Linear and Nonlinear Programming

Third Edition

David G. Luenberger
Stanford University

Yinyu Ye
Stanford University

Springer

CONTENTS

Hence α_k converges to some α and $\mathbf{y} = \mathbf{x} + \alpha\mathbf{d}$. Let

$$\phi(\mathbf{x},\ \mathbf{d},\ \alpha) = \frac{f(\mathbf{x}+\alpha\mathbf{d}) - f(\mathbf{x})}{\alpha\nabla f(\mathbf{x})\mathbf{d}}.$$

Then $\varepsilon \leqslant \phi(\mathbf{x}_k,\ \mathbf{d}_k,\ \alpha_k) \leqslant 1-\varepsilon$ for all k. By our assumptions on $f(\mathbf{x})$, ϕ is continuous. Thus $\phi(\mathbf{x}_k,\ \mathbf{d}_k,\ \alpha_k) \to \phi(\mathbf{x},\ \mathbf{d},\ \alpha)$ and $\varepsilon \leqslant \phi(\mathbf{x},\ \mathbf{d},\ \alpha) \leqslant 1-\varepsilon$, which implies $\mathbf{y} \in \mathbf{S}(\mathbf{x},\ \mathbf{d})$. ∎

Wolfe Test

If derivatives of the objective function, as well as its values, can be evaluated relatively easily, then the Wolfe test, which is a variation of the above, is sometimes preferred. In this case ε is selected with $0 < \varepsilon < 1/2$, and α is required to satisfy (24) and

$$\phi'(\alpha) \geqslant (1-\varepsilon)\phi'(0).$$

This test is illustrated in Fig. 8.8(c). An advantage of this test is that this last criterion is invariant to scale-factor changes, whereas (25) in the Goldstein test is not.

Backtracking

A simplified method of line search is available when a good estimate of a suitable step length is available. This is the case for the multi-dimensional Newton's method for minimization discussed in the next chapter. Here a good initial choice is $\alpha = 1$. *Backtracking* is defined by the initial guess α and two positive parameters $\eta > 1$ and $\varepsilon < 1$ (usually $\varepsilon < .5$). The stopping criterion used is the same as the first part of Amijo's rule or the Goldstein test. That is, defining $\phi(\alpha) \equiv f(\mathbf{x}_k + \alpha\mathbf{d}_k)$, the procedure is terminated at the current α if $\phi(\alpha) \leqslant \phi(0) + \varepsilon\phi'(0)\alpha$. If this criterion is not satisfied, then α is reduced by the factor $1/\eta$. That is, $\alpha_{\text{new}} = \alpha_{\text{old}}/\eta$. Often η of about 1.1 or 1.2 is used.

If the intial α (such as $\alpha = 1$) satisfies the test, then it is taken as the step size. Otherwise, α is reduced by $1/\eta$. Repeating this successively, the first α that satisfies the test is declared the final value. By definition it is known that the previous value $\alpha_{\text{old}} = \alpha_{\text{new}}\eta$ does not pass the first test, and this means that it passes the second condition of Amijo's rule.

8.6 THE METHOD OF STEEPEST DESCENT

One of the oldest and most widely known methods for minimizing a function of several variables is the method of steepest descent (often referred to as the gradient method). The method is extremely important from a theoretical viewpoint, since

it is one of the simplest for which a satisfactory analysis exists. More advanced algorithms are often motivated by an attempt to modify the basic steepest descent technique in such a way that the new algorithm will have superior convergence properties. The method of steepest descent remains, therefore, not only the technique most often first tried on a new problem but also the standard of reference against which other techniques are measured. The principles used for its analysis will be used throughout this book.

The Method

Let f have continuous first partial derivatives on E^n. We will frequently have need for the gradient vector of f and therefore we introduce some simplifying notation. The gradient $\nabla f(\mathbf{x})$ is, according to our conventions, defined as a n-dimensional *row* vector. For convenience we define the n-dimensional *column* vector $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})^T$. When there is no chance for ambiguity, we sometimes suppress the argument $\mathbf{x}$ and, for example, write $\mathbf{g}_k$ for $\mathbf{g}(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)^T$.

The method of steepest descent is defined by the iterative algorithm

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k,$$

where α_k is a nonnegative scalar minimizing $f(\mathbf{x}_k - \alpha \mathbf{g}_k)$. In words, from the point $\mathbf{x}_k$ we search along the direction of the negative gradient $-\mathbf{g}_k$ to a minimum point on this line; this minimum point is taken to be $\mathbf{x}_{k+1}$.

In formal terms, the overall algorithm $\mathbf{A} : E^n \to E^n$ which gives $\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k)$ can be decomposed in the form $\mathbf{A} = \mathbf{SG}$. Here $\mathbf{G} : E^n \to E^{2n}$ is defined by $\mathbf{G}(\mathbf{x}) = (\mathbf{x}, -\mathbf{g}(\mathbf{x}))$, giving the initial point and direction of a line search. This is followed by the line search $\mathbf{S} : E^{2n} \to E^n$ defined in Section 8.4.

