho = 1,$$

then (10) and (8) agree. Inequality (10), however, will allow us to obtain tighter bounds on descent algorithms of the form (1) than appear, for example, in Ref. 4. It should be noted, however, that these tighter bounds require an additional assumption (i.e., f must be continuously three times differentiable).

Major Results. The major results in this paper are given in five theorems. Theorem 3.1 reviews results (i) and (ii) with some extensions, and Theorem 3.2 reviews result (iii). Theorem 3.3 extends (iv) to convex, but not necessarily quadratic, f. The theorem also gives a tighter bound on convergence rate for descent methods of the form (1) using the extension of the Kantorovitch lemma. The theorem also shows that the convergence result still holds if Rule (M) is replaced with a Newton step, i.e.,

$$\lambda_i = \phi_i'(0)/\phi_i''(0),$$

where

$$\phi_i(\lambda) = f(x_i + \lambda h_i) - f(x_i).$$

Theorem 3.4 shows that (1) or (3) with Rules (G) and (A′) has a convergence rate that can be made as close to the Kantorovitch ratio (or extended Kantorovitch ratio) as desired by adjusting the stepsize parameters. Finally, an algorithm is defined which combines Rule (A) with a secant-type step. Theorem 3.5 shows that this algorithm has a rate of convergence equal to the Kantorovitch ratio, and eventually does not require a one-dimensional search at each step.

2. Assumptions on f

Throughout this paper, we shall make some or all of the following assumptions on $f: R^n \to R$.

Assumption (A1). f has continuous first and second partial derivatives.

Assumption (A2). f has continuous third partial derivatives.

Assumption (A3). For all $x, y \in R^n$, there exists an $m > 0$, such that

$$m|y|^2 \leq \langle y, f''(x)y \rangle.$$

Assumption (A4). For all $x, y \in R^n$, there exists an $M < \infty$, such that

$$\langle y, f''(x)y \rangle \leq M|y|^2.$$

These assumptions are made for all $x \in R^n$; however, since rate-of-convergence results are local results, these assumptions can be relaxed to be needed only in some neighborhood of the minimum.

3. Rate of Convergence Results

Descent Methods. We shall first show that the rate of convergence of descent methods of the form (1), with Rules (M), (G), (A'), are linear [the linear convergence of the descent method with Rule (A) needs some extra requirements]. Later in this paper, tighter bounds on the rate of convergence of these methods will be given. These bounds will, however, require an extra condition on f (f three times continuously differentiable, as opposed to twice). The convergence result follows easily from the following lemma.

Lemma 3.1. (*Ref. 1, p. 477, and Ref. 4, p. 245*). Suppose that f satisfies Assumptions (A1), (A3), (A4), $\{x_i\}$ converges to x^*, where $f'(x^*) = 0$, and that

$$f(x_i) - f(x_{i+1}) \geq \lambda |f'(x_i)|^2. \tag{11}$$

Then, $\{x_i\}$ converges linearly with rate at least

$$q = \sqrt{(1 - 2\lambda m^2/M)}.$$

Using this lemma we shall calculate convergence rates for the descent method with various stepsize rules.

Theorem 3.1. Suppose that f satisfies Assumptions (A1), (A3), (A4). Then, the descent method (1) converges linearly with rate[3]

(i) $q = \sqrt{[1 - (\rho m/M)^2]}$, if Rule (M) is used,

[3] For (i) and (ii), see Ref. 4.

(ii) $q = \sqrt{[1-(2\alpha\rho m/M)^2]}$, if Rule (G) is used,

(iii) $q = \sqrt{[1-\alpha(1-\alpha)\beta(2\rho m/M)^2]}$, if Rule (A′) is used.

Also, the descent method (3) with Rule (A) converges linearly with rate

(iv) $q = \sqrt{(1-2\gamma m^2 M)}$,

where

$$\gamma = \min(\alpha d g, 2\alpha(1-\alpha)\beta g^2/G^2 M).$$

Proof. In general,

$$f(x_i) - f(x_{i+1}) = -\lambda_i \langle f'(x_i), h_i\rangle - \tfrac{1}{2}\lambda_i^2 \langle h_i, H_i h_i\rangle, \tag{12}$$

where, for some $t \in (0, 1)$,

$$H_i = f''(x_i + t\lambda_i h_i). \tag{13}$$

Using (2), (12), and Assumption (A4), we have

$$f(x_i) - f(x_{i+1}) \geq \lambda_i \rho |f'(x_i)|\,|h_i| - (M\lambda_i^2/2)|h_i|^2. \tag{14}$$

