Reprinted from JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS Vol. 7, No. 1, January 1971
Printed in Belgium

Convergence Rate of a Penalty-Function Scheme[1]

DAVID G. LUENBERGER[2]

Communicated by M. R. Hestenes

Abstract. The penalty-function approach is an attractive method for solving constrained nonlinear programming problems, since it brings into play all of the well-developed unconstrained optimization techniques. If, however, the classical steepest-descent method is applied to the standard penalty-function objective, the rate of convergence approaches zero as the penalty coefficient is increased to yield a close approximation to the true solution.

In this paper, it is shown that, if $m + 1$ steps of the conjugate-gradient method are successively repeated (where m is the number of constraints), the convergence rate when applied to the penalty-function objective converges at a rate predicted by the second derivative of the Lagrangian. This rate is independent of the penalty coefficient and, hence, the scheme yields reasonable convergence for a first-order method.

1. Introduction

Consider the constrained minimization problem

$$\text{minimize} \quad f(x) \qquad \text{subject to} \quad h(x) = \theta, \tag{1}$$

where f is a real-valued functional on E^n, h is a mapping from E^n into E^m, $m \leqslant n$, and θ is the null vector in E^m. More detailed assumptions are introduced later. Courant (Ref. 1) introduced the penalty-function approach by which problem (1) is approximated by the unconstrained minimization problem

$$\text{minimize} \quad f(x) + \mu \| h(x)\|^2 \tag{2}$$

[1] Paper received January 19, 1970. This research was supported by National Science Foundation, Grant No. NSF-GK-1683.

[2] Associate Professor, Department of Engineering–Economic Systems, Stanford University, Stanford, California.

for some suitable positive constant μ. Under mild conditions, it can be shown (Refs. 2–3) that, as μ tends to infinity, the solution of problem (2) tends to the solution of problem (1).

There has been a good deal of analysis of penalty-function schemes (Refs. 4–5) and, partly as a result of these analyses, the penalty-function approach is today one of the most attractive general methods for solving constrained problems. For the most part, past analysis has been concerned with studying the convergence of the solution of (2) to the solution of (1) as μ is increased. Relationships between (2) and the classical Lagrange multipliers of (1) are among the important relationships that have been discovered in this way.

Although it is known (Ref. 6) that, for large μ, (2) is ill-conditioned, and therefore steepest-descent procedures applied to (2) converge slowly, no analytical attention has been directed toward the problem of solving (2) with gradient methods. Most reported computational experience on penalty-function methods has been based on the application of second-order schemes to (2), which, although excellent when applicable, cannot be applied to numerous important problems where second-order information is not directly available. Attention then naturally focuses on first-order methods.

In Section 3 of this paper, it is shown that, near the solution, the second-order approximation term of the penalty functional is a quadratic form having m of its n eigenvalues tending to infinity with μ. Since the convergence rate of the steepest-descent method decreases as the ratio of the largest to smallest eigenvalue increases, it follows that the convergence rate of the steepest-descent method applied to the penalty-function method decreases to zero as μ is increased without bound.

In the remainder of the paper, it is shown that the effect of the m large eigenvalues can be eliminated by employing a procedure that amounts to taking $m + 1$ conjugate-gradient steps at each iteration. The convergence rate of the resulting technique then depends only on the ratio of the $(m + 1)$th largest and smallest eigenvalue. Thus, since this ratio is independent of μ, so is the convergence rate of the proposed scheme. As a byproduct of this analysis, we obtain an interesting connection between the natural rates of convergence for solving (2) and the second-order sufficiency conditions for a local minimum of (1).

2. Properties of the Lagrangian

We write vectors $x \in E^n$ as $x = (x^1, x^2, \ldots, x^n)$, and the vector-valued function h is assumed to consist of m real-valued functionals $h^1(x)$,

$h^2(x),..., h^m(x)$. We assume throughout that f and h have continuous partial derivatives up to second order. We use the notation $\nabla f(x)$, $\nabla h(x)$ for the n-dimensional row vector and $m \times n$ matrix of first partial derivatives of f and h, respectively. We assume in this section that (1) has a unique solution x_0 at which the following classical regularity assumption is satisfied: the matrix

$$\nabla h = [\partial h^i / \partial x^j], \tag{3}$$

evaluated at the point x_0, has rank m.

