SIAM J. NUMER. ANAL.
Vol. 28, No. 2, pp. 545-572, April 1991

# A GLOBALLY CONVERGENT AUGMENTED LAGRANGIAN ALGORITHM FOR OPTIMIZATION WITH GENERAL CONSTRAINTS AND SIMPLE BOUNDS*

ANDREW R. CONN†, NICHOLAS I. M. GOULD‡, AND PHILIPPE L. TOINT§

**Abstract.** The global and local convergence properties of a class of augmented Lagrangian methods for solving nonlinear programming problems are considered. In such methods, simple bound constraints are treated separately from more general constraints and the stopping rules for the inner minimization algorithm have this in mind. Global convergence is proved, and it is established that a potentially troublesome penalty parameter is bounded away from zero.

**Key words.** constrained optimization, augmented Lagrangian, simple bounds, general constraints

**AMS(MOS) subject classifications.** 65K05, 90C30

**1. Introduction.** In this paper, we consider the problem of finding a local minimizer of the function

$$f(x), \tag{1.1}$$

where $x$ is required to satisfy the constraints

$$c_i(x) = 0, \qquad 1 \leqq i \leqq m, \tag{1.2}$$

and the simple bounds

$$l \leqq x \leqq u. \tag{1.3}$$

Here $f$ and $c_i$ map $R^n$ into $R$ and inequalities (1.3) are considered componentwise; we shall assume that the region $B = \{x \mid l \leqq x \leqq u\}$ is nonempty and may be infinite. We further assume that

(AS1) The functions $f(x)$ and $c_i(x)$ are twice continuously differentiable for all $x \in B$.

We assume that any general inequality constraints $c_i(x) \geqq 0$ have already been converted into equations by the introduction of slack variables (see, e.g., Fletcher (1981, p. 8)); we wish the combinatorial side of the minimization problem to be represented purely in terms of simple bound constraints. We shall attempt to solve our problem by means of a sequential minimization of the *augmented Lagrangian function*

$$\Phi(x, \lambda, S, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{1}{2\mu} \sum_{i=1}^{m} s_{ii} c_i(x)^2, \tag{1.4}$$

where the components $\lambda_i$ of the vector $\lambda$ are known as Lagrange multiplier estimates, where the entries $s_{ii}$ of the diagonal matrix $S$ are positive scaling factors, and where $\mu$ is known as the penalty parameter. Note that we *do not* include the simple bounds (1.3) in the augmented Lagrangian function; rather the intention is that the sequential minimization will automatically ensure that these constraints are always satisfied.

* Received by the editors October 31, 1988; accepted for publication (in revised form) April 4, 1990.

† Department of Combinatorics and Optimization, University of Waterloo, Ontario, Canada. The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada and by the Information Technology Research Centre, which is funded by the Province of Ontario.

‡ Computer Science and Systems Division, United Kingdom Atomic Energy Authority, Harwell, Oxford, United Kingdom.

§ Department of Mathematics, Facultés Universitaires Notre Dame de la Paix, B-5000, Namur, Belgium.

Our principal interest is in solving large-scale problems. With a few notable exceptions (see, for example, Murtagh and Saunders (1980), Lasdon (1982), Drud (1985)), there has been little progress in constructing algorithms for such problems; this is somewhat understandable in view of the lack of a consensus as to the "best" algorithm for solving small nonlinear programs. Nevertheless, there are many large-scale applications awaiting a suitable algorithm.