(i) Using Rule (M), we know that the decrease in f is at least as large as the maximum of the right-hand side of (14) with respect to λ_i. Thus,

$$f(x_i) - f(x_{i+1}) \geq (\rho^2/2M)|f'(x_i)|^2. \tag{15}$$

Using Lemma 3.1, we have

$$q = \sqrt{[1-(\rho m/M)^2]}. \tag{16}$$

(ii) From (12) and Rule (G), one has

$$-\lambda_i \langle f'(x_i), h_i\rangle - \tfrac{1}{2}\lambda_i^2 \langle h_i, H_i h_i\rangle \leq -\lambda_i(1-\alpha)\langle f'(x_i), h_i\rangle, \tag{17}$$

or

$$\lambda_i \geq -2\alpha \langle f'(x_i), h_i\rangle / \langle h_i, H_i h_i\rangle. \tag{18}$$

From Rule (G), Assumption (A4), and (2), we have

$$\begin{aligned} f(x_i) - f(x_i + \lambda_i h) &\geq -\alpha\lambda_i \langle f'(x_i), h_i\rangle \\ &\geq 2\alpha^2 \langle f'(x_i), h_i\rangle^2 / \langle h_i, H_i h_i\rangle \geq (2\alpha^2\rho^2/M)|f'(x_i)|^2. \end{aligned} \tag{19}$$

Using Lemma 3.1, one has

$$q = \sqrt{[1-(2\alpha\rho m/M)^2]}.$$

(iii) From (12) and (6), we have

$$-(\lambda_i/\beta)\langle f'(x_i), h_i\rangle - (\lambda_i^2/2\beta^2)\langle h_i, H_i h_i\rangle \leq -(\lambda_i/\beta)\alpha\langle f'(x_i), h_i\rangle, \tag{20}$$

or

$$\lambda_i \geq -2(1-\alpha)\beta\langle f'(x_i), h_i\rangle/\langle h_i, H_i h_i\rangle. \tag{21}$$

From (4), one has

$$\begin{aligned} f(x_i)-f(x_i+\lambda_i h) &\geq -\alpha\lambda_i\langle f'(x_i), h_i\rangle \\ &\geq 2\alpha(1-\alpha)\beta\langle f'(x_i), h_i\rangle^2/\langle h_i, H_i h_i\rangle \\ &\geq (2\alpha(1-\alpha)\beta\rho^2/M)|f'(x_i)|^2. \end{aligned} \tag{22}$$

Using Lemma 3.1, we have

$$q=\sqrt{[1-\alpha(1-\alpha)\beta(2\rho m/M)^2]}.$$

(iv) Noting that either $\lambda_i = d$ or (6) is satisfied, then using (21) we have that either $\lambda_i = d$ or

$$\lambda_i \geq 2(1-\alpha)\beta\langle f'(x_i), \Gamma_i f'(x_i)\rangle/\langle f'(x_i), \Gamma_i H_i \Gamma_i f'(x_i)\rangle.$$

Therefore, from (5),

$$\begin{aligned} f(x_i)-f(x_i+\lambda_i h) &\geq \alpha\lambda_i\langle f'(x_i), \Gamma_i f'(x_i)\rangle \\ &\geq \min\{\alpha d\langle f'(x_i), \Gamma_i f'(x_i)\rangle, \\ &\quad 2\alpha(1-\alpha)\beta\langle f'(x_i), \Gamma_i f'(x_i)\rangle^2/\langle f'(x_i), \Gamma_i H_i \Gamma_i f'(x_i)\rangle\}, \end{aligned}$$

where

$$\gamma = \min(\alpha dg, 2\alpha(1-\alpha)\beta g^2/G^2 M).$$

Using Lemma 3.1, we have

$$q=\sqrt{(1-2\gamma m^2/M)}.$$

Note that $q<1$, since

$$\alpha<\tfrac{1}{2}, \qquad \beta<1, \qquad g/G<1, \qquad m/M<1.$$

Notice the relations between the rates for the three rules (M), (G), (A′). When α approaches $\frac{1}{2}$, the rate of the descent algorithm with Rule (G) approaches the rate when Rule (M) is used. Similarly, when $\alpha=\frac{1}{2}$ in Rule (A′) and β approaches 1, the rate of Rule (A′) approaches that of Rules (M) and (G).

Deflected Gradient Method. We shall now consider algorithms of the form (3). Note that

$$h_i = -\Gamma_i f'(x_i)$$

satisfies (2) with

$$\rho = g/G.$$

We are, however, interested in obtaining tighter bounds on rate of convergence.