Under these conditions, it is well known that there is a vector $\lambda_0 \in E^m$ such that the Lagrangian

$$l(x, \lambda_0) = f(x) + \lambda_0' h(x) \tag{4}$$

is stationary at the solution x_0.

Now, introduce the subspace M of E^n

$$M = \{x: \ \nabla h(x_0)\, x = 0\}.$$

This is the tangent space of the constraint surface at x_0. Define $L(x_0, \lambda_0)$ to be the $n \times n$ second-derivative matrix of the Lagrangian l at x_0. We then obtain, under the same hypotheses as above, the second-order necessary condition $x' L(x_0, \lambda_0)\, x \geqslant 0$, $x \in M$; that is, the second derivative of the Lagrangian is positive-semidefinite on the tangent subspace M (Ref. 7).

For our convergence estimates, we shall assume the somewhat stronger condition that $L(x_0, \lambda_0)$ is actually positive-definite on M. Specifically, we assume the existence of two positive constants α and β, such that

$$\alpha I \leqslant L(x_0, \lambda_0) \leqslant \beta I \quad \text{on} \quad M, \tag{5}$$

that is, $L(x_0, \lambda_0) - \alpha I$ and $\beta I - L(x_0, \lambda_0)$ are positive-semidefinite on M. This corresponds to the standard sufficiency condition for a local minimum to (1). As we shall show, the properties of $L(x_0)$ on M determine the convergence rate for the solution of (2).

Now, consider the penalty-function method applied to (1). Let $\{\mu_i\}$ be an increasing sequence of positive constants tending toward infinity. We then have the following results (Refs. 2–5).

Theorem 2.1. For every i, let x_i be a point at which the functional $f(x) + \mu_i \| h(x) \|^2$ achieves its minimum. Then, if x_0 is any limit point of $\{x_i\}$, it follows that x_0 solves (1).

In addition, we have a connection to the Lagrange multiplier.

Theorem 2.2. Suppose that the sequence $\{x_i\}$ defined by Theorem 2.1 converges to x_0, the solution to (1). Assume further that x_0 satisfies the regularity condition (3). For each i, define $\lambda_i = 2\mu_i h(x_i)$. Then, the sequence $\{\lambda_i\}$ converges to the Lagrange multiplier for (1).

Proof. Since x_i minimizes $f(x) + \mu_i \| h(x)\|^2$, it follows immediately that, for each i,

$$\nabla f(x_i) + \lambda_i \nabla h(x_i) = 0.$$

Since $x_i \to x_0$ and since ∇f and ∇h are continuous, it follows that $\nabla f(x_i) \to \nabla f(x_0)$, $\nabla h(x_i) \to \nabla h(x_0)$. Also, since $\nabla h(x_0)$ has rank m, $\nabla h(x)$ has rank m in some neighborhood of x_0. From this, it follows that $\lambda_i \to \lambda_0$.

3. Eigenvalue Distribution

Since the convergence properties of gradient procedures are primarily determined by the second-order characteristics of the functional to which they are applied, we now turn to a study of the second-order properties of the functional in problem (2). For given μ, let x_μ be a solution of (2) and let $\lambda_\mu = 2\mu h(x_\mu)$. Then, it is easily seen that, near x_μ,

$$\begin{aligned} f(x) + \mu \| h(x)\|^2 = {} & f(x_\mu) + \nabla f(x_\mu)(x - x_\mu) \\ & + \tfrac{1}{2}(x - x_\mu)' F(x_\mu)(x - x_\mu) + \mu \| h(x_\mu) + \nabla h(x)(x_\mu - x_\mu)\|^2 \\ & + \tfrac{1}{2}(x - x_\mu)' \left[\sum_{i=1}^{m} \lambda_\mu{}^i H^i(x_\mu) \right] (x - x_\mu) + o(\| x - x_\mu \|^2), \end{aligned} \tag{6}$$

where $F(x_\mu)$ and $H^i(x_\mu)$ are the second-derivative matrices of f and h^i, respectively. Two times the matrix of the purely quadratic terms of (6) is

$$A_\mu = F(x_\mu) + \sum_{i=1}^{m} \lambda_\mu{}^i H^i(x_\mu) + 2\mu \nabla h(x_\mu)' \nabla h(x_\mu).$$

The convergence rates of various gradient schemes are primarily determined by the eigenvalue distribution A_μ. Since x_μ minimizes (2), it follows at once, by the second-order necessary conditions for a minimum, that A_μ is positive-semidefinite. We desire, however, additional information.