Global Convergence

It was shown in Section 8.4 that $\mathbf{S}$ is closed if $\nabla f(\mathbf{x}) \neq \mathbf{0}$, and it is clear that $\mathbf{G}$ is continuous. Therefore, by Corollary 2 in Section 7.7 $\mathbf{A}$ is closed.

We define the solution set to be the points $\mathbf{x}$ where $\nabla f(\mathbf{x}) = \mathbf{0}$. Then $Z(\mathbf{x}) = f(\mathbf{x})$ is a descent function for $\mathbf{A}$, since for $\nabla f(\mathbf{x}) \neq \mathbf{0}$

$$\lim_{0 \leqslant \alpha < \infty} f(\mathbf{x} - \alpha \mathbf{g}(\mathbf{x})) < f(\mathbf{x}).$$

Thus by the Global Convergence Theorem, if the sequence $\{\mathbf{x}_k\}$ is bounded, it will have limit points and each of these is a solution.

The Quadratic Case

Essentially all of the important local convergence characteristics of the method of steepest descent are revealed by an investigation of the method when applied to quadratic problems. Consider

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{x}^T\mathbf{b}, \tag{26}$$

where $\mathbf{Q}$ is a positive definite symmetric $n \times n$ matrix. Since $\mathbf{Q}$ is positive definite, all of its eigenvalues are positive. We assume that these eigenvalues are ordered: $0 < a = \lambda_1 \leqslant \lambda_2 \ldots \leqslant \lambda_n = A$. With $\mathbf{Q}$ positive definite, it follows (from Proposition 5, Section 7.4) that f is strictly convex.

The unique minimum point of f can be found directly, by setting the gradient to zero, as the vector $\mathbf{x}^*$ satisfying

$$\mathbf{Q}\mathbf{x}^* = \mathbf{b}. \tag{27}$$

Moreover, introducing the function

$$E(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{x} - \mathbf{x}^*), \tag{28}$$

we have $E(\mathbf{x}) = f(\mathbf{x}) + (1/2)\mathbf{x}^{*T}\mathbf{Q}\mathbf{x}^*$, which shows that the function E differs from f only by a constant. For many purposes then, it will be convenient to consider that we are minimizing E rather than f.

The gradient (of both f and E) is given explicitly by

$$\mathbf{g}(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{b}. \tag{29}$$

Thus the method of steepest descent can be expressed as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k\mathbf{g}_k, \tag{30}$$

where $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$ and where α_k minimizes $f(\mathbf{x}_k - \alpha\mathbf{g}_k)$. We can, however, in this special case, determine the value of α_k explicitly. We have, by definition (26),

$$f(\mathbf{x}_k - \alpha\mathbf{g}_k) = \tfrac{1}{2}(\mathbf{x}_k - \alpha\mathbf{g}_k)^T\mathbf{Q}(\mathbf{x}_k - \alpha\mathbf{g}_k) - (\mathbf{x}_k - \alpha\mathbf{g}_k)^T\mathbf{b},$$

which (as can be found by differentiating with respect to α) is minimized at

$$\alpha_k = \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{Q}\mathbf{g}_k}. \tag{31}$$

Hence the method of steepest descent (30) takes the explicit form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{Q}\mathbf{g}_k}\right)\mathbf{g}_k, \tag{32}$$

where $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$.

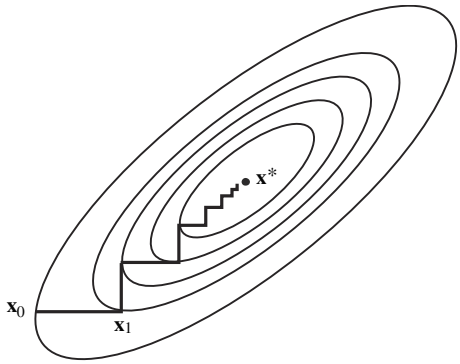

Fig. 8.9 Steepest descent

The function f and the steepest descent process can be illustrated as in Fig. 8.9 by showing contours of constant values of f and a typical sequence developed by the process. The contours of f are n-dimensional ellipsoids with axes in the directions of the n-mutually orthogonal eigenvectors of $\mathbf{Q}$. The axis corresponding to the ith eigenvector has length proportional to $1/\lambda_i$. We now analyze this process and show that the rate of convergence depends on the ratio of the lengths of the axes of the elliptical contours of f, that is, on the eccentricity of the ellipsoids.

Lemma 1. *The iterative process (32) satisfies*

$$E(\mathbf{x}_{k+1}) = \left\{1 - \frac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{(\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k)(\mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k)}\right\} E(\mathbf{x}_k). \tag{33}$$

Proof. The proof is by direct computation. We have, setting $\mathbf{y}_k = \mathbf{x}_k - \mathbf{x}^*$,

$$\frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} = \frac{2\alpha_k \mathbf{g}_k^T \mathbf{Q} \mathbf{y}_k - \alpha_k^2 \mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k}{\mathbf{y}_k^T \mathbf{Q} \mathbf{y}_k}.$$

Using $\mathbf{g}_k = \mathbf{Q}\mathbf{y}_k$ we have

$$\begin{aligned} \frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} &= \frac{\dfrac{2(\mathbf{g}_k^T \mathbf{g}_k)^2}{(\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k)} - \dfrac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{(\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k)}}{\mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k} \\ &= \frac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{(\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k)(\mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k)}. \blacksquare \end{aligned}$$

In order to obtain a bound on the rate of convergence, we need a bound on the right-hand side of (33). The best bound is due to Kantorovich and his lemma, stated below, is a useful general tool in convergence analysis.