A similar situation existed for unconstrained optimization in the early 1970s. However, during the past ten years, this deficiency has been redressed primarily through the development of three important ideas. The first is the recognition that large problems normally have considerable structure and that such structure usually manifests itself as sparsity or low rank of the relevant matrices. This has lead to suitable ways of storing and approximating problem data (function, gradient, and Hessian approximations), see, for example, Griewank and Toint (1982). The second development is the realization that, although Newton's method (or a good approximation to it) is necessary for rapid asymptotic convergence of an algorithm, in early iterations only very crude approximations to the solution of the Newton equations are needed to guarantee global convergence. In particular, the steepest descent method often makes very good initial progress towards a minimizer. This has led to a study of realistic conditions that suffice to guarantee global convergence of an algorithm and also of methods which satisfy such conditions, the truncated conjugate gradient method being a particularly successful example. This work is described, for example, by Toint (1981), Dembo, Eisenstat, and Steihaug (1982), and Steihaug (1983). Third, the development of trust-region methods (see, e.g., Moré (1983)) has allowed a sensible handling of negative curvature in the objective function; for large-scale problems whose second derivatives are available (contrary to popular belief, an extremely common circumstance in many problem areas), this enables meaningful steps towards the solution to be made when the Hessian matrix is indefinite. Significantly, these ideas have had an important impact on the design of algorithms not only for large problems but also for small ones (see, Toint (1988), and Dixon, Dolan, and Price (1988)).

One issue that is not present in unconstrained minimization, but is in evidence here, is the *combinatorial problem* of finding which of the variables lie at a bound at the solution (such bound constraints are said to be active). In active-set algorithms, the intention is to predict these variables and to minimize the function with respect to the remaining variables. Obviously, an incorrect prediction is undesirable, and it is then useful (indeed essential for large problems) to be able to make rapid changes in the active set to correct for wrong initial choices. Unfortunately, many existing algorithms for constrained optimization only allow very small changes in the active set at each iteration, and consequently, for large problems, there is the possibility of requiring a large number of iterations to find the solution. Fortunately, for simple bound constraints, it is easy to allow for rapid changes in the active set in the design of algorithms (see, e.g., Berksekas (1982b, pp. 76–92), and Conn, Gould, and Toint, (1988a)).

Our intention here is to develop a fairly general algorithm which may benefit from the above-mentioned advances. We have recently developed and tested (Conn et al. (1987), Conn, Gould, and Toint (1988a), (1988b)) an algorithm for solving *bound constrained minimization problems* (problems of the form minimize (1.1) subject to (1.3)) which is appropriate in the large-scale case. Our basic idea is now to use this algorithm within an augmented Lagrangian framework, that is to use the algorithm to find an *approximation* to a minimizer of the augmented Lagrangian function (1.4) *subject to the bounds* (1.3) for a sequence of different values of $S$, $\lambda$, and $\mu$. The novelty comes from being able to solve the augmented Lagrangian problems approximately and on being able to deal with the bounds in an efficient manner.

The augmented Lagrangian method was proposed independently by Hestenes (1969) and Powell (1969), partly as a reaction to the unfortunate side-effects associated with ill-conditioning of the simpler differentiable penalty and barrier functions (Murray (1971)). Indeed, Powell showed, using a very simple device, how to ensure that the penalty parameter does not converge to zero and hence that the resulting ill-conditioning does not occur. A similar device is employed in the algorithms with which we are concerned in this paper with the same consequence. A concise statement of the salient features of augmented Lagrangian methods, or multiplier methods as they are sometimes known, is given, for example, by Fletcher (1981). The most comprehensive references on augmented Lagrangians are the paper by Tapia (1977) and the book by Bertsekas (1982b). Globally convergent methods have been given by Powell (1969), Rockafellar (1976), Bertsekas (1982b), Polak and Tits (1980), Yamashita (1982), Bartholomew-Biggs (1987), and Hager (1987). Powell's method requires that the augmented Lagrangian be minimized exactly for fixed values of the multipliers and parameters. The multiplier estimates are guaranteed to be bounded, but convergence is only established in the case where the underlying nonlinear program has a unique solution. Rockafellar, Bertsekas, and Polak and Tits allow inexact minimization of the augmented Lagrangian function, but they require that the Lagrange multiplier estimates remain bounded—the methods differ in the stopping criteria used. Hager is slightly more restrictive in that he considers a particular multiplier update and specifies the method used for approximately minimizing the augmented Lagrangian function. His analysis also assumes that a subsequence of the Lagrange multiplier estimates converges. Yamashita and Bartholomew-Biggs are more specific in the method used for the inner minimization calculation—an appropriate quadratic program is solved—but their methods allow for more frequent updating of the penalty parameter and multiplier estimates. Yamashita establishes convergence under the assumption that the Lagrange multipliers for the quadratic programming problem stay bounded; the possibility of proving convergence for Bartholomew-Bigg's method under similar circumstances is only hinted at although encouraging numerical results are presented.