Observe that

$$A_\mu = L(x_\mu, \lambda_\mu) + 2\mu \nabla h(x_\mu)' \nabla h(x_\mu).$$

If, as $\mu \to \infty$, we have $x_\mu \to x_0$, then, by Theorems 2.1 and 2.2, $L(x_\mu, \lambda_\mu) \to L(x_0, \lambda_0)$. Thus, for sufficiently large μ, the eigenvalue structure of $L(x_\mu, \lambda_\mu)$ is close to that of $L(x_0, \lambda_0)$. The matrix A_μ differs from this by a matrix of rank m.

As $\mu \to \infty$, certainly m of the eigenvalues of A_μ tend to infinity with eigenvectors nearly in the subspace spanned by $\nabla h(x_0)$. The remaining eigenvectors of A_μ are orthogonal to these and, hence, lie in the tangent subspace M defined in Section 2. Thus, the remaining eigenvalue structure is, in the limit, precisely the same as that which is important for the second-order conditions for a constrained minimum, namely, that of $L(x_0, \lambda_0)$ restricted to M. We summarize these observations by the following theorem.

Theorem 3.1. Given $\epsilon > 0$, there is a μ_0 such that, for $\mu > \mu_0$, $n - m$ of the eigenvalues of A_μ lie in the interval $[\alpha - \epsilon, \beta + \epsilon]$. The remaining m eigenvalues tend to infinity, as $\mu \to \infty$.

4. Accelerated Gradient Method for Quadratic Problems

Motivated by the previous sections, we consider in this section the problem of minimizing a functional of the form

$$f(x) = x'Ax - 2b'x, \tag{7}$$

where the $n \times n$ matrix A is symmetric and positive-definite but has m large eigenvalues. The solution is, of course, $x_0 = A^{-1}b$. For the standard method of steepest descent, it can be shown that (Ref. 8)

$$\| x_k - x_0 \|^2 \leqslant c[(\gamma - \alpha)/(\gamma + \alpha)]^{2k} \| x_1 - x_0 \|^2, \tag{8}$$

where c is a constant depending on x_1, and α and γ are respectively the smallest and largest eigenvalues of A. Furthermore, this bound is in most cases exact (Ref. 9). Thus, the gradient method has an unacceptable convergence rate for our problem.

The method which we investigate is essentially that of successively repeating $m + 1$ steps of the conjugate-gradient procedure. We assume, similar to before, that the $(m + 1)$th largest eigenvalue of A is equal to β.

As is standard, define $r_k = b - Ax_k$. We consider an iteration scheme of the form

$$x_{k+1} = x_k + p_k(A)\, r_k, \tag{9}$$

where p_k is a polynomial of degree m. We select the coefficients of p_k so as to minimize $E(x_{k+1})$, where E is defined by

$$E(x) = (x - x_0)' A(x - x_0), \tag{10}$$

which implies that

$$E(x_k) = r_k' A^{-1} r_k .$$

Thus, x_{k+1} is obtained from x_k by minimizing the objective functional over the linear variety $x_k + [r_k , Ar_k ,..., A^m r_k]$. This point can be found by taking $m + 1$ conjugate-gradient steps from x_k or by explicitly determining the appropriate polynomial directly.