Kantorovich inequality: *Let* $\mathbf{Q}$ *be a positive definite symmetric* $n \times n$ *matrix. For any vector* $\mathbf{x}$ *there holds*

$$\frac{(\mathbf{x}^T\mathbf{x})^2}{(\mathbf{x}^T\mathbf{Q}\mathbf{x})(\mathbf{x}^T\mathbf{Q}^{-1}\mathbf{x})} \geqslant \frac{4aA}{(a+A)^2}, \tag{34}$$

where a *and* $\mathbf{A}$ *are, respectively, the smallest and largest eigenvalues of* $\mathbf{Q}$.

Proof. Let the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{Q}$ satisfy

$$0 < a = \lambda_1 \leqslant \lambda_2 \ldots \leqslant \lambda_n = A.$$

By an appropriate change of coordinates the matrix $\mathbf{Q}$ becomes diagonal with diagonal $(\lambda_1, \lambda_2, \ldots, \lambda_n)$. In this coordinate system we have

$$\frac{(\mathbf{x}^T\mathbf{x})^2}{(\mathbf{x}^T\mathbf{Q}\mathbf{x})(\mathbf{x}^T\mathbf{Q}^{-1}\mathbf{x})} = \frac{(\sum_{i=1}^n x_i^2)^2}{(\sum_{i=1}^n \lambda_i x_i^2)(\sum_{i=1}^n (x_i^2/\lambda_i))},$$

which can be written as

$$\frac{(\mathbf{x}^T\mathbf{x})^2}{(\mathbf{x}^T\mathbf{Q}\mathbf{x})(\mathbf{x}^T\mathbf{Q}^{-1}\mathbf{x})} = \frac{1/\sum_{i=1}^n \xi_i\lambda_i}{\sum_{i=1}^n (\xi_i/\lambda_i)} \equiv \frac{\phi(\xi)}{\psi(\xi)},$$

where $\xi_i = x_i^2/\sum_{i=1}^n x_i^2$. We have converted the expression to the ratio of two functions involving convex combinations; one a combination of λ_i's; the other a combination of $1/\lambda_i$'s. The situation is shown pictorially in Fig. 8.10. The curve in the figure represents the function $1/\lambda$. Since $\sum_{i=1}^n \xi_i\lambda_i$ is a point between λ_1 and λ_n, the value of $\phi(\xi)$ is a point on the curve. On the other hand, the value of $\psi(\xi)$ is a convex combination of points on the curve and its value corresponds to a point in the shaded region. For the same vector ξ both functions are represented by points on the same vertical line. The minimum value of this ratio is achieved for some $\lambda = \xi_1\lambda_1 + \xi_n\lambda_n$, with $\xi_1 + \xi_n = 1$. Using the relation $\xi_1/\lambda_1 + \xi_n/\lambda_n = (\lambda_1 + \lambda_n - \xi_1\lambda_1 - \xi_n\lambda_n)/\lambda_1\lambda_n$, an appropriate bound is

$$\frac{\phi(\xi)}{\psi(\xi)} \geqslant \lim_{\lambda_1 \leqslant \lambda \leqslant \lambda_n} \frac{(1/\lambda)}{(\lambda_1 + \lambda_n - \lambda)/(\lambda_1\lambda_n)}.$$

The minimum is achieved at $\lambda = (\lambda_1 + \lambda_n)/2$, yielding

$$\frac{\phi(\xi)}{\psi(\xi)} \geqslant \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}. \blacksquare$$

Combining the above two lemmas, we obtain the central result on the convergence of the method of steepest descent.

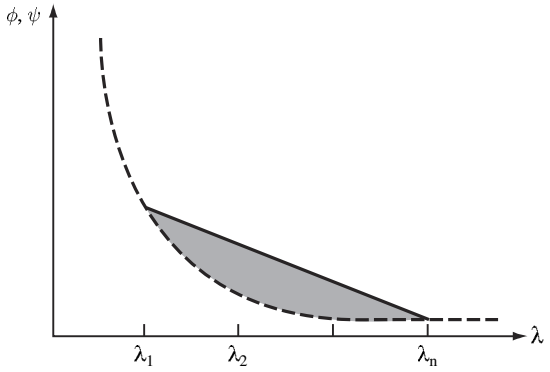

Fig. 8.10 Kantorovich inequality

Theorem. (Steepest descent—quadratic case). *For any* $\mathbf{x}_0 \in E^n$ *the method of steepest descent (32) converges to the unique minimum point* $\mathbf{x}^*$ *of f. Furthermore, with* $E(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\mathbf{Q}(\mathbf{x}-\mathbf{x}^*)$*, there holds at every step k*

$$E(\mathbf{x}_{k+1}) \leqslant \left(\frac{A-a}{A+a}\right)^2 E(\mathbf{x}_k). \tag{35}$$

