ON THE CONVERGENCE OF A SEQUENTIAL PENALTY FUNCTION METHOD FOR CONSTRAINED MINIMIZATION*

NICHOLAS IAN MARK GOULD†

Abstract. The convergence behaviour of a class of iterative methods for solving the constrained minimization problem is analysed. The methods are based on the sequential minimization of a simple differentiable penalty function. They are sufficiently general to ensure global convergence of the iterates to the solution of the problem at an asymptotic (two-step Q-) superlinear rate.

Key words. constrained optimization, quadratic penalty function, augmented Lagrangian function, convergence analysis

AMS(MOS) subject classification. 65K05

1. Introduction. We are concerned with solving the constrained minimization problem

$$(1.1a) \qquad \text{minimize} \quad f(x)(x \in R^n)$$

subject to the constraints

$$(1.1b) \qquad c_i(x) = 0 \quad \text{for all } i \in \mathscr{E} = \{1, \cdots, t\},$$

where the functions f and c_i are assumed to be twice continuously differentiable functions of the variables x. We suppose that u_i, $i \in \mathscr{E}$ are given fixed (and finite) constants and that μ is a positive parameter. Then we are interested in solving (1.1) by means of a sequential minimization of the penalty function

$$(1.2) \qquad \Phi(x, \mu) = f(x) + \sum_{i \in \mathscr{E}} c_i(x) u_i + 1/2\mu \sum_{i \in \mathscr{E}} c_i(x)^2,$$

for various values of μ approaching zero. Specific instances of such a function are the quadratic penalty function (Fiacco and McCormick [8]) and the augmented Lagrangian function (Powell [19], Hestenes [15]) although the latter is often used in a different context where μ is bounded away from zero but the u_i are allowed to vary.

It is well known that, under relatively weak conditions,

$$(1.3) \qquad \lim_{\mu \to 0_+} \lim_{x \to x(\mu)} x = x^*,$$

where $x(\mu)$ and x^* are, respectively, local minimizers of (1.2) and local solutions of (1.1). Recent work (Broyden and Attia [2], Gould [14]) has shown that the often perceived difficulty of minimizing Φ as μ approaches zero can be avoided. In this paper, we analyse the convergence of a particular scheme for solving (1.1) by way of the function (1.2). The scheme produces iterates that are globally and, potentially (two-step Q-), superlinearly convergent. The global convergence is assured by ensuring that μ approaches zero. The fast asymptotic convergence is achieved by ensuring that a good starting point for the minimization of Φ for a fixed value of μ is available.

The paper is organised as follows. In § 2, we define our notation and a description of the algorithm framework follows in § 3. In § 4, we describe the assumptions we will use in our convergence proofs and draw any immediate conclusions. The bulk of our

† Computer Science and Systems Division, Harwell Laboratory, United Kingdom Atomic Energy Authority, Oxfordshire, United Kingdom OX11 0RA.

convergence proof is given in § 5 and we make a number of comments and suggestions in § 6 along with a report on the results of some simple numerical experiments.

2. Notation. We denote the gradient and Hessian matrix of the objective function by

$$g(x) = \nabla_x f(x) \quad \text{and} \quad G(x) = \nabla_{xx} f(x),$$

respectively. Similarly, we denote the gradient and Hessian matrix of the ith constraint function by

$$a_i(x) = \nabla_x c_i(x) \quad \text{and} \quad G_i(x) = \nabla_{xx} c_i(x),$$

respectively. We shall let $c(x)$ be the vector whose ith component is $c_i(x)$ and let $A(x)$ denote the Jacobian of $c(x)$. Thus

$$A^T(x) = (a_1(x), \cdots, a_t(x)).$$

For any vector λ, we let

$$\bar{g}(x, \lambda) = g(x) + \sum_{i \in \mathscr{E}} \lambda_i a_i(x) = g(x) + A^T(x)\lambda, \quad \text{and}$$

$$\bar{G}(x, \lambda) = G(x) + \sum_{i \in \mathscr{E}} \lambda_i G_i(x)$$

be the gradient and Hessian matrix, respectively, of the Lagrangian function

$$\bar{f}(x, \lambda) = f(x) + \sum_{i \in \mathscr{E}} \lambda_i c_i(x).$$

We also define the augmented matrix

$$K(x, \lambda, \mu) = \begin{bmatrix} \bar{G}(x, \lambda) & A^T(x) \\ A(x) & -\mu I \end{bmatrix},$$

where I is the (appropriately dimensioned) identity matrix. We say that the matrix $K(x, \lambda, \mu)$ satisfies the *second-order condition* if K is nonsingular and has precisely t negative eigenvalues. Finally, we shall be especially concerned with the particular choice

$$\bar{\lambda} \equiv \bar{\lambda}(x, \mu) = u + c(x)/\mu$$

for λ. For such a choice of λ we have

(2.1) $$\nabla_x \Phi(x, \mu) = \bar{g}(x, \bar{\lambda}), \quad \text{and}$$

(2.2) $$\nabla_{xx} \Phi(x, \mu) = \Gamma(x, \bar{\lambda}, \mu), \quad \text{where}$$

$$\Gamma(x, \lambda, \mu) = \bar{G}(x, \lambda) + (1/\mu)A^T(x)A(x).$$

We shall say that a point x^* is a *Kuhn–Tucker (first-order stationary) point* for the problem (1.1) if there is a vector of Lagrange multipliers λ^* for which

(2.3) $$g(x^*) + A^T(x^*)\lambda^* = 0 \quad \text{and} \quad c(x^*) = 0.$$

Any Kuhn–Tucker point for which $K(x^*, \lambda^*, 0)$ satisfies the second-order condition is necessarily an isolated local solution to the problem (1.1) (see, e.g., Gould [13]). Last, we define the (right) generalized inverse of any full rank matrix B by

$$B^+ = B^T(BB^T)^{-1}.$$

We are concerned with the asymptotic behaviour of sequences of numbers and need to be able to compare the relative magnitudes of two such sequences. To this end, suppose that $\{a_k\}$ and $\{b_k\}$ are two sequences of numbers converging to zero. Then we say the following:

(i) $\{a_k\}$ is *asymptotically smaller* than $\{b_k\}$ if

$$\lim_{k \to \infty} |a_k/b_k| = 0$$

and write $a_k = o(b_k)$ as k tends to infinity.

(ii) $\{a_k\}$ is *asymptotically no larger* than $\{b_k\}$ if

$$|a_k| \leq c_1|b_k|,$$

for all $k \geq k_0$, and write $a_k = O(b_k)$ as k tends to infinity.

(iii) $\{a_k\}$ is *asymptotically similar* to $\{b_k\}$ if

$$c_2|b_k| \leq |a_k| \leq c_1|b_k|$$

for all $k \geq k_0$, and write $a_k = O_s(b_k)$ as k tends to infinity, for some constants c_1 and c_2 and integer k_0.

3. Algorithm. We propose to solve (1.1) by the following scheme:

Step 0 [Initialization and start of the outer iteration].
The starting point $x^{(0)}$ is given. The positive constants γ, τ, β_1, β_2, ε, $\mu^{(0)}$, and $\mu_{\min}$ are specified to satisfy $\beta_1 < 0.5$, $\beta_1 < \beta_2 < 1$, $\varepsilon \ll 1$, and $\mu_{\min} \ll 1$. Set $k = 0$ and $x^{(0,0)} = x^{(0)}$.

Step 1 [Inner iteration]. Here we find an approximate minimizer of $\Phi(x, \mu^{(k)})$ with the following iteration.

    Step 1.0 [Start of the inner iteration].
    Evaluate the problem functions, gradients, and Hessian matrices at $x^{(k,0)}$. Set $\bar{\lambda}^{(k,0)} = \bar{\lambda}(x^{(k,0)}, \mu^{(k)})$ and assemble $\bar{g}(x^{(k,0)}, \bar{\lambda}^{(k,0)})$, $\bar{G}(x^{(k,0)}, \bar{\lambda}^{(k,0)})$ and $K(x^{(k,0)}, \bar{\lambda}^{(k,0)}, \mu^{(k)})$. Set $l = 0$.

    Step 1.1 [Test for convergence of the inner iteration].
    If

(3.1)
$$\|\bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})\|_2 \leq \gamma\mu^{(k)},$$

    set

(3.2)
$$x^{*(k)} = x^{(k,l)} \quad \text{and} \quad \lambda^{*(k)} = \bar{\lambda}^{(k,l)}$$

    and go to Step 2.

    Step 1.2 [Compute a search direction].
    Find any vector $p^{(k,l)}$ satisfying

(3.3)
$$-\bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})^T p^{(k,l)} \geq \varepsilon\mu^{(k)}\|\bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})\|_2\|p^{(k,l)}\|_2.$$

    In particular, if $K(x^{(k,l)}, \bar{\lambda}^{(k,l)}, \mu^{(k)})$ satisfies the second-order condition, use the Newton direction $p^{(k,l)}$ for Φ, which may be computed from the augmented set of equations

(3.4)
$$\begin{bmatrix} \bar{G}(x^{(k,l)}, \bar{\lambda}^{(k,l)}) & A^T(x^{(k,l)}) \\ A(x^{(k,l)}) & -\mu^{(k)}I \end{bmatrix}\begin{bmatrix} p^{(k,l)} \\ r^{(k,l)} \end{bmatrix} = -\begin{bmatrix} \bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)}) \\ 0 \end{bmatrix}$$

    (see Gould [14]), whenever such a vector satisfies (3.3).

Step 1.3 [Compute a steplength].

Find a scalar $\alpha^{(k,l)}$ satisfying the Armijo condition (Armijo [1])

(3.5) $\Phi(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}, \mu^{(k)}) \leqq \Phi(x^{(k,l)}, \mu^{(k)}) + \beta_1 \alpha^{(k,l)} \bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})^T p^{(k,l)}$

and the Goldstein condition (Goldstein [11])

(3.6) $\bar{g}(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}, \bar{\lambda}(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}, \mu^{(k)}))^T p^{(k,l)} \geqq \beta_2 \bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})^T p^{(k,l)}$.