To find p_k explicitly, let us write

$$p_k(A) = c_1 + c_2 A + \cdots + c_{m+1} A^m. \tag{11}$$

We wish to find the constants c_i to substitute in (9) so as to minimize (10). This is an unconstrained, quadratic minimization problem which is solved by solving the normal equations

$$\begin{bmatrix} r_k'Ar_k & r_k'A^2r_k & \cdots & r_k'A^{m+1}r_k \\ r_k'A^2r_k & r_k'A^3r_k & \cdots & r_k'A^{m+2}r_k \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ r_k'A^{m+1}r_k & r_k'A^{m+2}r_k & \cdots & r_k'A^{2m+1}r_k \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{m+1} \end{bmatrix} = \begin{bmatrix} r_k'r_k \\ r_k'Ar_k \\ \vdots \\ r_k'A^m r_k \end{bmatrix}. \tag{12}$$

These are a set of linear equations involving the Gram matrix of the vectors $r_k , Ar_k ,..., A^m r_k$.

Let us now turn to the problem of determining the rate of convergence of the optimal scheme (9). We have

$$\begin{aligned} E(x_{k+1}) &= (x_{k+1} - x_0)' A(x_{k+1} - x_0) \\ &= (x_k - x_0 + p_k(A)\, r_k)' A(x_k - x_0 + p_k(A)\, r_k) \\ &= r_k'[I - Ap_k(A)]\, A^{-1}[I - Ap_k(A)]\, r_k . \end{aligned} \tag{13}$$

Let $e_1 , e_2 ,..., e_n$ be a complete orthonormal set of eigenvectors for A, with corresponding eigenvalues $\lambda_1 , \lambda_2 ,..., \lambda_n$. Let the decomposition for r_k be $r_k = \sum_{i=1}^n \xi_i^k e_i$. Then, from (13),

$$E(x_{k+1}) = \sum_{i=1}^n \lambda_i^{-1} \xi_i^2 [1 - \lambda_i p_k(\lambda_i)]^2 \leqslant \max_i\, [1 - \lambda_i p_k(\lambda_i)]^2 \cdot E(x_k). \tag{14}$$

The strict bound for the rate of convergence is determined by using the optimal p_k derived above. Using any mth-degree polynomial, however, gives a valid bound. We select p_k so that $1 - \lambda p_k(\lambda)$ vanishes at $\frac{1}{2}(\alpha + \beta)$ and at the m largest eigenvalues of A. Then, if all other eigenvalues of A are located in the interval $[\alpha, \beta]$, we obtain the bound

$$E(x_{k+1}) \leqslant \max_{\alpha \leqslant \lambda \leqslant \beta} [1 - \lambda p_k(\lambda)]^2 \cdot E(x_k).$$

Since the polynomial $q(\lambda) = 1 - \lambda p_k(\lambda)$ has $m + 1$ real roots, $q'(\lambda)$ will have m real roots which alternate between the roots of $q(\lambda)$ on the real axis. Likewise, $q''(\lambda)$ will have $m - 1$ real roots which alternate between the roots of $q'(\lambda)$. Thus, since $q(\lambda)$ has no root in the interval $(-\infty, \frac{1}{2}(\alpha + \beta))$, we see that $q''(\lambda)$ does not change sign in that interval and, since it is easily varified that $q''(0) > 0$, it follows that $q(\lambda)$ is convex for $\lambda < \frac{1}{2}(\alpha + \beta)$. Therefore, on $[0, \frac{1}{2}(\alpha + \beta)]$, $q(\lambda)$ lies below the straight line $1 - 2\lambda/(\alpha + \beta)$. Thus, we conclude that $q(\lambda) \leqslant 1 - 2\lambda/(\alpha + \beta)$ on $[0, \frac{1}{2}(\alpha + \beta)]$ and that $q'(\frac{1}{2}(\alpha + \beta)) \geqslant -1$.

We can also see that $q(\lambda) \geqslant 1 - 2\lambda/(\alpha + \beta)$ on $[\frac{1}{2}(\alpha + \beta), \beta]$, since for $q(\lambda)$ to cross first the line $1 - 2\lambda/(\alpha + \beta)$ and then the λ-axis would require at least two changes in sign of $q''(\lambda)$; whereas, at most, one root of $q''(\lambda)$ exists to the left of the second root of $q(\lambda)$. We see then that the inequality

$$| 1 - \lambda p_k(\lambda)| \leqslant 1 - 2\lambda/(\alpha + \beta)$$

is valid on the interval $[\alpha, \beta]$. Thus, we obtain the final result

$$E(x_{k+1}) \leqslant [(\beta - \alpha)/(\beta + \alpha)]^2 \, E(x_k).$$

This in turn implies that

$$E(x_{k+1}) \leqslant [(\beta - \alpha)/(\beta + \alpha)]^{2k} \, E(x_1). \tag{15}$$

Equation (15) is identical to the best bound for the method of steepest descent applied to a quadratic form having eigenvalues in $[\alpha, \beta]$.