Proof. By Lemma 1 and the Kantorovich inequality

$$E(\mathbf{x}_{k+1}) \leqslant \left\{1 - \frac{4aA}{(A+a)^2}\right\} E(\mathbf{x}_k) = \left(\frac{A-a}{A+a}\right)^2 E(\mathbf{x}_k).$$

It follows immediately that $E(\mathbf{x}_k) \to 0$ and hence, since $\mathbf{Q}$ is positive definite, that $\mathbf{x}_k \to \mathbf{x}^*$. ∎

Roughly speaking, the above theorem says that the convergence rate of steepest descent is slowed as the contours of f become more eccentric. If $a = A$, corresponding to circular contours, convergence occurs in a single step. Note, however, that even if $n-1$ of the n eigenvalues are equal and the remaining one is a great distance from these, convergence will be slow, and hence a single abnormal eigenvalue can destroy the effectiveness of steepest descent.

In the terminology introduced in Section 7.8, the above theorem states that with respect to the error function E (or equivalently f) the method of steepest descent converges linearly with a ratio no greater than $[(A-a)/(A+a)]^2$. The actual rate depends on the initial point $\mathbf{x}_0$. However, for some initial points the bound is actually achieved. Furthermore, it has been shown by Akaike that, if the ratio is unfavorable, the process is very likely to converge at a rate close to the bound. Thus, somewhat loosely but with reasonable justification, we say that the convergence ratio of steepest descent *is* $[(A-a)/(A+a)]^2$.

It should be noted that the convergence rate actually depends only on the ratio $r = A/a$ of the largest to the smallest eigenvalue. Thus the convergence ratio is

$$\left(\frac{A-a}{A+a}\right)^2 = \left(\frac{r-1}{r+1}\right)^2,$$

which clearly shows that convergence is slowed as r increases. The ratio r, which is the single number associated with the matrix $\mathbf{Q}$ that characterizes convergence, is often called the *condition number* of the matrix.

Example. Let us take

$$\mathbf{Q} = \begin{bmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{bmatrix}$$

$$\mathbf{b} = (0.76,\ 0.08,\ 1.12,\ 0.68).$$

For this matrix it can be calculated that $a = 0.52,\ A = 0.94$ and hence $r = 1.8$. This is a very favorable condition number and leads to the convergence ratio $[(A-a)/(A+a)]^2 = 0.081$. Thus each iteration will reduce the error in the objective by more than a factor of ten; or, equivalently, each iteration will add about one more digit of accuracy. Indeed, starting from the origin the sequence of values obtained by steepest descent as shown in Table 8.1 is consistent with this estimate.

The Nonquadratic Case

For nonquadratic functions, we expect that steepest descent will also do reasonably well if the condition number is modest. Fortunately, we are able to establish estimates of the progress of the method when the Hessian matrix is always positive definite. Specifically, we assume that the Hessian matrix is bounded above and below as $a\mathbf{I} \leqslant \mathbf{F}(\bar{\mathbf{x}}) \leqslant A\mathbf{I}$. (Thus f is *strongly* convex.) We present three analyses:

Table 8.1 Solution to Example

Step k	$f(\mathbf{x}_k)$
0	0
1	−2.1563625
2	−2.1744062
3	−2.1746440
4	−2.1746585
5	−2.1746595
6	−2.1746595

Solution point $\mathbf{x}^* = (1.534965,\ 0.1220097,\ 1.975156,\ 1.412954)$

1. **Exact line search**. Given a point $\mathbf{x_k}$, we have for any α

$$f(\mathbf{x}_k - \alpha \mathbf{g}(\mathbf{x}_k)) \leqslant f(\mathbf{x}_k)) - \alpha \mathbf{g}(\mathbf{x}_k)^T \mathbf{g}(\mathbf{x}_k) + \frac{A\alpha^2}{2} \mathbf{g}(\mathbf{x}_k)^T \mathbf{g}(\mathbf{x}_k). \tag{36}$$

Minimizing both sides separately with respect to α the inequality will hold for the two minima. The minimum of the left hand side is $f(\mathbf{x}_{k+1})$. The minimum of the right hand side occurs at $\alpha = 1/A$, yielding the result

$$f(\mathbf{x}_{k+1}) \leqslant f(\mathbf{x}_k) - \frac{1}{2A} |\mathbf{g}(\mathbf{x}_k)|^2.$$

where $|\mathbf{g}(\mathbf{x_k})|^2 \equiv \mathbf{g}(\mathbf{x_k})^T \mathbf{g}(\mathbf{x_k})$. Subtracting the optimal value $f^* = f(\mathbf{x}^*)$ from both sides produces

$$f(\mathbf{x_{k+1}}) - \mathbf{f}^* \leqslant \mathbf{f}(\mathbf{x}_k) - \mathbf{f}^* - \frac{1}{2\mathbf{A}} |\mathbf{g}(\mathbf{x}_k)|^2. \tag{37}$$