If $p^{(k,l)}$ is the Newton direction, always try a stepsize of unity as the first candidate for $\alpha^{(k,l)}$.

Step 1.4 [Compute the new iterate and evaluate problem functions].
Set

$$x^{(k,l+1)} = x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}.$$

Reset $l := l + 1$. Evaluate the problem functions, gradients, and Hessian matrices at $x^{(k,l)}$. Set $\bar{\lambda}^{(k,l)} = \bar{\lambda}(x^{(k,l)}, \mu^{(k)})$ and assemble $\bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})$, $\bar{G}(x^{(k,l)}, \bar{\lambda}^{(k,l)})$, and $K(x^{(k,l)}, \bar{\lambda}^{(k,l)}, \mu^{(k)})$. Go to Step 1.1.

Step 2 [Test for convergence of the outer iteration]. If $\mu^{(k)} < \mu_{\min}$, Stop with the approximations $x^{*(k)}$ to x^* and $\lambda^{*(k)}$ to λ^*. Otherwise, compute $\mu^{(k+1)}$ such that $0 < \mu^{(k+1)} < \mu^{(k)}$.

Step 3 [Find a starting point for the next sequential minimization]. If $K(x^{*(k)}, \lambda^{*(k)}, \mu^{(k)})$ satisfies the second-order condition, compute the solution $p^{(k)}$ to the equations

(3.7) $\begin{bmatrix} \bar{G}(x^{*(k)}, \lambda^{*(k)}) & A^T(x^{*(k)}) \\ A(x^{*(k)}) & -\mu^{(k)} I \end{bmatrix} \begin{bmatrix} p^{(k)} \\ r^{(k)} \end{bmatrix} = - \begin{bmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c(x^{*(k)}) - \mu^{(k+1)}(\lambda^{*(k)} - u) \end{bmatrix}$

and define

(3.8) $$x_a^{*(k)} = x^{*(k)} + p^{(k)}.$$

Otherwise, set $x_a^{*(k)} = x^{*(k)}$. Then, if

(3.9) $\|\bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}))\|_2 \leqq \max(\tau, \|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2)$,

set $x^{(k+1,0)} = x_a^{*(k)}$. Otherwise, set $x^{(k+1,0)} = x^{*(k)}$. Reset $k := k+1$ and go to Step 1.

[End of the algorithm].

Although much of this algorithm might be thought of as standard, a word of explanation of Step 3 is in order. The alternative starting point $x_a^{*(k)}$ is merely the result of a single iteration of Newton's method for a root of the nonlinear system of equations

(3.10) $g(x) + A^T(x)\lambda = 0$ and $c(x) - \mu^{(k+1)}(\lambda - u) = 0$

with a slightly perturbed Jacobian matrix—the bottom diagonal entry $\mu^{(k+1)}$ having been replaced by $\mu^{(k)}$—starting from the estimates $x^{*(k)}$ and $\lambda^{*(k)}$. The equations (3.10) give a stationary point for the function $\Phi(x, \mu^{(k+1)})$. The perturbed Jacobian is convenient in that it is already available from Step 1 of the algorithm. We might expect the alternative starting point to be superior to $x^{*(k)}$ when $\mu^{(k)}$ is small and indeed we will show that this is the case. The test (3.9) is merely to stop the alternative starting point from being used when, in the early stages of the algorithm, the gradient of the penalty function at $x_a^{*(k)}$ could be relatively large in comparison with that at $x^{*(k)}$.

Note that there is still considerable freedom in our algorithm. In particular, we have not indicated how the search direction $p^{(k,l)}$ in Step 1.2 should be chosen if the Newton direction is inappropriate, nor have we been specific as to the actual choice of the stepsize $\alpha^{(k,l)}$ in Step 1.3.

Although, in practice, we would expect to set $\mu_{\min}$ to some small positive constant (bearing in mind the asymptotic estimate $\|x^* - x^{*(k)}\| = O(\mu^{(k)})$, see, for example, Fletcher [10]), *here we shall be interested in the behaviour of the algorithm when* $\mu_{\min} = 0$, and we shall assume that this is so in the next three sections. Clearly, such an investigation is of both theoretical and practical interest. For although the limits of finite precision computer arithmetic prohibit us from observing in full any predicted asymptotic convergence results, in practice we benefit from a predicted rate when we are close to the solution. We discuss the limitations of our results in § 6.

4. General assumptions and consequences. We make the following assumptions about the problem functions and the iterates generated by the algorithm.

> *Assumption* 1 (AS.1)   The iterates generated by the algorithm all lie within a bounded domain Ω.

Remark. This condition is needed to ensure that inner iteration converges. It is entirely possible that the function Φ is unbounded from below for any value of the penalty parameter and (AS.1) is designed to prevent this possibility causing a problem within the algorithm. We consider this assumption to be rather strong but are somewhat placated in that it appears in almost every convergence proof for constrained optimization algorithms that we have encountered.

Under (AS.1), the line search assumptions (3.5) and (3.6), and the search direction condition (3.3), we must satisfy the inner iteration convergence criterion (3.1) in a finite number of iterations (see Theorem 6.3.3 of Dennis and Schnabel [7], where we may replace the assumption of local Lipschitz continuity of $\bar{g}$ in their proof by our stronger assumption of continuity of the second derivatives of Φ within the bounded region Ω).

> *Assumption* 2 (AS.2)   The sequence $\{\mu^{(k)}\}$ converges to zero as k tends to infinity.

Under (AS.2) we have the following important result, which is based on Theorem 12.1.2 of Fletcher [10].

THEOREM 4.1. *Suppose that* (AS.1) *and* (AS.2) *hold. Let* x^* *be any limit point of the sequence* $\{x^{*(k)}\}$ *generated by the algorithm of* § 3 *as k tends to infinity and suppose that* $A(x^*)$ *is of full row rank. Then we have the following*:

(i) x^* *is a Kuhn–Tucker (first-order stationary) point for problem* (1.1) *and the corresponding elements of the sequence* $\{\lambda^{*(k)}\}$ *converge to the Lagrange multipliers* λ^*.

(ii) *For those indices k for which* $\{x^{*(k)}\}$ *converges to* x^*, *we have the following estimates*:

$$(4.1) \qquad\qquad \lambda^{*(k)} = \lambda^* + o(1),$$

$$(4.2) \qquad\qquad c(x^{*(k)}) = \mu^{(k)}(\lambda^* - u) + o(\mu^{(k)})$$

as $\mu^{(k)} \to 0_+$.

Proof. We consider only those indices k for which a particular subsequence $\{x^{*(k)}\}$ converges to x^*. As $A(x^*)$ is of full rank, we may define

$$\lambda^* = -A(x^*)^{+T} g(x^*).$$

Furthermore, for k sufficiently large, $A(x^{*(k)})^+$ exists, is bounded, and converges to $A(x^*)^+$. From (3.1) and (3.2), we have that

$$(4.3) \qquad \|g(x^{*(k)}) + A(x^{*(k)})^T \lambda^{*(k)}\|_2 = \|\bar{g}(x^{*(k)}, \lambda^{*(k)})\|_2 \leq \gamma \mu^{(k)}.$$

Thus, we may deduce that

$$\|A(x^{*(k)})^{+T}g(x^{*(k)})+\lambda^{*(k)}\|_2 = \|A(x^{*(k)})^{+T}(g(x^{*(k)})+A(x^{*(k)})^T\lambda^{*(k)})\|_2$$

(4.4)

$$\leqq \gamma\mu^{(k)}\|A(x^{*(k)})^{+T}\|_2.$$

But then we may combine the identity

$$\lambda^{*(k)}-\lambda^* = (A(x^{*(k)})^{+T}g(x^{*(k)})+\lambda^{*(k)})+(A(x^*)^{+T}g(x^*)-A(x^{*(k)})^{+T}g(x^{*(k)}))$$

with (4.4) to obtain the bound

(4.5) $\|\lambda^{*(k)}-\lambda^*\|_2 \leqq \gamma\mu^{(k)}\|A(x^{*(k)})^{+T}\|_2 + \|A(x^*)^{+T}g(x^*)-A(x^{*(k)})^{+T}g(x^{*(k)})\|_2.$

Thus, as the right-hand side of (4.5) can be made arbitrarily close to zero by picking k large enough, $\lambda^{*(k)}$ is bounded for k sufficiently large and converges to λ^*. Then, taking the limit of (4.3) as k approaches infinity, we may deduce that

(4.6) $g(x^*)+A^T(x^*)\lambda^* = 0.$

Furthermore, multiplying (4.5) by $\mu^{(k)}$, we obtain the additional bound

$$\|c(x^{*(k)})-\mu^{(k)}(\lambda^*-u)\|_2 \leqq \gamma\mu^{(k)2}\|A(x^{*(k)})^{+T}\|_2$$

(4.7)

$$+ \mu^{(k)}\|A(x^*)^{+T}g(x^*)-A(x^{*(k)})^{+T}g(x^{*(k)})\|_2.$$

Again, taking the limit of (4.7) as k approaches infinity, we have that

(4.8) $c(x^*) = 0.$

Hence, (4.6) and (4.8) imply that x^* is a Kuhn–Tucker point and the (sub)sequence $\{\lambda^{*(k)}\}$ converges to the relevant vector of Lagrange multipliers. The asymptotic estimates (4.1) and (4.2) may be deduced from (4.5) and (4.7), respectively.   □

Remark. Theorem 4.1 remains true if the test (3.1) is replaced by the weaker test

$$\|\bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})\|_2 \leqq \varepsilon(\mu^{(k)}),$$

for any positive function $\varepsilon(\mu)$ which converges to zero as μ approaches zero.

Assumption 3 (AS.3)   The augmented matrix $K(x^*, \lambda^*, 0)$ satisfies the second-order condition at any limit point x^* of the sequence $\{x^{*(k)}\}$.

Remark. This is a sufficient condition for a constrained stationary point to be a solution of (1.1) (see Gould [13]). Note, in particular, that (AS.3) implies that $A(x^*)$ is of full row rank and therefore that the conclusions of Theorem 4.1 are valid.

Assumption 4 (AS.4)   The third derivatives of the functions $f(x)$ and $c_i(x)$ exist and are bounded for all points in Ω.