Interest in augmented Lagrangians declined with the introduction of successive quadratic programming (SQP) techniques but recently has gained in popularity. (See for example the papers of Schittkowski (1981) and Gill et al. (1986) which combine SQP with an augmented Lagrangian merit function. Both these methods are not *pure* augmented Lagrangian techniques since they perform a line search on the augmented Lagrangian as a function of both the position $x$ and the multipliers $\lambda$ in contrast to the method described in this paper.)

One strong disadvantage of SQP methods for large-scale problems is that, although there is a theory of how to truncate the solution process in the early iterations (see, Dembo and Tulowitzki (1984))—as is used so successfully in the unconstrained case—it is not clear to us how to construct an efficient algorithm that conforms to this theory. We feel that solving a quadratic programming to completion at each iteration is probably too expensive a calculation for large-scale problems in the same way that solving the Newton equations exactly is considered too expensive in large-scale unconstrained minimization. We thus feel there are compelling reasons for trying to use an alternative to the SQP approach.

Bertsekas (1982a) and others, however, have remarked that augmented Lagrangians are particularly attractive for large problems, where active set strategies are inappropriate, and we tend to agree with this sentiment. In particular, simple multiplier estimates may be used. In this paper we explore some of the issues involved in using an augmented Lagrangian approach for large-scale problems. We have deliberately not included the results of numerical testing as, in our view, the construction of appropriate software is by no means trivial and we wish to make a thorough job of it. We will report on our numerical experience in due course. We should comment, nonetheless, that at the heart of any method for solving nonlinear programming problems, there is a need to find an approximate solution to a system of linear equations. Linear equation solvers may be broadly categorised as either direct or iterative methods. In the former, a factorization of the relevant matrix is used, while in the latter, matrix-vector products involving the relevant matrix are required. Here the coefficient matrix for such systems will typically be symmetric submatrices of the Hessian of the augmented Lagrangian function (1.4). For iterative methods, sparsity in the derivatives of the objective function and constraints may be exploited in the matrix-vector products. For direct methods, there is often a concern that the Hessian of an augmented Lagrangian function may be less sparse than for the Lagrangian function because of the last term in (1.4). While this is certainly true, it is worth noting that variables that appear *nonlinearly* in the constraint functions give rise to nonzeros in the same positions in the Hessians of both the Lagrangian and augmented Lagrangian function. It may therefore be worth treating linear constraints in a different way from nonlinear ones; we are currently pursuing this line of research.

Our exposition will be considerably simplified if we consider the special case where $l_i = 0$ and $u_i = \infty$ for all $1 \leqq i \leqq n$ in (1.3). Although straightforward, the modification required to handle more general constraints will be indicated at the end of the paper. Thus we consider the problem:

$$\text{minimize} \, f(x), \tag{1.5}$$

subject to the constraints

$$c_i(x) = 0, \qquad 1 \leqq i \leqq m, \tag{1.6}$$

and the nonnegativity restrictions

$$x \in B = \{x \in R^n \,|\, x \geqq 0\}. \tag{1.7}$$

The paper is organised as follows. In § 2 we introduce concepts and definitions and then state a pair of related algorithms for solving (1.5)–(1.7) in § 3. Global convergence is established in § 4, while issues of asymptotic convergence follow in § 5. An example showing the importance of a certain assumption in § 5 is given in § 6, while in § 7 the consequences of satisfying second-order conditions are given. We conclude in § 8 by indicating how this theory applies to the original problem (1.1)–(1.3).