In a similar way, for any $\mathbf{x}$ there holds

$$f(\mathbf{x}) \geqslant f(\mathbf{x_k}) + \mathbf{g}(\mathbf{x_k})^T (\mathbf{x} - \mathbf{x_k}) + \frac{\mathbf{a}}{2} |\mathbf{x} - \mathbf{x_k}|^2.$$

Again we can minimize both sides separately. The minimum of the left hand side is f^* the optimal solution value. Minimizing the right hand side leads to the quadratic optimization problem. The solution is $\overline{\mathbf{x}} = \mathbf{x}_k - \mathbf{g}(\mathbf{x}_k)/a$. Substituting this $\overline{\mathbf{x}}$ in the right hand side of the inequality gives

$$f^* \geqslant f(\mathbf{x}_k) - \frac{1}{2a} |\mathbf{g}(\mathbf{x}_k)|^2. \tag{38}$$

From (38) we have

$$-|\mathbf{g}(\mathbf{x}_k)|^2 \leqslant 2a[f^* - f(\mathbf{x}_k)]. \tag{39}$$

Substituting this in (37) gives

$$f(\mathbf{x}_{k+1}) - f^* \leqslant (1 - a/A)[f(\mathbf{x}_k) - f^*]. \tag{40}$$

This shows that the method of steepest descent makes progress even when it is not close to the solution.

2. **Other stopping criteria**. As an example of how other stopping criteria can be treated, we examine the rate of convergence when using Amijo's rule with $\varepsilon < .5$ and $\eta > 1$. Note first that the inequality $t \geqslant t^2$ for $0 \leqslant t \leqslant 1$ implies by a change of variable that

$$-\alpha + \frac{\alpha^2 A}{2} \leq -\alpha/2$$

for $0 \leqslant \alpha \leqslant 1/A$. Then using (36) we have that for $\alpha < 1/A$

$$\begin{aligned} f(\mathbf{x}_k - \alpha \mathbf{g}(\mathbf{x}_k)) &\leq f(\mathbf{x}_k) - \alpha |\mathbf{g}(\mathbf{x}_k)|^2 + .5\alpha^2 A |\mathbf{g}(\mathbf{x}_l)|^2 \\ &\leq f(\mathbf{x}_k) - .5\alpha |\mathbf{g}(\mathbf{x}_k)|^2 \\ &< f(\mathbf{x}_k) - \varepsilon\alpha |\mathbf{g}(\mathbf{x}_k)|^2 \end{aligned}$$

since $\varepsilon < .5$. This means that the first part of the stopping criterion is satisfied for $\alpha < 1/A$.

The second part of the stopping criterion states that $\eta\alpha$ does not satisfy the first criterion and thus the final α must satisfy $\alpha \geq 1/(\eta A)$. Therefore the inequality of the first part of the criterion implies

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\varepsilon}{\eta A} |\mathbf{g}(\mathbf{x}_k)|^2.$$

Subtracting f^* from both sides,

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \frac{\varepsilon}{\eta A} |\mathbf{g}(\mathbf{x}_k)|^2.$$

Finally, using (39) we obtain

$$f(\mathbf{x}_{k+1}) - f^* \leq [1 - (2\varepsilon a/\eta A)](f(\mathbf{x}_k) - f^*).$$

Clearly $2\varepsilon a/\eta A < 1$ and hence there is linear convergence. Notice if that in fact ε is chosen very close to .5 and η is chosen very close to 1, then the stopping condition demands that the α be restricted to a very small range, and the estimated rate of convergence is very close to the estimate obtained above for exact line search.

3. **Asymptotic convergence**. We expect that as the points generated by steepest descent approach the solution point, the convergence characteristics will be close to those inherent for quadratic functions. This is indeed the case.

The general procedure for proving such a result, which is applicable to most methods having unity order of convergence, is to use the Hessian of the objective at the solution point as if it were the $\mathbf{Q}$ matrix of a quadratic problem. The particular theorem stated below is a special case of a theorem in Section 12.5 so we do not prove it here; but it illustrates the generalizability of an analysis of quadratic problems.

Theorem. *Suppose f is defined on E^n, has continuous second partial derivatives, and has a relative minimum at $\mathbf{x}^*$. Suppose further that the Hessian matrix of f, $\mathbf{F}(\mathbf{x}^*)$, has smallest eigenvalue $a > 0$ and largest eigenvalue $A > 0$. If $\{\mathbf{x}_k\}$ is a sequence generated by the method of steepest descent that converges to $\mathbf{x}^*$, then the sequence of objective values $\{f(\mathbf{x}_k)\}$ converges to $f(\mathbf{x}^*)$ linearly with a convergence ratio no greater than $[(A-a)/(A+a)]^2$.*

8.9 COORDINATE DESCENT METHODS

The algorithms discussed in this section are sometimes attractive because of their easy implementation. Generally, however, their convergence properties are poorer than steepest descent.