Remark. We may now extend Theorem 4.1 to the following result.

THEOREM 4.2. *Suppose that* (AS.1)-(AS.4) *hold. Let x^* be any limit point of the sequence $\{x^{*(k)}\}$ generated by the algorithm of §3 as k tends to infinity. Then the conclusions of Theorem 4.1 remain true and, for the appropriate subsequence of $\{x^{*(k)}\}$, we have the further estimates:*

(4.9) $x^{*(k)} = x^* + O(\mu^{(k)}),$

(4.10) $\lambda^{*(k)} = \lambda^* + O(\mu^{(k)}),$ *and*

(4.11) $c(x^{*(k)}) = \mu^{(k)}(\lambda^*-u) + O(\mu^{(k)2})$

as $\mu^{(k)} \to 0_+$.

Proof. Again, we consider only those indices k for which a particular subsequence $\{x^{*(k)}\}$ converges to x^*. The conclusions of Theorem 4.1 follow from (AS.1)-(AS.2).

To obtain the estimates (4.9) and (4.10), we use the relationships

$$g(x^{*(k)}) + A(x^{*(k)})^T \lambda^{*(k)} = \bar{G}(x^*, \lambda^*)(x^{*(k)} - x^*) + A^T(x^*)(\lambda^{*(k)} - \lambda^*)$$

(4.12)
$$+ O(\max(\|x^{*(k)} - x^*\|_2^2, \|x^{*(k)} - x^*\|_2 \|\lambda^{*(k)} - \lambda^*\|_2))$$

and

(4.13)
$$c(x^{*(k)}) = A(x^*)(x^{*(k)} - x^*) + O(\|x^{*(k)} - x^*\|_2^2).$$

Equations (4.12) and (4.13) follow from the relevant Taylor series expansions of their left-hand sides, using (AS.4), and from the relationships (2.3). Rearranging these equations, we obtain

(4.14)
$$\begin{bmatrix} \bar{G}(x^*, \lambda^*) & A^T(x^*) \\ A(x^*) & 0 \end{bmatrix} \begin{bmatrix} x^{*(k)} - x^* \\ \lambda^{*(k)} - \lambda^* \end{bmatrix} = \begin{bmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c(x^{*(k)}) \end{bmatrix}$$
$$+ O(\max(\|x^{*(k)} - x^*\|_2^2, \|x^{*(k)} - x^*\|_2 \|\lambda^{*(k)} - \lambda^*\|_2)).$$

The vector on the right-hand side of (4.14) is $O(\mu^{(k)})$, from (3.1) and (4.2). As the coefficient matrix $K(x^*, \lambda^*, 0)$ of (4.14) is nonsingular (see (AS.3)), the estimate (4.1) combines with (4.14) to give (4.9) and (4.10). The estimate (4.10) then implies (4.11). □

  Remark. In fact, Theorem 4.2 remains true if (AS.3) is replaced by the weaker assumption that $K(x^*, \lambda^*, 0)$ is nonsingular.

  5. Convergence results. In this section, we are concerned with the speed at which the algorithm of § 3 converges—to be more specific, with the speed at which the iterates converge to a limit point. Essentially two factors affect the rate of convergence. First, because of the estimate (4.9), the rate of convergence of the outer iteration is determined by the rate at which the sequence $\{\mu^{(k)}\}$ is allowed to converge to zero. For instance, a Q- or R-superlinear rate of convergence (in the sense of Ortega and Rheinboldt [18]) of $\{\mu^{(k)}\}$ implies at least as fast an R-rate of convergence for the sequence $\{x^{*(k)}\}$. Second, even if the outer iteration possesses a fast rate of convergence, the overall speed of the algorithm will be severely restricted if each outer iteration requires many inner iterations. In this section, we show that, under very mild restrictions on the sequence $\{\mu^{(k)}\}$, the inner iteration will cost at most two problem function evaluations for large enough k and that an overall two-step Q-superlinear asymptotic rate of convergence can be achieved. We do this by obtaining the following results:

  (i) The point $x_a^{*(k)}$ of (3.8) is a "better" starting point for the $(k+1)$st iteration than $x^{*(k)}$ if k is large enough.

  (ii) From this starting point, the Newton iteration with unit stepsize satisfies the linesearch conditions (3.3), (3.5), and (3.6) at every inner iteration provided that k is large enough.

  (iii) The starting point is sufficiently close to a minimizer of $\Phi(x, \mu^{(k)})$ that the first inner iterate satisfies the inner iteration convergence test (3.1) when k is large enough.

  In order to prove such results, we must restrict our attention to sequences $\{\mu^{(k)}\}$ that satisfy the further assumptions:

  Assumption 5 (AS.5)   The sequence $\{\mu^{(k)}\}$ is such that $\mu^{(k+1)} \leqq \sigma^{(k)} \mu^{(k)}$ where $\lim_{k \to \infty} \sigma^{(k)} = \sigma < 1$.

  Assumption 6 (AS.6)   The sequence $\{\mu^{(k)}\}$ is such that

  $$\mu^{(k)2} / \mu^{(k+1)} = o(1)$$

  as k approaches infinity.

Remark. Assumption (AS.5) essentially says that $\{\mu^{(k)}\}$ converges to zero at least linearly and implies (AS.2). Assumption (AS.6) restricts the sequence to have a Q-order of convergence of anything less than two. Both assumptions thus restrict the overall rate of convergence of the algorithm.

As we shall be concerned with convergence to a limit point of the sequence $\{x^{*(k)}\}$, we will start by restricting our attention to an infinite subsequence of iterates $\{x^{*(k)}\}$, $k \in \mathcal{K}$, whose limit point is x^*. We caution the reader that we often use qualifiers such as "for all k" and "as k approaches infinity" when we strictly mean "for all $k \in \mathcal{K}$" and "as $k \in \mathcal{K}$ approaches infinity." Our assumptions are sufficiently strong that we will ultimately be able to show that the algorithm actually has only a single limit point. We start by obtaining the following theorem.

THEOREM 5.1. *Suppose that* (AS.1)-(AS.5) *hold, that* $t \geqq 1$, *and that* $u \neq \lambda^*$. *Then we have the following estimates:*

$$(5.1) \qquad \bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) = O_s(\mu^{(k)}/\mu^{(k+1)}),$$

$$(5.2) \qquad \bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) = O(\mu^{(k)2}/\mu^{(k+1)})$$

as $k \in \mathcal{K}$ *tends to infinity.*

Proof. To verify (5.1), first we have that the estimate (4.2) yields

$$
\begin{aligned}
(5.3) \qquad \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)} &= c(x^{*(k)})(1/\mu^{(k+1)} - 1/\mu^{(k)}) \\
&= (\mu^{(k)}/\mu^{(k+1)} - 1)(\lambda^* - u) + o(\mu^{(k)}/\mu^{(k+1)})
\end{aligned}
$$

as k tends to infinity. From (AS.5), we have that

$$(5.4) \qquad \tfrac{1}{2}(1-\sigma)\mu^{(k)}/\mu^{(k+1)} \leqq |\mu^{(k)}/\mu^{(k+1)} - 1| \leqq \mu^{(k)}/\mu^{(k+1)}$$

for all large k. Therefore, combining (5.3) and (5.4), we have

$$
\begin{aligned}
(5.5) \qquad (\tfrac{1}{2}(1-\sigma)(1-\varepsilon_1)\|\lambda^* - u\|_2)\mu^{(k)}/\mu^{(k+1)} &\leqq \|\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}\|_2 \\
&\leqq ((1+\varepsilon_1)\|\lambda^* - u\|_2)\mu^{(k)}/\mu^{(k+1)}
\end{aligned}
$$

for all k sufficiently large—the terms $(1-\varepsilon_1)$ and $(1+\varepsilon_1)$ $(0 < \varepsilon_1 \ll 1)$ accounting for the asymptotically smaller terms in (5.3). Now (4.3) and (5.5) give

$$
\begin{aligned}
(5.6) \qquad \bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) &= \bar{g}(x^{*(k)}, \lambda^{*(k)}) + A^T(x^{*(k)})(\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) \\
&= A^T(x^{*(k)})(\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) + O(\mu^{(k)}) \\
&= A^T(x^{*(k)})(\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) + o(\mu^{(k)}/\mu^{(k+1)}).
\end{aligned}
$$

Then (5.5), (5.6), and the continuity of $A(x)$ give the bound

$$(5.7) \quad \|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2 \leqq (2(1+\varepsilon_1)(1+\varepsilon_2)\|A^T(x^*)\|_2\|\lambda^* - u\|_2)\mu^{(k)}/\mu^{(k+1)}$$

for all k sufficiently large; the term $(1+\varepsilon_2)$ $(0 < \varepsilon_2 \ll 1)$ accounts for the asymptotically smaller terms in (5.6) and the constant two occurs because of the bound $\|A^T(x^{*(k)})\|_2 \leqq 2\|A^T(x^*)\|_2$. Premultiplying (5.6) by $A(x^{*(k)})^{+T}$ gives

$$\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)} = A(x^{*(k)})^{+T}\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) + o(\mu^{(k)}/\mu^{(k+1)});$$

this combines with (5.5) and the continuity of $A(x)^{+T}$ in some neighbourhood of x^* to give the bound

$$(5.8) \quad \|\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}\|_2 \leqq 2(1+\varepsilon_2)\|A(x^*)^{+T}\|_2\|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2$$

for all k sufficiently large—the term $(1+\varepsilon_2)$ once again accounts for the asymptotically smaller term in (5.7). Inequalities (5.5) and (5.8) combine to give the bound

$$(5.9) \quad \begin{aligned} (\tfrac{1}{4}(1-\sigma)(1-\varepsilon_1)\|\lambda^*-u\|_2/(1+\varepsilon_2)\|A(x^*)^{+T}\|_2)\mu^{(k)}/\mu^{(k+1)} \\ \leqq \|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2 \end{aligned}$$

for large k. The bounds (5.7) and (5.9) then imply (5.1).