Assumption (A5). For these results, we require that f satisfy Assumptions (A3) and (A4) and that the eigenvalues of $\Gamma_i f''(x_i)$ lie in the interval $[r, R]$, where $r > 0$ and $R < \infty$. This requirement on the eigenvalues of $\Gamma_i f''(x_i)$ is equivalent to requiring either

$$0 < r \leq \langle y, \Gamma_i^{1/2} f''(x_i) \Gamma_i^{1/2} y \rangle / \langle y, y \rangle \leq R$$

or

$$0 < r \leq \langle y, [f''(x_i)]^{1/2} \Gamma_i [f''(x_i)]^{1/2} y \rangle / \langle y, y \rangle \leq R,$$

for all $y \in R^n$.

Convergence Rate with Predetermined Stepsize. The following theorem, which derives rate of convergence for algorithms of the form (3) with stepsize in a predetermined interval, is well known; however, to the author's knowledge, the proof does not appear in the literature.

Theorem 3.2. Suppose that f satisfies Assumption (A1) and that f'' is positive definite for all x. Let $\{x_i\}$ be a sequence defined by (3) converging to a root of f', where Γ_i is positive definite and $\Gamma_i f''(x)$ has eigenvalues between $r > 0$ and $R < \infty$, for all $x \in R^n$ and for all i. Then, for any $\epsilon > 0$ and $\lambda_i \in [\epsilon, 2/R - \epsilon]$, $\{x_i\}$ converges with linear rate. In particular, if $\lambda_i = 2/(R + r)$, for all i, then $\{x_i\}$ converges with linear rate at least $(R - r)/(R + r)$.

Proof. One has

$$x_{i+1} - x^* = x_i - x^* - \lambda_i \Gamma_i f'(x_i) = x_i - x^* - \lambda_i \Gamma_i H_i (x_i - x^*), \tag{23}$$

where

$$H_i = \int_0^1 f''(x_i + t(x^* - x_i))\, dt.$$

So, we have

$$|x_{i+1} - x^*| \leq |I - \lambda_i \Gamma_i H_i|\, |x_i - x^*| \leq [\max(|1 - \lambda_i r|, |1 - \lambda_i R|)] \cdot |x_i - x^*|. \tag{24}$$

Thus, x_i converges linearly with rate at least

$$q = \overline{\lim_{i \to \infty}} \{\max(|1 - \lambda_i r|, |1 - \lambda_i R|)\},$$

if $q<1$. Clearly, $q<1$, if $\lambda_i \in [\epsilon, 2/R-\epsilon]$, for all i, where $\epsilon>0$, and

$$\lambda_i = 2/(R+r)$$

minimizes

$$\max(|1-\lambda_i r|, |1-\lambda_i R|)$$

to obtain

$$q=(R-r)/(R+r).$$

Tighter Convergence Rate Bounds. In this subsection, we will obtain new bounds on the rate of convergence of descent methods and deflected gradient methods using the revised Kantarovitch ratio. We must first introduce a lemma which is useful in proving the remaining convergence results.

Lemma 3.2. Suppose that f satisfies Assumptions (A1), (A2), (A3), (A4). Let x_i be a sequence converging to x^*, where $f'(x^*)=0$, defined by either

(a) Equation (1), where h_i satisfies (2),

or

(b) Equation (3), where Γ_i satisfies Assumption (A5).

Furthermore, suppose that x_i satisfies

$$\phi_i(\lambda_i)=f(x_{i+1})-f(x_i)\leq -[(\phi_i'(0))^2/2\phi_i''(0)](1+s_i)+o(|f'(x_i)|^2), \tag{25}$$

where

$$\phi_i(\lambda)=f(x_i+\lambda h_i)-f(x_i).$$

Then, there exists constants K and $S>0$, such that $\{x_i\}$ converges at least linearly, if $|s_i|<S$ for $i\geq K$. Also, if $s_i\to 0$, then $\{x_i\}$ converges with linear rate at least

(a) $$d_1 \triangleq \frac{(M-m)+(M+m)\sqrt{(1-\rho^2)}}{(M+m)+(M-m)\sqrt{(1-\rho^2)}}, \qquad \text{if Eq. (1) is used,}$$

or

(b) $$d_2 \triangleq (R-r)/(R+r), \qquad \text{if Eq. (3) is used.}$$

The proof of Lemma 3.2 appears in the Appendix.

Convergence Rate Using Newton Step or Stepsize Rule (M). The convergence rate

$$q = (R-r)/(R+r)$$

is the best bound on rate of convergence known for methods of the form of (3). However, it is not practical to try to use the stepsize

$$\lambda_i = 2/(R+r),$$

since r and R are typically unknown. We shall now show that method (3), where λ_i is picked according to either Newton's method or Rule (M), converges with linear rate at least $(R-r)/(R+r)$. We shall also give a tight convergence rate bound for descent method (1) with the same choice of stepsize rules.