Let f be a function on E^n having continuous first partial derivatives. Given a point $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, descent with respect to the coordinate x_i (i fixed) means that one solves

$$\underset{x_i}{\text{minimize}} \; f(x_1, x_2, \ldots, x_n).$$

Thus only changes in the single component x_i are allowed in seeking a new and better vector $\mathbf{x}$. In our general terminology, each such descent can be regarded as a descent in the direction $\mathbf{e}_i$ (or $-\mathbf{e}_i$) where $\mathbf{e}_i$ is the ith unit vector. By sequentially minimizing with respect to different components, a relative minimum of f might ultimately be determined.

There are a number of ways that this concept can be developed into a full algorithm. The *cyclic coordinate descent* algorithm minimizes f cyclically with respect to the coordinate variables. Thus x_1 is changed first, then x_2 and so forth through x_n. The process is then repeated starting with x_1 again. A variation of this is the *Aitken double sweep method*. In this procedure one searches over $x_1, x_2, \ldots, x_n$, in that order, and then comes back in the order $x_{n-1}, x_{n-2}, \ldots, x_1$. These cyclic methods have the advantage of not requiring any information about ∇f to determine the descent directions.

If the gradient of f is available, then it is possible to select the order of descent coordinates on the basis of the gradient. A popular technique is the *Gauss–Southwell Method* where at each stage the coordinate corresponding to the largest (in absolute value) component of the gradient vector is selected for descent.

Global Convergence

It is simple to prove global convergence for cyclic coordinate descent. The algorithmic map $\mathbf{A}$ is the composition of $2n$ maps

$$\mathbf{A} = \mathbf{S}\mathbf{C}^n\mathbf{S}\mathbf{C}^{n-1}\ldots\mathbf{S}\mathbf{C}^1,$$

where $\mathbf{C}^i(\mathbf{x}) = (\mathbf{x}, \mathbf{e}_i)$ with $\mathbf{e}_i$ equal to the ith unit vector, and $\mathbf{S}$ is the usual line search algorithm but over the doubly infinite line rather than the semi-infinite line. The map $\mathbf{C}^i$ is obviously continuous and $\mathbf{S}$ is closed. If we assume that points are restricted to a compact set, then $\mathbf{A}$ is closed by Corollary 1, Section 7.7. We define the solution set $\Gamma = \{\mathbf{x} : \nabla f(\mathbf{x}) = \mathbf{0}\}$. If we impose the mild assumption on f that a search along any coordinate direction yields a unique minimum point, then the function $Z(\mathbf{x}) \equiv f(\mathbf{x})$ serves as a continuous descent function for $\mathbf{A}$ with respect to Γ. This is because a search along any coordinate direction either must yield a decrease or, by the uniqueness assumption, it cannot change position. Therefore,

if at a point $\mathbf{x}$ we have $\nabla f(\mathbf{x}) \neq \mathbf{0}$, then at least one component of $\nabla f(\mathbf{x})$ does not vanish and a search along the corresponding coordinate direction must yield a decrease.

Local Convergence Rate

It is difficult to compare the rates of convergence of these algorithms with the rates of others that we analyze. This is partly because coordinate descent algorithms are from an entirely different general class of algorithms than, for example, steepest descent and Newton's method, since coordinate descent algorithms are unaffected by (diagonal) scale factor changes but are affected by rotation of coordinates—the opposite being true for steepest descent. Nevertheless, some comparison is possible.

It can be shown (see Exercise 20) that for the same quadratic problem as treated in Section 8.6, there holds for the Gauss–Southwell method

$$E(\mathbf{x}_{k+1}) \leqslant \left(1 - \frac{a}{A(n-1)}\right) E(\mathbf{x}_k), \tag{57}$$

where a, A are as in Section 8.6 and n is the dimension of the problem. Since

$$\left(\frac{A-a}{A+a}\right)^2 \leqslant \left(1 - \frac{a}{A}\right) \leqslant \left(1 - \frac{a}{A(n-1)}\right)^{n-1}, \tag{58}$$

we see that the bound we have for steepest descent is better than the bound we have for $n-1$ applications of the Gauss–Southwell scheme. Hence we might argue that it takes essentially $n-1$ coordinate searches to be as effective as a single gradient search. This is admittedly a crude guess, since (47) is generally not a tight bound, but the overall conclusion is consistent with the results of many experiments. Indeed, unless the variables of a problem are essentially uncoupled from each other (corresponding to a nearly diagonal Hessian matrix) coordinate descent methods seem to require about n line searches to equal the effect of one step of steepest descent.

The above discussion again illustrates the general objective that we seek in convergence analysis. By comparing the formula giving the rate of convergence for steepest descent with a bound for coordinate descent, we are able to draw some general conclusions on the relative performance of the two methods that are not dependent on specific values of a and A. Our analyses of local convergence properties, which usually involve specific formulae, are always guided by this objective of obtaining general qualitative comparisons.