5. Accelerated Gradient Method for Nonquadratic Problems

The process of finding the minimum of a functional over an m-dimensional subspace, which was discussed in Section 4, can be approximated by various procedures in the general nonquadratic problem. One method is to use m steps of the generalized conjugate-gradient procedure due to Fletcher and

Reeves (Ref. 10). Another is to use m steps of the method of parallel tangents (Ref. 11). Still another, and this is the method we shall analyze, although our choice is admittedly based somewhat on analytical expediency rather than practical superiority, is to use the local second-order information to determine the next point. This is similar in spirit to the nonlinear generalization of the conjugate-gradient method studied by Daniel (Ref. 12).

For the problem of minimizing a functional f (and, here, although it does not matter, we have in mind that f arises from a penalty-function approach to a constrained problem), we use the iteration

$$x_{k+1} = x_k + p_k(A_k)\, r_k\,, \tag{16}$$

where $r_k = -\nabla f(x_k)$, $A_k = F(x_k)$, and p_k is determined from (11) and (12).

To examine the convergence properties of (16), we establish some bounds relating the various quantities of the nonlinear process. We do not always derive the tightest bound possible, but favor instead simple bounds sufficient to establish our main result, Theorem 4.1.

For the next three lemmas, we invoke the following standing assumptions: (i) there is a convex region D throughout which

$$\alpha I \leqslant F(x) \leqslant \gamma I, \tag{17}$$

i.e., the second-derivative operator F is uniformly bounded and uniformly positive-definite; (ii) the $(m + 1)$th largest eigenvalue of $F(x)$ is less than β; and (iii)

$$\|f'''(x)\| \leqslant B, \tag{18}$$

i.e., the third derivative of f is uniformly bounded. We shall always assume in the lemmas below that $x_k \in D$.

Lemma 5.1. $\| x_{k+1} - x_k \| \leqslant (2/\alpha)\, \| r_k \|$, if $x_{k+1} \in D$.

Proof. From (16), we have $x_{k+1} - x_k = p_k(A_k)\, r_k$. Hence, it follows immediately that $\| x_{k+1} - x_k \| \leqslant \max_{\alpha \leqslant \lambda \leqslant \gamma} | p_k(\lambda)| \cdot \| r_k \|$. From the definition of $p_k(\lambda)$, it is clear that $| 1 - \lambda p_k(\lambda)|^2 < 1$ for λ any eigenvalue of A_k. Hence, p_k must satisfy $| p_k(\lambda)| < 2/\alpha$ for λ any eigenvalue of A_k.

Lemma 5.2. $\| r_{k+1} \| \leqslant (1 + 2\gamma/\alpha)\, \| r_k \|$, if x_{k+1} is in D.

Proof. By Lemma 5.1,

$$\| r_{k+1} \| \leqslant \| r_{k+1} - r_k \| + \| r_k \| \leqslant \gamma \, \| x_{k+1} - x_k \| + \| r_k \| \leqslant (1 + 2\gamma/\alpha)\, \| r_k \|.$$

Following Daniel (Ref. 9), we now introduce the sequence of error functionals

$$E_k(x_k) = r_k' A_k^{-1} r_k\,, \tag{19}$$

and we shall prove that this sequence has properties similar to $E(x_k)$ in Section 4. Note that

$$E_k(x_k) = (x_k - h_k)' A_k(x_k - h_k),$$

where $h_k = x_k + A_k^{-1} r_k$ is the local Newton's approximation to the solution. For convenience, we write $e_k^2 = E_k(x_k)$.