Theorem 3.3. Suppose that f satisfies Assumptions (A1), (A2), (A3), (A4). Let $\{x_i\}$ converging to x^* be a sequence defined by either Eq. (1), where h_i satisfies (2), or by Eq. (3), where Γ_i satisfies Assumption (A5) and λ_i is chosen according to Newton's method or Rule (M). Then, $\{x_i\}$ converges with linear rate at least

$$d_1 = \frac{(M-m)-(M+m)\sqrt{(1-\rho^2)}}{(M+m)+(M-m)\sqrt{(1-\rho^2)}}, \qquad \text{if Eq. (1) is used,}$$

or

$$d_2 = (R-r)/(R+r), \qquad \text{if Eq. (3) is used.}$$

Proof. Let

$$\phi_i(\lambda_i) = f(x_i + \lambda_i h_i) - f(x_i).$$

Then, by Taylor's theorem,

$$\phi_i(\lambda_i) = \lambda_i \phi_i'(0) + \tfrac{1}{2}\lambda_1^2 \phi_i''(0) + o(|\lambda_i h_i|^2). \tag{26}$$

If λ_i is chosen by Newton's method, then

$$\lambda_i = -\phi_i'(0)/\phi_i''(0) = -\langle h_i, f'(x_i)\rangle/\langle h_i, f''(x_i)h_i\rangle \tag{27}$$

and

$$|\lambda_i h_i| \leq (1/m)|f'(x_i)|. \tag{28}$$

Thus,

$$\phi(\lambda_i) = -(\phi_i'(0))^2/2\phi_i''(0) + o(|f'(x_i)|^2). \tag{29}$$

If λ_i is chosen by Rule (M), then

$$f(x_{i+1}) - f(x_i) \leq f(\bar{x}_{i+1}) - f(x_i), \tag{30}$$

where $\bar{x}_{i+1}$ is a Newton's step along h_i from x_i, that is,

$$\bar{x}_{i+1} = x_i - [\phi_i'(0)/\phi_i''(0)]h_i. \tag{31}$$

Thus, the right-hand side of (30) is equal to (29). Therefore, for λ_i chosen by Newton's method (29) or Rule (M),

$$\phi(\lambda_i) = f(x_{i+1}) - f(x_i) \leq -[\phi_i'(0)]^2/2\phi_i''(0) + o(|f'(x_i)|^2). \tag{32}$$

The theorem now follows from Lemma 3.2.

Asymptotic Properties of Rules (G) and (A'). We shall now show that the rate of convergence of either (1) or (3) with either Rule (G) or (A'), with appropriate choices of α and β, can be made to approach the Kantorovitch ratio. To be precise, we have the following theorem.

Theorem 3.4. Suppose that f satisfies Assumptions (A1), (A2), (A3), (A4). Let $\{x_i\}$ be a sequence converging to x^* defined either by Eq. (1), where h_i satisfies (2), or Eq. (3), where Γ_i satisfies Assumption (A5). Also, let λ_i be chosen according to either Rule (G), with $\alpha = \frac{1}{2} - \epsilon$, or Rule (A'), with $\alpha = \frac{1}{2} - \epsilon$ and $\beta = \frac{1}{2} + \epsilon$, $0 < \epsilon < \frac{1}{2}$. Then, $\{x_i\}$ converges linearly with rate $q_1(\epsilon)$ if Eq. (1) is used or $q_2(\epsilon)$ if Eq. (3) is used, where

$$\lim_{\epsilon \to 0} q_k(\epsilon) = d_k$$

and

$$d_1 = \frac{(M-m) + (M+m)\sqrt{(1-\rho^2)}}{(M+m) + (M-m)\sqrt{(1-\rho^2)}},$$

$$d_2 = (R-r)/(R+r).$$

Proof. For Rule (G), using (4) and (26), we have

$$\tfrac{1}{2} - \epsilon \leq 1 + [\lambda_i \phi_i''(0)/2\phi_i'(0)][o_1(|\lambda_i h_i|)^2/\lambda_i \phi_i'(0)] < \tfrac{1}{2} + \epsilon, \tag{33}$$

where we recall that

$$\phi_i(\lambda) = f(x_i + \lambda h_i) - f(x_i).$$

From (5), the left-hand inequality in (33) holds for Rule (A'). From (6) and (26),

$$1 + \lambda_i \phi_i''(0)/2\beta\phi_i'(0) + \beta\, o_2(|\lambda_i h_i|^2)/\lambda_i \phi_i'(0) < \tfrac{1}{2} - \epsilon. \tag{34}$$