To obtain the estimate (5.2), first observe that the coefficient matrix $K(x^{*(k)}, \lambda^{*(k)}, \mu^{(k)})$ of (3.7) satisfies the second-order condition (and hence is nonsingular) for large enough k from assumption (AS.3) and Theorem 4.2. Hence $x_a^{*(k)}$ is defined by (3.8). Now define

$$(5.10) \quad \lambda_a^{*(k)} = \lambda^{*(k)} + r^{(k)},$$

where $r^{(k)}$ is given by (3.7). Then, a straightforward Taylor's expansion and (3.7) give that

$$(5.11) \quad \begin{aligned} \begin{bmatrix} \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) \\ c(x_a^{*(k)}) - \mu^{(k+1)}(\lambda_a^{*(k)}-u) \end{bmatrix} &= \begin{bmatrix} \bar{G}(x^{*(k)}, \lambda^{*(k)}) & A^T(x^{*(k)}) \\ A(x^{*(k)}) & -\mu^{(k+1)}I \end{bmatrix} \begin{bmatrix} p^{(k)} \\ r^{(k)} \end{bmatrix} \\ &\quad + \begin{bmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c(x^{*(k)}) - \mu^{(k+1)}(\lambda^{*(k)}-u) \end{bmatrix} \\ &\quad + O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2) \\ &= \begin{bmatrix} 0 \\ (\mu^{(k)}-\mu^{(k+1)})r^{(k)} \end{bmatrix} + O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2) \\ &= O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2) + O(\mu^{(k)}\|r^{(k)}\|_2). \end{aligned}$$

Moreover, equations (3.1), (4.1), and (4.2) ensure that the right-hand side of (3.7) is $O(\mu^{(k)})$. Thus $\|p^{(k)}\|_2 = O(\mu^{(k)}) = \|r^{(k)}\|_2$ and (5.11) yields

$$(5.12) \quad \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) = O(\mu^{(k)2}), \quad \text{and}$$

$$(5.13) \quad c(x_a^{*(k)}) - \mu^{(k+1)}(\lambda_a^{*(k)}-u) = O(\mu^{(k)2}).$$

But then, (5.13) and the definition of $\bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})$ give

$$\mu^{(k+1)}(\bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}) = c(x_a^{*(k)}) - \mu^{(k+1)}(\lambda_a^{*(k)}-u) = O(\mu^{(k)2}),$$

and hence

$$(5.14) \quad \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)} = O(\mu^{(k)2}/\mu^{(k+1)}).$$

Equations (5.12) and (5.14) then combine to give

$$\begin{aligned} \bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) &= \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) + A^T(x_a^{*(k)})(\bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}) \\ &= O(\mu^{(k)2}/\mu^{(k+1)}), \end{aligned}$$

which establishes (5.2). $\quad \square$

Theorem 5.1 shows that the gradient of the penalty function at the alternative starting point $x_a^{*(k)}$ is asymptotically smaller than at $x^{*(k)}$. In fact, if $\{\mu^{(k)}\}$ is chosen as a superlinearly convergent sequence, the gradient at $x^{*(k)}$ will become arbitrarily large as k approaches infinity. Under (AS.6), quite the opposite will happen to the gradient at $x_a^{*(k)}$, which asymptotically converges to zero. There is therefore good reason to prefer the alternative starting point.

If, by design or chance, the parameters u are chosen to be the Lagrange multipliers at the solution, we have the following theorem.

THEOREM 5.2. *Suppose that* (AS.1)–(AS.5) *hold, that* $t \geqq 1$, *and that* $u = \lambda^*$. *Then the estimates*

(5.15) $$\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) = O(\mu^{(k)2}/\mu^{(k+1)})$$

and (5.2) *hold as* $k \in \mathcal{K}$ *tends to infinity.*

Proof. The estimate (4.11) ensures that $c(x^{*(k)}) = O(\mu^{(k)2})$. This then implies that

(5.16) $$\bar{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)} = c(x^{*(k)})(1/\mu^{(k+1)} - 1/\mu^{(k)}) = O(\mu^{(k)2}/\mu^{(k+1)}).$$

Equation (5.6) is still valid and combines with (5.16) to give (5.15). The estimate (5.2) is obtained exactly as in the proof of Theorem 5.1. □

In the unlikely event we can choose $u = \lambda^*$, we might feel that (5.15) does not indicate starting the $(k+1)$st inner iteration from the alternative point as the gradient at $x^{*(k)}$ is already very small; indeed, it is not clear whether (5.2) or (5.15) yields the smaller gradient. There is, however, a more subtle reason for preferring $x_a^{*(k)}$—not only is the gradient of the penalty function at this point small, but the point is in a very definite sense actually closer to $x(\mu^{(k+1)})$ than $x^{*(k)}$ is. We discuss this further after a slight digression but conclude that, in our algorithmic framework, we have the following corollary.

COROLLARY 5.3. *Suppose that* (AS.1)–(AS.6) *hold, and that* $t \geqq 1$. *Then the* $(k+1)$st *inner iteration will be started at* $x_a^{*(k)}$, *as defined by* (3.8), *rather than* $x^{*(k)}$ *for all* $k \in \mathcal{K}$ *sufficiently large.*

Proof. This follows immediately from the test (3.9) in the algorithm in conjunction with the estimate (5.2), which holds regardless of the value of u, and (AS.6). □

When there are no constraints we have the following more conclusive result.

THEOREM 5.4. *Suppose that* (AS.1)–(AS.6) *hold and that* $t = 0$. *Then* $x^{(k+1,0)} = x_a^{*(k)}$ *and satisfies the inner iteration convergence test* (3.1) *for all* $k \in \mathcal{K}$ *sufficiently large.*

Proof. If there are no constraints, $\bar{g}(x, \lambda) = g(x)$ and $\bar{G}(x, \lambda) = G(x) = K(x, \lambda, \mu)$. The matrix $K(x^{*(k)}, \lambda^{*(k)}, \mu^{(k)})$ satisfies the second-order condition for all large k, from (AS.3) and the estimates (4.9) and (4.10), and $x^{*(k)}$ will then be computed from (3.8). The exact reasoning that allowed us to derive (5.12) gives the estimate

$$g(x_a^{*(k)}) \equiv \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) = O(\mu^{(k)2})$$

for large k. Therefore $x_a^{*(k)}$ satisfies the test (3.9) and $x^{(k+1,0)} = x_a^{*(k)}$. But then we may invoke (AS.6) to obtain

$$\|g(x^{(k+1,0)})\|_2 = O(\mu^{(k)2}) = o(\mu^{(k+1)}).$$

This implies that (3.1) will be satisfied at $x^{(k+1,0)}$ for all k sufficiently large. □

We see from Corollary 5.3 that if constraints are present in the problem, the inner iterations will ultimately all be started from the alternative starting point $x_a^{*(k)}$. The first inner iteration of the $(k+1)$st outer iteration involves the computation of a search direction $p^{(k+1,0)}$ from (3.4). Considering this equation in combination with the estimate $\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}) = O(\mu^{(k)2}/\mu^{(k+1)})$ from Theorem 5.1 might lead us to suspect that $p^{(k+1,0)}$ is of the same order. In fact, $p^{(k+1,0)}$ is actually smaller. For we have the following lemma.

LEMMA 5.5. *Suppose that* (AS.1)–(AS.6) *hold and that* $t \geqq 0$. *Then we have the following*:

   (i) $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ *satisfies the second-order condition.*
   (ii) *If* $p^{(k+1,0)}$ *is found from* (3.4), *we have the estimate*

(5.17) $$p^{(k+1,0)} = O(\mu^{(k)2})$$

for all $k \in \mathcal{K}$ *sufficiently large.*

Proof. Corollary 5.3 gives that $x^{(k+1,0)} = x_a^{*(k)}$. As the right-hand side of (3.7) is $O(\mu^{(k)})$, (3.7), (3.8), and (4.9) imply that

$$x^{(k+1,0)} = x^* + O(\mu^{(k)}) \tag{5.18}$$

as k approaches infinity. Furthermore, the estimates (4.1), (5.10), and (5.14) (under (AS.6)) show that

$$\bar{\lambda}^{(k+1,0)} = \lambda^* + o(1) \tag{5.19}$$

as k approaches infinity. Thus $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ converges to $K(x^*, \lambda^*, 0)$ and, by continuity, satisfies the second-order condition for all k sufficiently large.

Now, we have that

$$
\begin{aligned}
\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}) &= \bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) \\
&= \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) + A^T(x^{(k+1,0)})(\bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}).
\end{aligned}
\tag{5.20}
$$

If we substitute (5.20) into the right-hand side of (3.4) and then rearrange terms, we obtain that $p^{(k+1,0)}$ may be computed from the alternative set of augmented equations

$$
\begin{aligned}
&\begin{bmatrix} \bar{G}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}) & A^T(x^{(k+1,0)}) \\ A(x^{(k+1,0)}) & -\mu^{(k+1)}I \end{bmatrix} \begin{bmatrix} p^{(k+1,0)} \\ s^{(k+1,0)} \end{bmatrix} \\
&= -\begin{bmatrix} \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) \\ \mu^{(k+1)}(\bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}) \end{bmatrix}.
\end{aligned}
\tag{5.21}
$$

Equations (5.12) and (5.14) ensure that the right-hand side of (5.21) is $O(\mu^{(k)2})$. The nonsingularity of $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ for large k (and the boundedness of its inverse) and (5.21) then imply the estimate (5.17). $\square$

We will also need to be able to bound the condition number of the Hessian matrix of Φ. To this end, we have the following lemma.

LEMMA 5.6 (Murray [17]). *Suppose that $t \geq 1$, that the matrices $\bar{G}(x, \lambda)$ and $A(x)$ of §2 are bounded in some domain Λ, and that $A(x)$ is of full row rank there. Then, for all $(x, \lambda) \in \Lambda$, we have the following:*

(i) *If $t < n$, the smallest eigenvalue, $\sigma_{\min}(x, \lambda, \mu)$, of the matrix $\Gamma(x, \lambda, \mu) \equiv \bar{G}(x, \lambda) + (1/\mu)A^T(x)A(x)$ satisfies*

$$\sigma_{\min}(x, \lambda, \mu) = \sigma_{\min}(x, \lambda) + o(1),$$

where $\sigma_{\min}(x, \lambda)$ is the smallest eigenvalue of the matrix $Z^T(x)\bar{G}(x, \lambda)Z(x)$ and where $Z(x)$ is any matrix satisfying $A(x)Z(x) = 0$ and $Z^T(x)Z(x) = I$.