Lemma 5.3. $E_{k+1}(x_{k+1}) \leqslant \{[(\beta - \alpha)/(\beta + \alpha)]^2 + s_k\} E_k(x_k)$, if x_{k+1} is in D, where

$$s_k = (2B/\alpha^3)(1 + 2\gamma/\alpha)^2 \gamma^{3/2} e_k + (4B/\alpha^3)(1 + 2\gamma/\alpha) \gamma^{3/2} e_k\,.$$

Proof. By Lemmas 5.1 and 5.3,

$$\begin{aligned} E_{k+1}(x_{k+1}) &= r'_{k+1} A_{k+1}^{-1} r_{k+1} \\ &\leqslant r'_{k+1} A_k^{-1} r_{k+1} + (B/\alpha^2) \| r_{k+1} \|^2 \| x_{k+1} - x_k \| \\ &\leqslant r'_{k+1} A_k^{-1} r_{k+1} + (2B/\alpha^3)(1 + 2\gamma/\alpha)^2 \| r_k \|^3. \end{aligned} \tag{20}$$

Now, let $y_k = r_{k+1} - r_k - A_k(x_{k+1} - x_k)$. Clearly, $\| y_k \| \leqslant \frac{1}{2} B \| x_{k+1} - x_k \|^2 \leqslant (2B/\alpha^2) \| r_k \|^2$. Thus, we have

$$\begin{aligned} r'_{k+1} A_k^{-1} r_{k+1} &= [r_k + A_k(x_{k+1} - x_k)]' A_k^{-1} [r_k + A_k(x_{k+1} - x_k)] \\ &\quad + y_k' A_k^{-1} r_{k+1} + r'_{k+1} A_k^{-1} y_k - y_k' A_k^{-1} y_k\,. \end{aligned}$$

The first term is the local quadratic approximation to E_k, and the scheme (16) selects x_{k+1} to minimize this. Hence, from the analysis of the quadratic case and the above inequalities, we have

$$r'_{k+1} A_k^{-1} r_{k+1} \leqslant [(\beta - \alpha)/(\alpha + \beta)]^2 E_k(x_k) + (4B/\alpha^3)(1 + 2\gamma/\alpha) \| r_k \|^3. \tag{21}$$

Finally, using the inequality $\| r_k \| \leqslant e_k \sqrt{\gamma}$ and combining (20), we obtain

$$E_{k+1}(x_{k+1}) \leqslant \{[(\beta - \alpha)^2/(\beta + \alpha)]^2 + s_k\} E_k(x_k).$$

We now prove the theorem establishing the rate of convergence near the solution.

Theorem 5.1. Let the functional f defined on E^n have continuous partial derivatives up to third order. Let x_1 be arbitrary and assume that, in a ball $S(x_1, R_1)$, (i) $\alpha I \leqslant F(x) \leqslant \gamma I$ and (ii) $\|f'''(x)\| \leqslant B$, where $R_1 = (2e_1 \sqrt{\gamma})/2(1 - q_1)$. Further, assume that e_1 is so small that $q_1{}^2 = [(\beta - \alpha)/(\beta + \alpha)]^2 + s_1 < 1$. Then, the algorithm (16) generates a well-defined sequence $\{x_k\}$ converging to $x_0 \in S(x_1, R_1)$ with $\nabla f(x_0) = 0$. Also, $q_k{}^2 = [(\beta - \alpha)/(\beta + \alpha)]^2 + s_k$ decreases to $[(\beta - \alpha)/(\beta + \alpha)]^2$, and $\| x_k - x_0 \| \leqslant (2e_k \sqrt{\gamma})/\alpha(1 - q_k)$.

Proof. We take $D = S(x_1, R_1)$ in Lemmas 5.1–5.3. For $\| x - x_1 \| \leqslant (2/\alpha)\| r_1 \| \leqslant (2e_1 \sqrt{\gamma})/\alpha$, we have $x \in S(x_1, R_1)$ and, hence, D is large enough so that $x_2 \in D$ and Lemmas 5.1–5.3 hold for the first iteration. Let $R_2 = (2e_2 \sqrt{\gamma})/\alpha(1 - q_2)$. We see that $S(x_2, R_2) \subset S(x_1, R_1)$. For, if $x \in S(x_2, R_2)$,

$$\begin{aligned}\| x - x_1 \| \leqslant \| x - x_2 \| + \| x_2 - x_1 \| &\leqslant (2e_2 \sqrt{\gamma})/\alpha(1 - q_2) + (2e_1 \sqrt{\gamma})/\alpha \\ &\leqslant (2q_1 e_1 \sqrt{\gamma})/\alpha(1 - q_1) + (2e_1 \sqrt{\gamma})/\alpha = R_1 .\end{aligned}$$

Thus, all lemmas now hold in $S(x_2, R_2)$ and $e_2 < e_1$, $s_2 < s_1$, $q_2 < q_1$. This process obviously continues on by induction.