Multiplying through by

$$\beta = (\tfrac{1}{2}-\epsilon)/(\tfrac{1}{2}+\epsilon),$$

and noting that

$$\beta = 1-2\epsilon/(\tfrac{1}{2}+\epsilon),$$

we get

$$1+\lambda_i\phi_i''(0)/2\phi_i'(0)+\beta^2 o_2(|\lambda_i h_i|^2)/\lambda_i\phi_i'(0) \leq \tfrac{1}{2}-\epsilon+2\epsilon/(\tfrac{1}{2}+\epsilon). \quad (35)$$

Combining (33), (34), (35), there exists a function o_3, which depends on which stepsize is used, such that

$$\tfrac{1}{2}-\epsilon \leq 1+\lambda_i\phi_i''(0)/2\phi_i'(0)+o_3(|\lambda_i h_i|^2)/\lambda_i\phi_i'(0) \leq \tfrac{1}{2}+3\epsilon, \quad (36)$$

when either Rule (G) or Rule (A′) is used. Since, by assumption, x_i converges to x^*, $\lambda_i h_i$ converges to zero. Therefore,

$$\tfrac{1}{2}-\epsilon \leq 1+\lambda_i\phi_i''(0)/2\phi_i'(0)+\lambda_i\nu_i \leq \tfrac{1}{2}+4\epsilon,$$

where $\nu_i \to 0$, which yields

$$\lambda_i = \frac{-\phi_i'(0)(1+H(\epsilon))}{\phi_i''(0)\{1+[\phi_i'(0)/\phi_i''(0)]\nu_i\}}, \quad (37)$$

where

$$|H(\epsilon)| < 8\epsilon.$$

Using (26), we have

$$\begin{aligned} f(x_{i+1})-f(x_i) &= -[(\phi_i'(0))^2/2\phi_i''(0)][[1-H^2(\epsilon)] \\ &\quad -2[1+H(\epsilon)]\nu_i)/\{1+[\phi_i'(0)/\phi_i''(0)]\nu_i\}^2+o(|f'(x_i)|^2) \\ &= -[(\phi_i'(0))^2/2\phi_i''(0)][1-H^2(e)](1+\nu_i')+o(|f'(x_i)|^2), \quad (38) \end{aligned}$$

where $\nu_i' \to 0$ [the boundedness of λ_i follows from (37)]. The theorem now follows from Lemma 3.2 with s_i in (25) equal to $\nu_i'[1-H^2(\epsilon)]-H^2(\epsilon)$.

Combined Stepsize Rules. It is desirable to have stepsize rules so that the overall minimization algorithm (i) is guaranteed to converge, (ii) has rate equal to the Kantorovitch ratio, and (iii) does not require extensive numbers of functional evaluations at each step. This can be accomplished by combining interval stepsize rules of the form (4) with stepsizes that minimize polynomials interpolating

$$\phi_i(\lambda) = f(x_i+\lambda h_i)-f(x_i).$$

Goldstein (Ref. 2) essentially used this idea in his quasi-Newton method; he noted that, if Γ_i in (3) converged to $f''(x^*)$, where x^* is the minimum of f,

then $\lambda_i = 1$ would satisfy the Goldstein stepsize rule [Rule (G)] for large enough i, thereby not requiring a search to satisfy (4). In this section, we shall present one such stepsize rule. This algorithm combines Rule (A) with a secant-type step to yield an algorithm which converges under very general assumptions, which converges with linear convergence rate equal to the Kantorovitch ratio, and which eventually (i.e., for large i) does not require a search along a line.

The secant-type step referred to is the step that minimizes the parabola defined by $\phi_i(0)$, $\phi_i'(0)$, $\phi_i(\xi)$, where

$$\phi_i(\xi) = f(x_i - \xi\Gamma_i f'(x_i)) - f(x_i)$$

and $\epsilon > 0$ are given. Letting $b(\xi)$ equal the step, it is easy to show that

$$b(\xi) = -\phi_i'(0)\xi^2/2[\phi_i(\xi) - \phi_i'(0)\xi].$$

The new algorithm is basically (3) with Rule (A), with the exception that, at each step of Rule (A) [i.e., each time, assuming that (5) is not satisfied, the prospective step length is reduced from $\beta^i d$ to $\beta^{i+1}d$], $f(x_i - b(\xi)\Gamma_i f'(x_i))$ is also calculated, where

$$\xi = \beta^{i+1}d.$$

The steplength ξ or $b(\xi)$ that gives the lower value of f, say λ, is then substituted into (5). If the test passes, then λ is used as the steplength; if not, ξ is again multiplied by β, and the process repeats. The complete algorithm is given below.