(ii) *If $t = n$, the smallest eigenvalue of $\Gamma(x, \lambda, \mu)$ satisfies*

$$\sigma_{\min}(x, \lambda, \mu) = \sigma_{\min}(x)/\mu + o(1/\mu),$$

where $\sigma_{\min}(x)$ is the smallest eigenvalue of the matrix $A^T(x)A(x)$.

(iii) *The largest eigenvalue, $\sigma_{\max}(x, \lambda, \mu)$, of $\Gamma(x, \lambda, \mu)$ satisfies*

$$\sigma_{\max}(x, \lambda, \mu) = \sigma_{\max}(x)/\mu + o(1/\mu),$$

where $\sigma_{\max}(x)$ is the largest eigenvalue of the matrix $A^T(x)A(x)$, as μ approaches zero from above.

It is known that the matrix $Z(x)$ may not be a continuous function of x (see, Coleman and Sorensen [6] and Gill et al. [12]). However, Byrd and Schnabel [5, Thm. 2.3] show how to construct a matrix $Z(x)$ that is continuous in some open neighbourhood of x^*. For this we require that $A(x^*)$ has a positive smallest singular value (see (AS.3)) and that $A(x)$ is continuous in a neighbourhood of x^*. With such a choice of $Z(x)$, Lemma 5.6 leads to Lemma 5.7.

LEMMA 5.7. *Suppose that* (AS.1)–(AS.6) *hold and that* $t \geqq 1$. *Then* (*using the notation of Lemma 5.6*), *for all* $k \in \mathcal{H}$ *sufficiently large, we have the following*:

(i) *The smallest eigenvalue of* $\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ *satisfies*

$$(5.22) \qquad 0 < \tfrac{1}{2}\sigma_{\min} < \sigma_{\min}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$$

where $\sigma_{\min}$ *is the smallest eigenvalue of the matrix* $Z^T(x^*)\bar{G}(x^*, \lambda^*)Z(x^*)$ *if* $t < n$ *or the smallest eigenvalue of* $A^T(x^*)A(x^*)$ *otherwise.*

(ii) *The largest eigenvalue of* $\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ *satisfies*

$$(5.23) \qquad \sigma_{\max}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)}) < 2\sigma_{\max}/\mu^{(k+1)}$$

where $\sigma_{\max} > 0$ *is the largest eigenvalue of the matrix* $A^T(x^*)A(x^*)$.

Proof. The results follow immediately from Lemma 5.6, the estimates (5.18) and (5.19), the (local) continuity of $\bar{G}(x, \lambda)$, $A(x)$, and $Z(x)$, and the consequent continuity of the eigenvalues of the symmetric matrices $Z^T(x)\bar{G}(x, \lambda)Z(x)$ and $A^T(x)A(x)$—the constants $\tfrac{1}{2}$ and 2 being arbitrary numbers smaller than and larger than one, respectively. The eigenvalue $\sigma_{\min}$ is positive from (AS.3) (see, Gould [14]). □

The results of Lemmas 5.5 and 5.7 are of fundamental importance, for they allow us to prove Theorem 5.8.

THEOREM 5.8. *Suppose that* (AS.1)–(AS.6) *hold and that* $t \geqq 1$. *Then, for all* $k \in \mathcal{H}$ *sufficiently large, we have the following*:

(i) *The Newton direction* $p^{(k+1,0)}$ *obtained from equation* (3.4) *satisfies condition* (3.3).

(ii) *A stepsize* $\alpha^{(k+1,0)} = 1$, *used in conjunction with the Newton direction, satisfies the linesearch conditions* (3.5) *and* (3.6).

Proof. The results are trivial if $p^{(k+1,0)} = 0$, and so we only consider the case $p^{(k+1,0)} \neq 0$.

(i) In our algorithm, the Newton direction is computed from the system of equations (3.4). However, this direction also satisfies the (potentially badly-conditioned) equations

$$(5.24) \qquad \Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)} = -\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}).$$

This equation (5.24) is, of course, merely

$$(5.25) \qquad \nabla_{xx}\Phi(x^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)} = -\nabla_x\Phi(x^{(k+1,0)}, \mu^{(k+1)}).$$

It follows from (5.24) that

$$(5.26) \qquad -\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} = p^{(k+1,0)T}\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)}.$$

Lemma 5.5(i) ensures that $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ satisfies the second-order condition for all k sufficiently large and thus that $\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ is positive definite (Gould [14]). Therefore (5.26) gives that

$$(5.27) \qquad -\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)}/\|\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T\|_2\|p^{(k+1,0)}\|_2 > 0$$

for all k sufficiently large. Moreover, we have from (5.24) that

$$
\begin{aligned}
(\bar{g}^{(k+1,0)T}&p^{(k+1,0)})^2/\|\bar{g}^{(k+1,0)}\|_2^2\|p^{(k+1,0)}\|_2^2 \\
&= (p^{(k+1,0)T}\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)}/p^{(k+1,0)T}p^{(k+1,0)}) \\
&\quad \cdot (\bar{g}^{(k+1,0)T}\Gamma(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})^{-1}\bar{g}^{(k+1,0)}/\bar{g}^{(k+1,0)T}\bar{g}^{(k+1,0)}) \\
&> \sigma_{\min}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})/\sigma_{\max}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)}),
\end{aligned}
$$

(5.28)

where we have written $\bar{g}^{(k+1,0)}$ for $\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})$. Therefore, for all k sufficiently large, (5.22), (5.23), (5.27), and (5.28) combine to give

(5.29)
$$-\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} / \|\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})\|_2 \|p^{(k+1,0)}\|_2$$
$$> \tfrac{1}{2}\sqrt{(\sigma_{\min}/\sigma_{\max})\mu^{(k+1)}}.$$

If we choose k large enough, we must have that

(5.30)
$$\tfrac{1}{2}\sqrt{(\sigma_{\min}/\sigma_{\max})\mu^{(k+1)}} > \varepsilon\mu^{(k+1)}.$$

Inequalities (5.29) and (5.30) then imply that (3.3) is satisfied for all k sufficiently large.

(ii) To verify (3.5), observe that a Taylor series expansion, (2.1), and (5.24) give

$$\Phi(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)})$$
$$= \Phi(x^{(k+1,0)}, \mu^{(k+1)}) + \nabla_x \Phi(x^{(k+1,0)}, \mu^{(k+1)})^T p^{(k+1,0)}$$
$$+ \tfrac{1}{2}p^{(k+1,0)T} \nabla_{xx} \Phi(x^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)} + O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)})$$
$$= \Phi(x^{(k+1,0)}, \mu^{(k+1)}) + \tfrac{1}{2}\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} + O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)})$$

under (AS.4). Then

(5.31)
$$\Phi(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}) - \Phi(x^{(k+1,0)}, \mu^{(k+1)}) - \beta_1 \bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)}$$
$$= (\tfrac{1}{2} - \beta_1)\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} + O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)}).$$

But, by assumption (AS.6) and the estimate (5.17),

$$\|p^{(k+1,0)}\|_2/\mu^{(k+1)} = O(\mu^{(k)2}/\mu^{(k+1)}) = o(1);$$

this implies that the second term on the right-hand side of (5.31) is $o(\|p^{(k+1,0)}\|_2^2)$. Furthermore, (5.22) and (5.25) give the inequality

(5.32)
$$\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} < -\sigma_{\min}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)T}p^{(k+1,0)}$$
$$< -\tfrac{1}{2}\sigma_{\min}\|p^{(k+1,0)}\|_2^2 < 0.$$

As $\beta_1 < 0.5$, (5.31) and (5.32) give

$$\Phi(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}) - \Phi(x^{(k+1,0)}, \mu^{(k+1)}) - \beta_1 \bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)}$$
$$< -\tfrac{1}{2}(\tfrac{1}{2} - \beta_1)(1 - \varepsilon_3)\sigma_{\min}\|p^{(k+1,0)}\|_2^2 < 0$$

for all large k; the term $(1 - \varepsilon_3)(0 < \varepsilon_3 \ll 1)$ accounts for the asymptotically smaller second term in (5.31), which implies that the Newton direction with a unit stepsize satisfies (3.5).

The inequality (3.6) is established in much the same fashion. A Taylor series expansion, (2.1), and (5.24) give

$$\bar{g}(x^{(k+1,0)} + p^{(k+1,0)}, \bar{\lambda}(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}))^T p^{(k+1,0)}$$
$$= \nabla_x \Phi(x^{(k+1,0)}, \mu^{(k+1)})^T p^{(k+1,0)} + p^{(k+1,0)T}\nabla_{xx}\Phi(x^{(k+1,0)}, \mu^{(k+1)})p^{(k+1,0)}$$
$$+ O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)}) = O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)})$$

under (AS.4). Then

(5.33)
$$\bar{g}(x^{(k+1,0)} + p^{(k+1,0)}, \bar{\lambda}(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}))^T p^{(k+1,0)}$$
$$- \beta_2 \bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)}$$
$$= -\beta_2 \bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)} + O(\|p^{(k+1,0)}\|_2^3/\mu^{(k+1)}).$$

Once again, the second term on the right-hand side of (5.33) is $o(\| p^{(k+1,0)} \|_2^2)$. As $\beta_2 > 0$, (5.32), and (5.33) give

$$\bar{g}(x^{(k+1,0)} + p^{(k+1,0)}, \bar{\lambda}(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}))^T p^{(k+1,0)} - \beta_2 \bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})^T p^{(k+1,0)}$$

$$> \beta_2(1 - \varepsilon_4)\sigma_{\min} \| p^{(k+1,0)} \|_2^2 > 0$$

for all large k—the term $(1 - \varepsilon_4)$ $(0 < \varepsilon_4 \ll 1)$ again accounts for the asymptotically smaller second term in (5.33), and hence the Newton direction with $\alpha^{(k+1,0)} = 1$ satisfies (3.6). $\quad\square$

We then have the immediate corollary.