As a result of this process, we have

$$(1/\gamma) \| \nabla f(x_k)\|^2 = (1/\gamma) \| r_k \|^2 \leqslant e_k{}^2 \leqslant q_1{}^2 q_2{}^2 \cdots q_{k-1}^2 e_1{}^2,$$

and thus $\nabla f(x_k)$ converges to zero. Also, setting $R_k = (2e_k \sqrt{\gamma})/\alpha(1 - q_k)$, we have $S(x_{k+1}, R_{k+1}) \subset S(x_k, R_k)$ for all k. Therefore, for fixed k, we have $x_{k+p} \in S(x_k, R_k)$ for all $p \geqslant 0$ and, hence, $\| x_{k+p} - x_k \| \leqslant R_k \to 0$. Therefore, $\{x_k\}$ is a Cauchy sequence converging to a limit x_0. Since $\nabla f(x_k) \to 0$, we have $\nabla f(x_0) = 0$. Since $x_0 \in S(x_k, R_k)$ for each k, we have $\| x_k - x_0 \| \leqslant R_k = (2e_k \sqrt{\gamma})/\alpha(1 - q_k)$.

Finally, we point out a main corollary for the technique of penalty functions.

Corollary 5.1. Let f, h, α, β be as defined in Sections 1–2. Let $\{\mu_i\}$ be a sequence of positive constants tending to infinity and let x_{μ_i} be the solution to the problem

$$\text{minimize } f(x) + \mu_i \| h(x)\|^2. \tag{22}$$

Suppose that algorithm (16) is applied to (22), yielding the sequence $\{x_{\mu_i,k}\}$, $x_{\mu_i,k} \to x_{\mu_i}$. Suppose also that $x_{\mu_i} \to x_0$, the solution to (1). Then, given

$\epsilon > 0$, there is a μ_0 such that, for $\mu_i > \mu_0$, the tail of the sequence $\{x_{\mu_i,k}\}$ satisfies

$$\| x_{\mu_i,k+n} - x_{\mu_i} \| \leqslant c_k[(\beta - \alpha)/(\beta + \alpha) + \epsilon]^{2n}$$

for some constant c_k.

Proof. By selecting μ_0 large enough, x_{μ_i} can be made arbitrarily close to x_0 and, hence, the first $n - m$ eigenvalues of the second derivative of $f(x) + \mu_i h(x)$ will be arbitrarily close to the interval $[\alpha, \beta]$. Thus, by Theorem 5.1, the tail of the sequence $x_{\mu_i,k}$ satisfies

$$\| x_{\mu_i,k+n} - x_{\mu_i} \| \leqslant d e_{k+n},$$

where d is a constant depending on the eigenvalue structure at x_{μ_i}. By Lemma 5.3, we then obtain

$$\| x_{\mu_i,k+n} - x_{\mu_i} \| \leqslant d[(\beta - \alpha)/(\beta + \alpha) + \epsilon]^{2n} e_k = c_k[(\beta - \alpha)/(\beta + \alpha) + \epsilon]^{2n}.$$

6. Example

The method proposed in this paper has in one form or another been applied to a variety of optimization problems. In most applications, it has been found most convenient to take $m + 1$ PARTAN steps at each iteration rather than to use the procedure analyzed in Section 5. The specialist will recognize that, except for the term s_k; the convergence rates for any number of conjugate-gradient type algorithms, modified so as to take only $m + 1$ steps before restarting, are identical to that derived in this paper.

The method has been particularly successful in treating constrained optimal control problems which are typically characterized as having a large number of variables (200 or so), only a few constraints, and often no possibility of directly obtaining second-order information. In addition, the method has been used to solve smaller mathematical programming problems. An example is given below.