Algorithm

Step 0. Given x_0, $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, $d > 0$, $i = 0$, $j = 0$.

Step 1. Set

$$x = x_i - \beta^j d\Gamma_i f'(x_i).$$

Step 2. Calculate

$$\phi(\beta^j d) = f(x) - f(x_i)$$

and

$$\phi'(0) = \langle f'(x_i), \Gamma_i f'(x_i)\rangle.$$

Step 3.

$$b_j = -\phi'(0)(\beta^j d)^2/2[\phi(\beta^j d) - \phi'(0)\beta^j d].$$

Step 4. Calculate $f(x - b_j\Gamma_i f'(x_i))$.

Step 5. If

$$f(x - b_j\Gamma_i f'(x_i)) < f(x),$$

set $\lambda = b_j$. If not, set $\lambda = \beta^j d$.

Step 6. Does λ satisfy Rule (A), i.e., (5)? If yes, go to Step 7; if not, go to Step 8.

Step 7. Set $\lambda_i = \lambda$, $x_{i+1} = x_i - \lambda_i \Gamma_i f'(x_i)$, $i = i+1$, $j = 0$; go to Step 1.

Step 8. Let $j = j+1$; go to Step 1.

Theorem 3.5. Suppose that f satisfies Assumptions (A1), (A2), (A3), (A4). Let $\{x_i\}$ be a sequence converging to x^* defined by the above algorithms, and suppose that Γ_i satisfies Assumption (A5). Then, there exists a K such that, for $i > K$, $\lambda_i = b_0$ or $\lambda_i = d$ (i.e., for each i, the algorithm will pass through Step 5 only once) and $\{x_i\}$ converges linearly with rate $q = (R-r)/(R+r)$.

Proof. It is clear that $\{x_i\}$ converges with rate at least as large as would occur if only Rule (A) were used, since the decrease at each step is, by Step 5, at least as large as the decrease using Rule (A). From Theorem 3.1, $\{x_i\}$ converges linearly. Let

$$b = b_0 = -\phi_i'(0)d^2/2(\phi_i(d) - \phi_i'(0)d).$$

Using (26), we have

$$[\phi_i(d) - \phi'(0)d]/d^2 = \tfrac{1}{2}\phi_i''(0) + o_1(|f'(x_i)|^2), \tag{39}$$

and so

$$b = -\phi_i'(0)/[\phi_i''(0) + 2o_1(|f'(x_i)|^2)]. \tag{40}$$

Again, using (26), we have

$$\begin{aligned}\phi_i(b)/b\phi_i'(0) &= 1 + b\phi_i''(0)/2\phi_i'(0) + o_2(|f'(x_i)|^2)/b\phi_i'(0)\\ &= 1 - \phi_i''(0)/[2\phi_i''(0) + 4o_1(|f'(x_i)|^2)]\\ &\quad - \{[\phi_i''(0) + 2o_1(|f'(x_i)|^2)]/(\phi_i'(0))^2\}o_2(|f'(x_i)|^2),\end{aligned} \tag{41}$$

which converges to $\frac{1}{2}$ as $i \to \infty$. Therefore, since $\alpha < \frac{1}{2}$, there exists a K such that (5) will be satisfied for $i \geq K$, and thus $\lambda_i = b_0$ or $\lambda_i = d$ for $i \geq K$. Combining (40) and (41), we have

$$\phi_i(b) = (\tfrac{1}{2} + \epsilon_i)b\phi_i'(0) = -[(\phi_i'(0))^2/2\phi_i''(0)](1 + 2\epsilon_i) + o_3(|f'(x_i)|^2), \tag{42}$$

where $\epsilon_i \to 0$ as $i \to \infty$. Suppose that $i > K$, so that b satisfies (5). Then, from Step 5 in the algorithm,

$$f(x_{i+1}) \leq f(x - b\Gamma_i f'(x_i)),$$

or

$$\phi_i(\lambda) \leq \phi_i(b).$$

The theorem, therefore, follows from (42) and Lemma 3.2.

4. Conclusions

In this paper, we have investigated the convergence properties of descent methods with various rules for choosing the stepsize. We have shown that many such methods have linear convergence properties. The convergence of the methods has been compared to the convergence of the descent method where the stepsize is the minimum along a line [i.e., Rule (M)]. It has been shown that this rule can be replaced by rules of the form due to Goldstein–Armijo [i.e., Rules (G), (A), (A')], with a controllable degradation in convergence rate. Furthermore, we have exhibited a descent algorithm that combines a Goldstein–Armijo stepsize rule and a secant-type step to yield an algorithm that converges with the same rate as the descent method with Rule (M) and which eventually (i.e., close to the minimum) does not require a search to determine an acceptable stepsize.