COROLLARY 5.9. *Suppose that* (AS.1)-(AS.6) *hold and that* $t \geq 1$. *Then,*

$$x^{(k+1,1)} = x^{(k+1,0)} + p^{(k+1,0)},$$

with $p^{(k+1,0)}$ *being the Newton direction* (3.4), *for all* $k \in \mathcal{K}$ *sufficiently large.*

Proof. This follows directly from Lemma 5.5 (i) and Theorem 5.8. $\quad\square$

Having shown that a full Newton step is possible at the first inner iteration, we can now reap the benefits of such a step. Therefore, we show Lemma 5.10.

LEMMA 5.10. *Suppose that* (AS.1)-(AS.6) *hold and that* $t \geq 1$. *Then,*

$$(5.34) \qquad\qquad \bar{g}(x^{(k+1,1)}, \bar{\lambda}^{(k+1,1)}) = O(\mu^{(k)4}/\mu^{(k+1)})$$

for all $k \in \mathcal{K}$ *sufficiently large.*

Proof. A Taylor series expansion, (2.1), (5.17), (5.20), and (5.27) give

$$\bar{g}(x^{(k+1,1)}, \bar{\lambda}^{(k+1,1)}) = \bar{g}(x^{(k+1,0)} + p^{(k+1,0)}, \bar{\lambda}(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)}))$$

$$= \nabla_x \Phi(x^{(k+1,0)} + p^{(k+1,0)}, \mu^{(k+1)})$$

$$= \nabla_x \Phi(x^{(k+1,0)}, \mu^{(k+1)}) + \nabla_{xx} \Phi(x^{(k+1,0)}, \mu^{(k+1)}) p^{(k+1,0)}$$

$$+ O(\| p^{(k+1,0)} \|_2^2 / \mu^{(k+1)})$$

$$= O(\| p^{(k+1,0)} \|_2^2 / \mu^{(k+1)}) = O(\mu^{(k)4}/\mu^{(k+1)})$$

under (AS.4), which establishes (5.34). $\quad\square$

THEOREM 5.11. *Suppose that* (AS.1)-(AS.6) *hold. Then inequality* (3.1) *is satisfied with* $l \leq 1$ *for all* $k \in \mathcal{K}$ *sufficiently large.*

Proof. This follows immediately from Lemma 5.10 in conjunction with (AS.6), when $t \geq 1$, and from Theorem 5.4, when $t = 0$. $\quad\square$

If we say that a *step* of the algorithm is the computation that takes place between the start of one function/derivative evaluation and the start of the next, we deduce the following corollary.

COROLLARY 5.12. *Suppose that* (AS.1)-(AS.6) *hold and that the whole sequence of iterates* $\{x^{*(k)}\}$ *generated by the algorithm converges to* x^*. *Then the iterates converge at least two-step R-linearly if* $\{\mu^{(k)}\}$ *converges Q-linearly and at least two-step R-superlinearly if* $\{\mu^{(k)}\}$ *converges Q-superlinearly.*

It is interesting to note that the actual rate of superlinear convergence of the algorithm is entirely controlled by the sequence $\{\mu^{(k)}\}$, which may be assigned before the computation commences.

We conclude the section by showing that our assumptions actually imply that the sequence of iterates has a single limit point and that the R-rates of convergence referred to in Corollary 5.12 may be tightened to be Q-rates.

LEMMA 5.13. *Suppose that* (AS.1)-(AS.6) *hold. Then we have the estimate*

$$(5.35) \qquad\qquad x^{*(k+1)} = x^* + \mu^{(k+1)}v + o(\mu^{(k+1)}),$$

where the vector v satisfies the equation

$$(5.36) \qquad \begin{bmatrix} \bar{G}(x^*, \lambda^*) & A^T(x^*) \\ A(x^*) & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda^* - u \end{bmatrix}$$

as $k \in \mathcal{K}$ approaches infinity.

Proof. We may obtain the estimate

$$(5.37) \qquad \begin{bmatrix} \bar{G}(x^*, \lambda^*) & A^T(x^*) \\ A(x^*) & 0 \end{bmatrix} \begin{bmatrix} x_a^{*(k)} - x^* \\ \lambda_a^{*(k)} - \lambda^* \end{bmatrix} = \begin{bmatrix} \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) \\ c(x_a^{*(k)}) \end{bmatrix}$$
$$+ O(\max (\|x_a^{*(k)} - x^*\|_2^2, \|x_a^{*(k)} - x^*\|_2 \|\lambda_a^{*(k)} - \lambda^*\|_2)),$$

in exactly the way that we derived (4.14) in the proof of Theorem 4.2. Equation (5.12) shows that $\bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) = O(\mu^{(k)2})$ and (5.13) combines with (4.1), and (5.10) to give $c(x_a^{*(k)}) = \mu^{(k+1)}(\lambda^* - u) + O(\mu^{(k)2}) + o(\mu^{(k+1)})$. Then, as $\mu^{(k)2} = o(\mu^{(k+1)})$ (see (AS.6)), we may rewrite (5.37) as

$$\begin{bmatrix} \bar{G}(x^*, \lambda^*) & A^T(x^*) \\ A(x^*) & 0 \end{bmatrix} \begin{bmatrix} x_a^{*(k)} - x^* \\ \lambda_a^{*(k)} - \lambda^* \end{bmatrix} = \mu^{(k+1)} \begin{bmatrix} 0 \\ \lambda^* - u \end{bmatrix}$$
$$+ o(\mu^{(k+1)}) + O(\max (\|x_a^{*(k)} - x^*\|_2^2, \|x_a^{*(k)} - x^*\|_2 \|\lambda_a^{*(k)} - \lambda^*\|_2)).$$

This then gives the estimates

$$(5.38) \qquad x_a^{*(k)} = x^* + \mu^{(k+1)} v + o(\mu^{(k+1)}), \quad \text{and}$$

$$(5.39) \qquad \lambda_a^{*(k)} = \lambda^* + \mu^{(k+1)} w + o(\mu^{(k+1)}).$$

But, from Corollary 5.3, $x^{(k+1,0)} = x_a^{*(k)}$; Theorem 5.11 gives that $x^{*(k+1)}$ is either $x^{(k+1,0)}$ or $x^{(k+1,1)}$ where Corollary 5.9 implies that $x^{(k+1,1)} = x^{(k+1,0)} + p^{(k+1,0)}$ and where Lemma 5.5 gives $p^{(k+1,0)} = O(\mu^{(k)2}) = o(\mu^{(k+1)})$. Thus (5.35) is true.   □

THEOREM 5.14. *Suppose that (AS.1)-(AS.6) hold. Then the whole sequence of iterates $\{x^{*(k)}\}$ generated by the algorithm converges to a single limit point x^*. Furthermore, the iterates converge at least two-step Q-linearly if $\{\mu^{(k)}\}$ converges Q-linearly and at least two-step Q-superlinearly if $\{\mu^{(k)}\}$ converges Q-superlinearly.*

Proof. Let x^* be a limit point of the sequence $\{x^{*(k)}\}$. As x^* is a limit point, there is some infinite subsequence of iterates $\mathcal{K}_1$ and, from Theorem 4.2, some bound κ for which $\|x^{*(k)} - x^*\|_2 \leq \kappa \mu^{(k)}$ for all $k \in \mathcal{K}_1$. Without loss of generality we may assume that $\kappa \geq 2\|v\|_2$, where v is given by (5.36).

Now consider the set

$$\mathcal{K} = \{k : \|x^{*(k)} - x^*\|_2 \leq \kappa \mu^{(k)}\}.$$

Clearly $\mathcal{K}_1 \subseteq \mathcal{K}$ and $\mathcal{K}$ defines a convergent subsequence of the iterates. Then Lemma 5.13 applies to this subsequence. In particular, (5.35) shows that there is an index k_0 such that

$$(5.40) \qquad \|x^{*(k+1)} - x^*\|_2 \leq 2\|v\|_2 \mu^{(k+1)}$$

for all $k \geq k_0$, $k \in \mathcal{K}$; the factor two accounts for the asymptotically smaller term in equation (5.35). But then (5.40) implies that $k + 1 \in \mathcal{K}$ provided that $k \geq k_0$. Hence, inductively, $\mathcal{K}$ contains all $k \geq k_0$ and the entire sequence $\{x^{*(k)}\}$ converges to x^*. Finally, we now have that the estimate (5.35) holds for *all* k large enough and thus the Q-rate of convergence of the penalty parameters is inherited by the iterates $x^{*(k)}$.   □

6. **Comments.** We conclude the paper with a number of comments and remarks.

(i) This paper was motivated by a need to analyse the algorithm sketched in Gould [14]. In that paper, the sequence of penalty parameters $\{\mu^{(k)}\}$ was constructed so that $\mu^{(k+1)} = 0.01\mu^{(k)}$ and the convergence of the iterates was observed to be at a fast two-step linear rate. We have performed a number of experiments with superlinearly convergent sequences $\{\mu^{(k)}\}$, but the numerical results obtained are never substantially better than those given in Gould [14] for the linearly convergent sequence. We suspect this really indicates that a fast linearly convergent algorithm is quite adequate on a finite precision computer where the analytic advantages of superlinear convergence only start to occur as the computation is terminated. However, for completeness, we give details of some of the numerical experiments we have made.

All of the linear algebraic processes in the algorithm of § 3 may be performed in a stable fashion. In particular, the block linear equations (3.4) and (3.7) may be solved using the symmetric indefinite matrix techniques of Bunch and Parlett [4], Fletcher [9], and Bunch and Kaufman [3]. The only time we need to compute carefully is in the calculation of the constraint values. In particular, the calculation of $\bar{\lambda}$ requires that we form the ratio $c(x)/\mu$, and we should expect both $c(x)$ and μ to approach zero (in an analytically well-behaved fashion). However, computer rounding errors and in particular cancellation errors in forming $c(x)$ can be magnified in an unfortunate way by μ. We suggest that the constraint values themselves be computed in a higher precision than the rest of the computation to alleviate this difficulty.

In order to illustrate the performance of the new algorithm, we consider the following simple test problem.