Example. The quadratic problem considered in Section 8.6 with

$$\mathbf{Q} = \begin{bmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{bmatrix}$$

$$\mathbf{b} = (0.76,\ 0.08,\ 1.12,\ 0.68)$$

Table 8.3 Solutions to Example

	Value of f for various methods		
Iteration no.	Gauss-Southwell	Cyclic	Double sweep
0	0.0	0.0	0.0
1	−0.871111	−0.370256	−0.370256
2	−1.445584	−0.376011	−0.376011
3	−2.087054	−1.446460	−1.446460
4	−2.130796	−2.052949	−2.052949
5	−2.163586	−2.149690	−2.060234
6	−2.170272	−2.149693	−2.060237
7	−2.172786	−2.167983	−2.165641
8	−2.174279	−2.173169	−2.165704
9	−2.174583	−2.174392	−2.168440
10	−2.174638	−2.174397	−2.173981
11	−2.174651	−2.174582	−2.174048
12	−2.174655	−2.174643	−2.174054
13	−2.174658	−2.174656	−2.174608
14	−2.174659	−2.174656	−2.174608
15	−2.174659	−2.174658	−2.174622
16		−2.174659	−2.174655
17		−2.174659	−2.174656
18			−2.174656
19			−2.174659
20			−2.174659

was solved by the various coordinate search methods. The corresponding values of the objective function are shown in Table 8.3. Observe that the convergence rates of the three coordinate search methods are approximately equal but that they all converge about three times slower than steepest descent. This is in accord with the estimate given above for the Gauss-Southwell method, since in this case $n-1=3$.

8.10 SPACER STEPS

In some of the more complex algorithms presented in later chapters, the rule used to determine a succeeding point in an iteration may depend on several previous points rather than just the current point, or it may depend on the iteration index k. Such features are generally introduced in order to obtain a rapid rate of convergence but they can grossly complicate the analysis of global convergence.

If in such a complex sequence of steps there is inserted, perhaps irregularly but infinitely often, a step of an algorithm such as steepest descent that is known to converge, then it is not difficult to insure that the entire complex process converges. The step which is repeated infinitely often and guarantees convergence is called a *spacer step*, since it separates disjoint portions of the complex sequence. Essentially

Convergence

Global convergence of the line search methods is established by noting that a pure steepest descent step is taken every n steps and serves as a spacer step. Since the other steps do not increase the objective, and in fact hopefully they decrease it, global convergence is assured. Thus the restarting aspect of the algorithm is important for global convergence analysis, since in general one cannot guarantee that the directions $\mathbf{d}_k$ generated by the method are descent directions.

The local convergence properties of both of the above, and most other, nonquadratic extensions of the conjugate gradient method can be inferred from the quadratic analysis. Assuming that at the solution, $\mathbf{x}^*$, the matrix $\mathbf{F}(\mathbf{x}^*)$ is positive definite, we expect the asymptotic convergence rate per step to be at least as good as steepest descent, since this is true in the quadratic case. In addition to this bound on the single step rate we expect that the method is of order two with respect to each complete cycle of n steps. In other words, since one complete cycle solves a quadratic problem exactly just as Newton's method does in one step, we expect that for general nonquadratic problems there will hold $|\mathbf{x}_{k+n} - \mathbf{x}^*| \leqslant c|\mathbf{x}_k - \mathbf{x}^*|^2$ for some c and $k = 0, n, 2n, 3n, \ldots$. This can indeed be proved, and of course underlies the original motivation for the method. For problems with large n, however, a result of this type is in itself of little comfort, since we probably hope to terminate in fewer than n steps. Further discussion on this general topic is contained in Section 10.4.

Scaling and Partial Methods

Convergence of the partial conjugate gradient method, restarted every $m+1$ steps, will in general be linear. The rate will be determined by the eigenvalue structure of the Hessian matrix $\mathbf{F}(\mathbf{x}^*)$, and it may be possible to obtain fast convergence by changing the eigenvalue structure through scaling procedures. If, for example, the eigenvalues can be arranged to occur in $m+1$ bunches, the rate of the partial method will be relatively fast. Other structures can be analyzed by use of Theorem 2, Section 9.4, by using $\mathbf{F}(\mathbf{x}^*)$ rather than $\mathbf{Q}$.

9.7 PARALLEL TANGENTS

In early experiments with the method of steepest descent the path of descent was noticed to be highly zig-zag in character, making slow indirect progress toward the solution. (This phenomenon is now quite well understood and is predicted by the convergence analysis of Section 8.6.) It was also noticed that in two dimensions the solution point often lies close to the line that connects the zig-zag points, as illustrated in Fig. 9.5. This observation motivated the *accelerated gradient method* in which a complete cycle consists of taking two steepest descent steps and then searching along the line connecting the initial point and the point obtained after the two gradient steps. The method of parallel tangents (PARTAN) was developed through an attempt to extend this idea to an acceleration scheme involving all

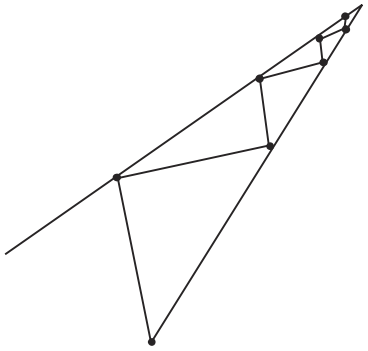

Fig. 9.5 Path of gradient method

previous steps. The original development was based largely on a special geometric property of the tangents to the contours of a quadratic function, but the method is now recognized as a particular implementation of the method of conjugate gradients, and this is the context in which it is treated here.