5. Appendix: Proof of Lemma 3.2

Equation (25) implies that

$$f(x_{i+1})-f(x^*)\leq f(x_i)-f(x^*)-[(\phi_i'(0))^2(1+s_i)/2\phi_i''(0)] + o(|f'(x_i)|^2). \tag{43}$$

Since $f'(x^*)=0$,

$$f'(x_i)=f''(x^*)(x_i-x^*)+o(|x_i-x^*|^2). \tag{44}$$

Using Taylor's theorem repeatedly, we have

$$\begin{aligned} f(x_i)-f(x^*) &= \tfrac{1}{2}\langle (x_i-x^*), f''(x^*)(x_i-x^*)\rangle + o_1(|x_i-x^*|^2) \\ &= \tfrac{1}{2}\langle (x_i-x^*), f'(x_i)\rangle + o_2(|x_i-x^*|^2) \\ &= \tfrac{1}{2}\langle f'(x_i), (f''(x^*))^{-1}f'(x_i)\rangle + o_3(|x_i-x^*|^2). \end{aligned} \tag{45}$$

Since f is three times continuously differentiable,

$$f''(x^*)=f''(x_i)+o_4(|x_i-x^*|). \tag{46}$$

So, using Assumption (A3), we have

$$f(x_i)-f(x^*)=\tfrac{1}{2}\langle f'(x_i), (f''(x_i))^{-1}f'(x_i)\rangle + o_5(|x_i-x^*|^2). \tag{47}$$

Combining (43) and (47), we have

$$\begin{aligned} &[f(x_{i+1})-f(x^*)]/[f(x_i)-f(x^*)] \\ &\quad\leq 1-(\phi_i'(0))^2(1+s_i)/\phi_i''(0)[\langle f'(x_i), (f''(x_i))^{-1}f'(x_i)\rangle + o_5(|x_i-x^*|^2)] \\ &\qquad + o(|f'(x_i)|^2)/\langle f'(x_i), (f''(x_i))^{-1}f'(x_i)\rangle + o_5(|x_i-x^*|^2). \end{aligned} \tag{48}$$

Using (44) and Assumption (A3), we see that the last term on the right-hand side of (48) goes to zero as $i \to \infty$. So, there exists a sequence $\epsilon_{1,i} \to 0$ and a K such that, for $i > K$,

$$[f(x_{i+1}) - f(x^*)]/[f(x_i) - f(x^*)]$$
$$\leq 1 - (\phi_i'(0))^2(1+s_i)/\phi_i''(0)[\langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle + o_5(|x_i - x^*|^2)]$$
$$+ \epsilon_{1,i}. \tag{49}$$

If $\{x_i\}$ is a sequence formed by (1),

$$(\phi_i'(0))^2/\phi_i''(0)\langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle$$
$$= \langle h_i, f'(x_i)\rangle^2/\langle h_i, f''(x_i)h_i\rangle \langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle$$
$$\geq \rho^2 |h_i|^2 \cdot |f'(x_i)|^2/\langle h_i, f''(x_i)h_i\rangle \langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle$$
$$\geq 4Mm\rho^2/[M+m+(M-m)\sqrt{(1-\rho^2)}] \triangleq C_1, \tag{50}$$

where (2), (10), and the Schwartz inequality were used. If $\{x_i\}$ is a sequence formed by (3),

$$(\phi_i'(0))^2/\phi_i''(0)\langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle$$
$$= \langle f'(x_i), \Gamma_i f'(x_i)\rangle^2/\langle f'(x_i), \Gamma_i f''(x_i)\Gamma_i f'(x_i)\rangle \langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle$$
$$\geq 4rR/(r+R)^2 \triangleq C_2, \tag{51}$$

where we used (8), with

$$s = \Gamma_i^{1/2} f'(x_i), \qquad A = \Gamma_i^{1/2} f''(x_i) \Gamma_i^{1/2}.$$