PROBLEM 6.1. Minimize $\sum_{i=1}^{6} ix_i^2$ subject to the constraints

$$(x_1 + x_3 + x_5)^2 = 1, \quad (x_2 + x_3 + x_4)^2 = 1, \quad x_1 x_6 = 1,$$

starting from the initial point $(-2, 1.5, 2, -1, -1, 3)^T$.

We compare the method suggested here with the stabilized use of the quadratic penalty function as given by Gould [14]. All of our experiments were performed on the CRAY 2 computer at Harwell, using extended precision arithmetic, with the sequence of penalty parameters 10^{-i}, with $i = 1, 2, 3, 4, 6, 9, 14$, and 22. The other parameters used in our implementation of the new algorithm were chosen to be $\gamma = 1$, $\tau = 0.1$, $\varepsilon = 10^{-10}$, $\beta_1 = 10^{-4}$, and $\beta_2 = 0.1$, and all components of u were set to zero. Both algorithms take six iterations to reduce the gradient of the penalty function to 3.3D − 2 for the value $\mu = 10^{-1}$, and we describe the progress from there onward—unit stepsizes were always acceptable for both methods with such values of μ. The results for the method proposed in § 3 are summarized in Table 6.1 and those for the method that does not use the alternative starting point in Table 6.2. We show the progress of each inner iteration as the penalty parameter is reduced. For the new method, we observe (Note 1) that the gradient of the penalty function at iteration 1 with $\mu = 10^{-14}$ is starting to suffer from numerical cancellation when evaluating the constraint values and the magnification of this error through dividing by μ. We would expect to have only about 28—14 significant figures here (the machine precision is roughly 10^{-28} on the CRAY for extended precision arithmetic). We see this effect more seriously for $\mu = 10^{-22}$ (Note 2), where we would only expect to have six figures correct in the constraint values, and the subsequent loss in accuracy when forming the gradient. The figures for larger values of μ follow closely the theory predicted in the paper—indeed it was figures like these that actually suggested the theory!

Again, in Table 6.2, we observe (Notes 3 and 4) the effects of finite precision arithmetic when calculating the gradients for small values of the penalty parameter.

TABLE 6.1
The progress of new method on problem 6.1.

μ	Inner iteration	Calculate derivatives? ($\times$ = yes)	Gradient norm	Comments
1.0D–2	0	$\times$	9.4D + 0	alternative starting point
	1	$\times$	2.4D + 0	
	2	$\times$	1.1D – 1	
	3	$\times$	4.6D – 4	
1.0D–3	0	$\times$	1.4D + 0	alternative starting point
	1	$\times$	5.9D – 4	
1.0D–4	0	$\times$	2.9D + 0	alternative starting point
	1	$\times$	5.6D – 7	
1.0D–6	0	$\times$	1.8D – 1	alternative starting point
	1	$\times$	8.8D – 9	
1.0D–9	0	$\times$	1.9D – 2	alternative starting point
	1	$\times$	9.1D – 14	
1.0D–14	0	$\times$	1.9D – 3	alternative starting point
	1	$\times$	3.2D – 15	(see Note 1, p. 122)
1.0D–22	0	$\times$	1.8D – 5	alternative starting point
	1	$\times$	5.9D – 6	(see Note 2, p. 122)

Significantly, we do not observe the same pleasing behaviour in Table 6.2 that we saw in Table 6.1 with respect to the number of inner iterations performed for each outer iteration. How many inner iterations will be required per outer iteration for the older method is not obvious. We observe that the (norm of the) gradient at the end of the first iteration of the older method appears to be of the same order of magnitude as that at the alternative starting point for the new algorithm. However, subsequent Newton steps (with unit stepsizes), only approximately square the size of the gradient at each step (as we might expect from Newton's method). This is in contrast to the behaviour predicted for the new scheme in Lemma 5.10, where the alternative starting point leads to a greater reduction in the size of the gradient at the next step than might be expected just from the use of Newton's method.

In Table 6.3 we show how the new algorithm performs on the subset of Hock and Schittkowski [16] test problems used by Gould [14]—except that problems 68 and 69 are not used now as they require an accurate computation of the error function erf(x), which is unfortunately not available on the CRAY 2 at Harwell. For every test problem, we record the number of gradient evaluations required by the algorithm for each value of the penalty parameter used. The sequence of penalty parameters and other constants used is exactly as described above. We note the widely differing asymptotic regions—the regions within which two gradient evaluations are sufficient for the convergence of the inner iteration—for the various test examples, but also note that the predicted asymptotic behaviour takes place in all cases.

(ii) Although we have concentrated on equality constraints in this paper, our results may be extended to inequality constrained problems in a natural way. If we consider the inequality constrained problem

(6.1a)   minimize $f(x)(x \in R^n)$

TABLE 6.2
The progress of the method in Gould [14] *on problem* 6.1.

μ	Inner iteration	Calculate derivatives? ($\times$ = yes)	Gradient norm	Comments
1.0D–2	0		2.3D + 1	
	1	$\times$	3.9D + 2	
	2	$\times$	1.1D + 2	
	3	$\times$	2.2D + 1	
	4	$\times$	2.7D + 0	
	5	$\times$	3.3D + 0	
	6	$\times$	3.1D − 1	
	7	$\times$	9.6D − 4	
1.0D–3	0		3.9D + 1	
	1	$\times$	2.9D + 0	
	2	$\times$	4.5D − 2	
	3	$\times$	3.9D − 3	
	4	$\times$	3.1D − 8	
1.0D–4	0		4.1D + 1	
	1	$\times$	2.6D − 1	
	2	$\times$	4.2D − 3	
	3	$\times$	1.9D − 6	
1.0D–6	0		4.5D + 2	
	1	$\times$	3.2D − 1	
	2	$\times$	8.7D − 3	
	3	$\times$	5.9D − 6	
	4	$\times$	2.8D − 12	
1.0D–9	0		4.6D + 3	
	1	$\times$	3.3D − 2	
	2	$\times$	8.9D − 4	
	3	$\times$	6.3D − 9	
	4	$\times$	3.2D − 17	
1.0D–14	0		4.6D + 5	
	1	$\times$	3.3D − 3	
	2	$\times$	8.9D − 5	
	3	$\times$	6.3D − 12	
	4	$\times$	2.8D − 15	(see Note 3, p. 122)
1.0D–22	0		4.6D + 8	
	1	$\times$	3.3D − 5	
	2	$\times$	1.6D − 6	(see Note 4, p. 122)

subject to the constraints

(6.1b) $$c_i(x) \geqq 0, \quad \text{for all } i \in \mathscr{I} = \{t+1, \cdots, m\},$$

then a suitable penalty function for this problem (see, Rockafellar [20]) is the function

(6.2) $$\Phi(x, \mu) = f(x) + \sum_{i \in \mathscr{I}} \begin{cases} c_i(x)u_i + (1/2\mu)c_i(x)^2 & \text{if } c_i(x) = \mu u_i \leqq 0, \\ -(\mu/2)u_i^2 & \text{otherwise,} \end{cases}$$

where, as before, the u_i, $i \in \mathscr{I}$ are fixed, finite scalars. Once again, it is well known

TABLE 6.3
The performance of the new algorithm on some test problems of Hock and Schittkowski [16].

Example number	Number of gradient evaluations for the particular penalty parameter value							
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-6}	10^{-9}	10^{-14}	Total
43	9	3	2	2	2	2	2	22
64	19	4	3	3	3	3	2	37
74	7	2	2	2	2	2	2	19
75	8	5	5	4	3	3	2	30
78	2	6	8	4	4	3	2	29
80	6	5	2	2	2	2	2	21
81	5	2	2	2	2	2	2	17
83	3	3	5	5	3	3	2	24
86	4	3	3	3	2	2	2	19
106	52	23	23	3	3	2	2	108
111	11	8	8	4	3	2	2	38
112	11	2	5	3	3	2	2	28
117	37	34	3	2	2	2	2	82

that, under relatively mild conditions,

$$\lim_{\mu \to 0_+} \lim_{x \to x(\mu)} x = x^*,$$

where $x(\mu)$ and x^* are, respectively, local minimizers of (6.2) and local solutions of (6.1). If we make a strict complementary slackness assumption—that all of the Lagrange multipliers corresponding to active constraints at x^* are strictly negative—it is straightforward to extend the results of § 4 to show that limit points of iterates generated by the obvious adaptation of the algorithm of § 3 converge to Kuhn-Tucker points of (6.1). Moreover, for k large enough, the algorithm correctly identifies the active set for the solution (in the sense that, for k large enough, $c_i(x^{*(k)}) + \mu^{(k)} u_i \leq 0$ if and only if the ith constraint is active at x^*). Even more importantly, for k large enough, we can show that the use of (6.2) is equivalent to the use of (2.2) with the set $\mathscr{E}$ chosen to be the indices of those constraints active at a limit point of the algorithm. Briefly, when μ is small and x is in some neighbourhood of x^*, the two functions differ only at points for which $c_i(x) + \mu u_i > 0$ for some $i \in \mathscr{E}$. For k large enough, we have shown that the only iterates generated by the algorithm as applied to (2.2) are $\{x^{*(k)}\}$ and $\{x_a^{*(k-1)}\}$. At the former points we have

$$c_i(x^{*(k)}) + \mu^{(k)} u_i = \mu^{(k)} \lambda_i^* + o(\mu^{(k)}), \qquad i \in \mathscr{E}$$

from (4.2); for the latter (5.15), (5.39), and (AS.6) combine to give

$$c_i(x_a^{*(k-1)}) + \mu^{(k)} u_i = \mu^{(k)} \lambda_i^* + o(\mu^{(k)}), \qquad i \in \mathscr{E}.$$

Thus in all cases the strict complementary slackness assumption ensures that $c_i(x) + \mu u_i < 0$ for all $i \in \mathscr{E}$, and hence (1.2) and (6.2) are the same for all iterates encountered for all sufficiently large k. Thus the rates of convergence of the iterates obtained in § 5 apply equally to the obvious modification of the algorithm of § 3 when applied to the problem (6.1). In particular, a "popular" concern that second derivative discontinuities of Φ will cause inefficiencies is avoided as the iterates generated by our algorithm all lie within a region in which the discontinuities do not occur when k is large enough.