The algorithm is defined by reference to Fig. 9.6. Starting at an arbitrary point $\mathbf{x}_0$ the point $\mathbf{x}_1$ is found by a standard steepest descent step. After that, from a point $\mathbf{x}_k$ the corresponding $\mathbf{y}_k$ is first found by a standard steepest descent step from $\mathbf{x}_k$, and then $\mathbf{x}_{k+1}$ is taken to be the minimum point on the line connecting $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$. The process is continued for n steps and then restarted with a standard steepest descent step.

Notice that except for the first step, $\mathbf{x}_{k+1}$ is determined from $\mathbf{x}_k$, not by searching along a single line, but by searching along two lines. The direction $\mathbf{d}_k$ connecting two successive points (indicated as dotted lines in the figure) is thus determined only indirectly. We shall see, however, that, in the case where the objective function is quadratic, the $\mathbf{d}_k$'s are the same directions, and the $\mathbf{x}_k$'s are the same points, as would be generated by the method of conjugate gradients.

PARTAN Theorem. *For a quadratic function, PARTAN is equivalent to the method of conjugate gradients.*

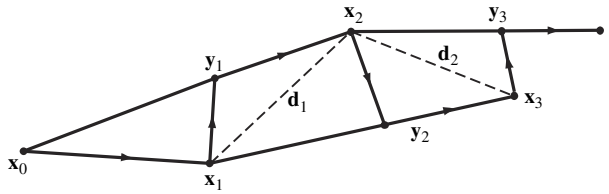

Fig. 9.6 PARTAN

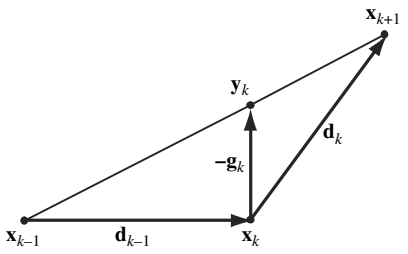

Fig. 9.7 One step of PARTAN

Proof. The proof is by induction. It is certainly true of the first step, since it is a steepest descent step. Suppose that $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$ have been generated by the conjugate gradient method and $\mathbf{x}_{k+1}$ is determined according to PARTAN. This single step is shown in Fig. 9.7. We want to show that $\mathbf{x}_{k+1}$ is the same point as would be generated by another step of the conjugate gradient method. For this to be true $\mathbf{x}_{k+1}$ must be that point which minimizes f over the plane defined by $\mathbf{d}_{k-1}$ and $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^T$. From the theory of conjugate gradients, this point will also minimize f over the subspace determined by $\mathbf{g}_k$ and all previous $\mathbf{d}_i$'s. Equivalently, we must find the point $\mathbf{x}$ where $\nabla f(\mathbf{x})$ is orthogonal to both $\mathbf{g}_k$ and $\mathbf{d}_{k-1}$. Since $\mathbf{y}_k$ minimizes f along $\mathbf{g}_k$, we see that $\nabla f(\mathbf{y}_k)$ is orthogonal to $\mathbf{g}_k$. Since $\nabla f(\mathbf{x}_{k-1})$ is contained in the subspace $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}]$ and because $\mathbf{g}_k$ is orthogonal to this subspace by the Expanding Subspace Theorem, we see that $\nabla f(\mathbf{x}_{k-1})$ is also orthogonal to $\mathbf{g}_k$. Since $\nabla f(\mathbf{x})$ is linear in $\mathbf{x}$, it follows that at every point $\mathbf{x}$ on the line through $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$ we have $\nabla f(\mathbf{x})$ orthogonal to $\mathbf{g}_k$. By minimizing f along this line, a point $\mathbf{x}_{k+1}$ is obtained where in addition $\nabla f(\mathbf{x}_{k+1})$ is orthogonal to the line. Thus $\nabla f(\mathbf{x}_{k+1})$ is orthogonal to both $\mathbf{g}_k$ and the line joining $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$. It follows that $\nabla f(\mathbf{x}_{k+1})$ is orthogonal to the plane. ∎

There are advantages and disadvantages of PARTAN relative to other methods when applied to nonquadratic problems. One attractive feature of the algorithm is its simplicity and ease of implementation. Probably its most desirable property, however, is its strong global convergence characteristics. Each step of the process is at least as good as steepest descent; since going from $\mathbf{x}_k$ to $\mathbf{y}_k$ is exactly steepest descent, and the additional move to $\mathbf{x}_{k+1}$ provides further decrease of the objective function. Thus global convergence is not tied to the fact that the process is restarted every n steps. It is suggested, however, that PARTAN should be restarted every n steps (or $n+1$ steps) so that it will behave like the conjugate gradient method near the solution.

An undesirable feature of the algorithm is that two line searches are required at each step, except the first, rather than one as is required by, say, the Fletcher–Reeves method. This is at least partially compensated by the fact that searches need not be as accurate for PARTAN, for while inaccurate searches in the Fletcher–Reeves method may yield nonsensical successive search directions, PARTAN will at least do as well as steepest descent.