Therefore, for either algorithm [i.e., (1) or (3)],

$$\phi_i''(0)(\langle f'(x_i), (f''(x_i))^{-1} f'(x_i)\rangle + o_5(|x_i - x^*|^2))/(\phi_i'(0))^2(1+s_i)$$
$$\leq 1/(1+s_i)C_k + \phi_i''(0) o_5(|x_i - x^*|^2)/(\phi_i'(0))^2(1+s_i)$$
$$\leq 1/(1+s_i)C_k + 2M\, o_5(|x_i - x^*|^2)/|f'(x_i)|^2$$
$$\leq 1/(1+s_i)C_k + \epsilon_{2,i}, \tag{52}$$

where $\epsilon_{2,i} \to 0$. Combining (49) and (52), we have

$$[f(x_{i+1}) - f(x^*)]/[f(x_i) - f(x^*)] \leq 1 - C_k(1+s_i) + \epsilon_{3,i}, \tag{53}$$

where $\epsilon_{3,i} \to 0$. Now,

$$1 - C_1 = 1 - \frac{4Mm\rho^2}{[M+m+(M-m)\sqrt{(1-\rho^2)}]^2} = \frac{[M-m+(M+m)\sqrt{(1-\rho^2)}]^2}{[M+m+(M-m)\sqrt{(1-\rho^2)}]^2}$$
$$= d_1^2, \tag{54}$$

$$1 - C_2 = 1 - 4rR/(r+R)^2 = [(R-r)/(R+r)]^2 = d_2^2. \tag{55}$$

Let

$$q_k^2(s_i) = d_k^2 - C_k s_i,$$

and let $S \in [0, \frac{1}{2}]$ be small enough, so that $|s_i| \leq S$ for $i \geq K$ implies

$$q_k(s_i) \leq Q_k < 1$$

(note that d_k is always less than 1). Then,

$$[f(x_{i+n}) - f(x^*)]/[f(x_i) - f(x^*)] \leq (q_k^2(s_{i+n-1}) + \epsilon_{3,i+n-1}) \cdots (q_k^2(s_i) + \epsilon_{3,i})$$

$$\leq Q_k^{2n} + \epsilon_{4,i}^n \tag{56}$$

where

$$\epsilon_{4,i}^n \to 0, \qquad \text{as} \quad i \to \infty.$$

From (45) and Assumptions (A3) and (A4),

$$m|x_i - x^*|^2 \leq f(x_i) - f(x^*) \leq M|x_i - x^*|^2. \tag{57}$$

Therefore,

$$|x_{i+n} - x^*|^2/|x_i - x^*|^2 \leq (M/m)Q_k^{2n} + \epsilon_{4,i}^n \tag{58}$$

and so

$$\overline{\lim_{i \to \infty}}[|x_{i+n} - x^*|/|x_i - x^*|] \leq Q_k^n \sqrt{(M/m)}, \tag{59}$$

where $Q_k < 1$, and so $\{x_i\}$ converges at least linearly. Also, $q_k(s_i) \to d_k$ whenever $s_i \to 0$; therefore,

$$\overline{\lim_{i \to \infty}}[|x_{i+n} - x^*|/|x_i - x^*|] \leq d_k^n, \tag{60}$$

if $s_i \to 0$. This proves the remainder of the lemma.

References

1. ORTEGA, J. M., and RHEINBOLT, W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, New York, 1970.
2. GOLDSTEIN, A. A., *Constructive Real Analysis*, Harper and Row, New York, New York, 1967.
3. ARMIJO, L., *Minimization of Functions Having Continuous Partial Derivatives*, Pacific Journal of Mathematics, Vol. 16, pp. 1–3, 1966.
4. WOLFE, P., *Convergence Conditions for Ascent Methods*, SIAM Review, Vol. 11, pp. 226–235, 1969.

5. POLAK, E., *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, New York, 1971.
6. CROCKETT, J. B., and CHERNOFF, H., *Gradient Methods of Maximization*, Pacific Journal of Mathematics, Vol. 5, pp. 33–50, 1955.
7. GOLDSTEIN, A., *Cauchy's Method of Minimization*, Numerische Mathematik, Vol. 4, pp. 146–150, 1962.
8. POLYAK, B. T., *Gradient Methods for the Minimization of Functionals*, USSR Computational Mathematics and Mathematical Physics, Vol. 3, pp. 864–878, 1963.
9. GREENSTADT, J., *On the Relative Efficiencies of Gradient Methods*, Mathematics of Computation, Vol. 21, pp. 360–367, 1967.
10. LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1973.
11. AKAIKE, H., *On the Successive Transformation of Probability Distributions and Its Application to the Analysis of the Optimum Gradient Method*, Annals of Institute of Mathematical Statistics, Tokyo, Japan, Vol. 11, pp. 1–17, 1959.
12. BAUER, F. L., and HOUSEHOLDER, A. S., *Some Inequalities Involving the Euclidean Condition of a Matrix*, Numerische Mathematik, Vol. 2, pp. 308–311, 1960.