(iii) In this paper we have only considered using the penalty function (1.2) with u fixed. If we allow u to vary, we move into the realms of the classical augmented Lagrangian methods. However, such methods normally operate under the condition that μ is bounded away from zero and convergence is encouraged by means of forcing u to approach the vector of Lagrange multipliers at a Kuhn–Tucker point. Although it is beyond the scope of this paper, it is interesting to speculate what would happen if we allow μ to approach zero at the same time as encouraging u to converge to λ^{*}. A preliminary perusal of the results in §§ 4 and 5 indicates that the theory still holds if we allow different u, $u^{(k)}$, for each outer iteration, provided that the $\{u^{(k)}\}$ remain bounded and satisfy the asymptotic estimate

(6.3) $$u^{(k+1)} - u^{(k)} = O(\mu^{(k)2}/\mu^{(k+1)})$$

as $k \in \mathcal{K}$ tends to infinity. If in addition we ensure that $u^{(k)}$ is a sufficiently good approximation to λ^{*} we would hope to improve upon the estimates (4.9) and (4.10). In particular it seems likely that if we pick $u^{(k+1)} = \lambda^{*(k)}$, then we will obtain the improved estimate

(6.4) $$x^{*(k)} = x^{*} + O(\mu^{(k-1)}\mu^{(k)})$$

while if we choose $u^{(k+1)} = \lambda_a^{*(k)}$, then we will obtain the estimate

(6.5) $$x^{*(k)} = x^{*} + O(\mu^{(k)2}),$$

and we conjecture that (6.4) and (6.5) are true. Although such estimates do not imply a faster Q-rate of convergence, they would improve on the size of the error for a given value of $\mu^{(k)}$. We are currently pursuing this line of research.

(iv) We have intentionally chosen only to analyse methods for which exact second derivative information has been available. In some cases such information is hard, impossible or extremely expensive to obtain and our algorithm is inappropriate. It is of course easy to formulate an algorithm along the lines of that given in § 3 in which all occurrences of the second derivative matrix $\bar{G}(x^{(k,l)}, \bar{\lambda}^{(k,l)})$ are replaced by a suitable approximation $B^{(k,l)}$. Although we have not analysed such a situation in much detail, we make the following observations; we stress that these comments only apply to our method of proof in § 5 and may be invalid for other proof techniques.

First, it is crucial that the approximations should all asymptotically satisfy the second-order condition (in the sense that the matrix obtained by replacing $\bar{G}(x^{(k,l)}, \bar{\lambda}^{(k,l)})$ in $K(x^{(k,l)}, \bar{\lambda}^{(k,l)}, \mu^{(k)})$ by $B^{(k,l)}$ should be uniformly nonsingular and have precisely t negative eigenvalues for large k). Failure to do so may inhibit the algorithm from picking the alternative starting point, an important ingredient in the algorithm we analysed in § 5.

Second, the results of Theorem 5.1 are unaffected by approximations to the second derivatives although the estimate (5.12) used in its proof is weakened to

(6.6) $$\bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) = O(\|(\bar{B}^{(k)} - \bar{G}(x^{*(k)}, \lambda^{*(k)}))p^{(k)}\|_2),$$

where $\bar{B}^{(k)}$ is the appropriate approximation to $\bar{G}(x^{*(k)}, \lambda^{*(k)})$. However, our proof of Theorem 5.4, which shows a one-step Q-superlinear rate of the method when applied to unconstrained problems, depends crucially upon the estimate (5.12), and thus (6.6), and may therefore no longer be true. In particular, one-dimensional examples that use the secant method to approximate the second derivatives can be constructed for which

(6.7) $$(\bar{B}^{(k)} - \bar{G}(x^{*(k)}, \lambda^{*(k)}))p^{(k)} = O_S(\mu^{(k)v}),$$

for some constant $1 < v < 2$ as k approaches infinity. Therefore, unless (AS.6) is strengthened so that the sequence $\{\mu^{(k)}\}$ satisfies the estimate

$$(6.8) \qquad\qquad \mu^{(k)v}/\mu^{(k+1)} = o(1)$$

as k approaches infinity, we may be unable to derive a one-step Q-superlinear rate of convergence for the algorithm in the unconstrained case for an important class of second-derivative approximations. We should not expect things to improve in the constrained case and believe that this indicates that we need to be more conservative in how we pick $\mu^{(k+1)}$ when we have approximate second derivatives.

Third, we cannot obtain the estimate (5.17) using the technique given in the proof of Lemma 5.5 unless

$$(6.9) \qquad\qquad (\bar{B}^{(k)} - \bar{G}(x^{*(k)}, \lambda^{*(k)}))p^{(k)} = O(\mu^{(k)2}).$$

In particular, if

$$(\bar{B}^{(k)} - \bar{G}(x^{*(k)}, \lambda^{*(k)}))p^{(k)} = O(\mu^{(k)v}) \quad \text{then } p^{(k+1,0)} = O(\mu^{(k)v})$$

for $1 < v < 2$. However, the principal use of Lemma 5.5 is in the proof of Theorem 5.8 where we need the estimate

$$\|p^{(k+1,0)}\|_2 / \mu^{(k+1)} = o(1);$$

this estimate is obtained by combining Lemma 5.5 with (AS.6). The same estimate is possible if

$$(6.10) \qquad\qquad (\bar{B}^{(k)} - \bar{G}(x^{*(k)}, \lambda^{*(k)}))p^{(k)} = O(\mu^{(k)v})$$

and if we restrict $\mu^{(k+1)}$ by (6.8).

Fourth, Lemma 5.6 remains true for any bounded approximation B to $\bar{G}(x, \lambda)$ (in the sense that all occurrences of $\bar{G}$ in the statement of the lemma are replaced by B) as does the relationship (5.23). The results of Theorem 5.8 then remain true provided that

$$(6.11) \qquad p^{(k+1,0)T}(B^{(k+1,0)} - \bar{G}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}))p^{(k+1,0)} = o(p^{(k+1,0)T}p^{(k+1,0)})$$

for all sufficiently large k. This relationship is needed to establish (5.28) and to show that the change in the values of (5.31) and (5.33) due to the approximation of second-derivative information is negligible in comparison to the dominant terms in the two expressions.

Finally, if we assume that

$$(6.12) \qquad\qquad (B^{(k+1,0)} - \bar{G}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}))p^{(k+1,0)} = O(\mu^{(k)v}),$$

(6.8) combines with (6.12) to give Theorem 5.11; the quantity $\bar{g}(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)})$ is now $O(\mu^{(k)v})$. The same assumptions are also needed to extend Theorem 5.14 for approximate second derivatives.

In summary, it would appear that we can use approximate second derivatives within our framework provided that the approximations satisfy the conditions outlined above—of course we have not proved this rigorously here but believe it to be correct. Whether these conditions are necessary in general is unknown; moreover, precisely which approximations (if any) satisfy such conditions is at present unclear but under investigation.

(v) The determination of an alternative starting point is the key to the success of the algorithm suggested in this paper. It has essentially three beneficial effects: it gives us a point which is close to a stationary point of the new merit function, a point at

which the gradient of the merit function is also small (in view of the ill-conditioning of the penalty function these are not necessarily equivalent), and enables us to maintain bounded Lagrange multiplier estimates $\bar{\lambda}$. There is every reason to believe that algorithms based upon other simple penalty functions or barrier functions would benefit from alternative starting points and that such points may be generated from the appropriate version of (3.10).

7. Acknowledgement. The author thanks the referees for their helpful comments on an earlier version of this paper.

REFERENCES

[1] L. ARMIJO, *Minimization of functions having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.
[2] C. G. BROYDEN AND N. F. ATTIA, *A smooth sequential penalty function method for solving nonlinear programming problems*, in System Modelling and Optimization, A. V. Balakrishnan and M. Thomas, eds., Lecture Notes in Control and Information Sciences 59, Springer–Verlag, Berlin, 1984.
[3] J. R. BUNCH AND L. C. KAUFMAN, *Some stable methods for calculating inertia, and solving symmetric linear equations*, Math. Comp., 31 (1977), pp. 163-179.
[4] J. R. BUNCH AND B. N. PARLETT, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639-655.
[5] R. H. BYRD AND R. B. SCHNABEL, *Continuity of the null space basis and constrained optimization*, Math. Programming, 35 (1986), pp. 32-41.
[6] T. F. COLEMAN AND D. C. SORENSEN, *A note on the computation of an orthogonal basis for the null-space of a matrix*, Math. Programming, 29 (1984), pp. 234-242.
[7] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[8] A. V. FIACCO AND G. P. McCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
[9] R. FLETCHER, *Factorizing symmetric indefinite matrices*, Linear Algebra Appl., 14 (1976), pp. 257-272.
[10] ———, *Practical Methods of Optimization, Vol.* 2, John Wiley, Chichester, New York, 1981.
[11] A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper and Row, New York, 1967.
[12] P. E. GILL, W. MURRAY, M. A. SAUNDERS, G. W. STEWART, AND M. H. WRIGHT, *Properties of a representation of a basis for the null space*, Math. Programming, 33 (1985), pp. 172-186.
[13] N. I. M. GOULD, *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem*, Math. Programming, 32 (1985), pp. 90-99.
[14] ———, *On the accurate determination of search directions for simple differentiable penalty functions*, IMA J. Numer. Anal., 6 (1986), pp. 357-372.
[15] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303-320.
[16] W. HOCK AND K. SCHITTKOWSKI, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Economics and Mathematical Systems 187, Springer–Verlag, Berlin, 1981.
[17] W. MURRAY, *Analytical expressions for eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions*, J. Optim. Theory Appl., 7 (1971), pp. 189-196.
[18] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, London, New York, 1970.
[19] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, London, New York, 1969, pp. 283-298.
[20] R. T. ROCKAFELLAR, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control Optim., 12 (1974), pp. 268-285.
[21] D. M. RYAN, *Penalty and barrier functions*, in Numerical Methods for Constrained Optimization, P. E. Gill and W. Murray, eds., Academic Press, London, 1974, pp. 175-190.