For purposes of comparison, consider the problem of minimizing

$$f(x_1, x_2, \ldots, x_{10}) = \sum_{k=1}^{10} k x_k^2$$

Table 1

μ	Steps per cycle, p	Number of cycles to convergence	Number of steps	Value of modified objective function
10	1	90	90	388.565
	3	8	24	388.563
	5	3	15	388.563
	7	3	21	388.563
100	1	230 (*)	230	488.607
	3	21	63	487.446
	5	4	20	487.438
	7	2	14	487.433
1000	1	260 (*)	260	525.238
	3	45 (*)	135	503.550
	5	3	15	500.910
	7	3	21	500.882

(*) Program not run to convergence due to excessive time.

subject to

$$
\begin{aligned}
1.5x_1 + x_2 + x_3 + 0.5x_4 + 0.5x_5 &= 5.5,\\
2.0x_6 - 0.5x_7 - 0.5x_8 + x_9 - x_{10} &= 2.0,\\
x_1 + x_3 + x_5 + x_7 + x_9 &= 10.0,\\
x_2 + x_4 + x_6 + x_8 + x_{10} &= 15.0.
\end{aligned}
$$

This problem was treated by the penalty-function approach and the resulting composite functional was then solved for various values of μ by using various cycle lengths of a conjugate-gradient algorithm. In Table 1, p is the number of conjugate-gradient steps in a cycle. Thus, $p = 1$ corresponds to the ordinary steepest-descent method; $p = 5$ corresponds, by the theory of this paper, to the smallest value of p for which the rate of convergence is independent of μ; and $p = 10$ is the standard conjugate-gradient method. Note that, for $p < 5$, the convergence rate does indeed depend on μ, while it is more or less constant for $p \geqslant 5$. The values of μ selected are not artificially large since, for $\mu = 1000$, the constraints are satisfied only to within 0.5 %.

References

1. Courant, T., *Calculus of Variations and Supplementary Notes and Exercises*, Supplementary Notes by M. Kruskal and H. Rubin, Revised and Amended by J. Moser, New York University, New York, 1962.

2. Butler, T., and Martin, A. V., *On a Method of Courant for Minimizing Functionals*, Journal of Mathematics and Physics, Vol. 41, No. 1, 1962.
3. Luenberger, D. G., *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.
4. Fiacco, A. V., and McCormick, G. P., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968.
5. Zangwill, W. I., *Nonlinear Programming Via Penalty Functions*, Management Science, Vol. 13, No. 5, 1967.
6. Bryson, A. E., and Ho, Y. C., *Applied Optimal Control*, Blaisdell Publishing Company, Waltham, Massachusetts, 1969.
7. Goldstine, H. H., *Minimum Problems in the Functional Calculus*, Bulletin of the American Mathematical Society, Series 2, Vol. 46, No. 1, 1940.
8. Antosiewicz, H. A., and Rheinboldt, W. C., *Numerical Analysis and Functional Analysis*, Survey of Numerical Analysis, Edited by J. Todd, McGraw-Hill Book Company, New York, 1962.
9. Akaike, H., *On a Successive Transformation of Probability Distributions and its Application to the Analysis of the Optimum Gradient Method*, Annals of Mathematical Statistics, Vol. 11, No. 3, 1959–1960.
10. Fletcher, R., and Reeves, C. M., *Functional Minimization by Conjugate Gradients*, Computer Journal, Vol. 7, No. 2, 1964.
11. Shah, B. V., Buehler, R. J., and Kempthorne, O., *Some Algorithms for Minimizing a Function of Several Variables*, SIAM Journal on Applied Mathematics, Vol. 12, No. 1, 1964.
12. Daniel, J. W., *The Conjugate Gradient Method for Linear and Nonlinear Operator Equations*, Stanford University, Department of Mathematics, Ph.D. Thesis, 1965.
13. Hestenes, M. R., and Stiefel, E., *Method of Conjugate Gradients for Solving Linear Systems*, Journal of Research for the National Bureau of Standards, Vol. 49, No. 6, 1952.

PRINTED IN BELGIUM BY THE ST. CATHERINE PRESS, TEMPELHOF, 37, BRUGES, LTD.