**2. Notation.** In this section we introduce the notation to be used throughout the paper. We will use the projection operator defined componentwise by

$$(P[x])_i = \begin{cases} 0 & \text{if } x_i \leqq 0, \\ x_i & \text{otherwise.} \end{cases} \tag{2.1}$$

This operator projects the point $x$ onto the region $B$. Furthermore, we will make use of the "projection"

$$P(x, v) = x - P[x - v]. \tag{2.2}$$

Let $g(x)$ denote the gradient $\nabla_x f(x)$ of $f(x)$ and let $H(x)$ denote its Hessian matrix $\nabla_{xx} f(x)$. Let $A(x)$ denote the $m$-by-$n$ Jacobian of $c(x)$, where

$$c(x) = [c_1(x), \cdots, c_m(x)]^T, \tag{2.3}$$

and let $H_i(x)$ denote the Hessian matrix $\nabla_{xx}c_i(x)$ of $c_i(x)$. Finally, let $g_L(x, \lambda)$ and $H_L(x, \lambda)$ denote the gradient and Hessian matrix (taken with respect to its first argument) of the Lagrangian function

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x). \tag{2.4}$$

We note that $L(x, \lambda)$ is the Lagrangian function with respect to the $c_i$ constraints only. If we define *first-order Lagrange multiplier estimates*

$$\bar{\lambda}(x, \lambda, S, \mu) = \lambda + Sc(x)/\mu, \tag{2.5}$$

we shall make much use of the identity

$$\nabla_x \Phi(x, \lambda, S, \mu) = g_L(x, \bar{\lambda}(x, \lambda, S, \mu)). \tag{2.6}$$

Now suppose that $\{x^{(k)} \geqq 0\}$ and $\{\lambda^{(k)}\}$ are infinite sequences of $n$-vectors and $m$-vectors, respectively, that $\{S^{(k)}\}$ is an infinite sequence of positive-definite diagonal matrices, and that $\{\mu^{(k)}\}$ is an infinite sequence of positive scalars. For any function $F$, we shall use the notation that $F^{(k)}$ denotes $F$ evaluated with arguments $x^{(k)}$, $\lambda^{(k)}$, $S^{(k)}$, or $\mu^{(k)}$ as appropriate. So, for instance, using the identity (2.6), we have

$$\nabla_x \Phi^{(k)} = \nabla_x \Phi(x^{(k)}, \lambda^{(k)}, S^{(k)}, \mu^{(k)}) = g_L(x^{(k)}, \bar{\lambda}^{(k)}), \tag{2.7}$$

where we have written

$$\bar{\lambda}^{(k)} = \bar{\lambda}(x^{(k)}, \lambda^{(k)}, S^{(k)}, \mu^{(k)}). \tag{2.8}$$

For any $x^{(k)}$, we have two possibilities for each component $x_i^{(k)}$, namely,

$$\begin{array}{ll} \text{(i)} & 0 \leqq x_i^{(k)} \leqq (\nabla_x \Phi^{(k)})_i, \quad \text{or} \\ \text{(ii)} & (\nabla_x \Phi^{(k)})_i < x_i^{(k)}. \end{array} \tag{2.9}$$

In case (i) we then have

$$(P(x^{(k)}, \nabla_x \Phi^{(k)}))_i = x_i^{(k)}, \tag{2.10}$$

whereas in case (ii) we have

$$(P(x^{(k)}, \nabla_x \Phi^{(k)}))_i = (\nabla_x \Phi^{(k)})_i. \tag{2.11}$$

We shall refer to an $x_i^{(k)}$ which satisfies (i) as a *dominated* variable; a variable which satisfies (ii) is known as a *floating* variable. The algorithms we are about to develop construct iterates which force $P(x^{(k)}, \nabla_x \Phi^{(k)})$ to zero as $k$ increases. The dominated variables are thus pushed to zero, while the floating variables are allowed to find their own level.

If, in addition, there is a convergent subsequence $\{x^{(k)}\}$, $k \in K$, with limit point $x^*$, we wish to partition the set $N = \{1, 2, \cdots, n\}$ into the following four subsets which are related to the two possibilities (i) and (ii) above and to the corresponding $x^*$:

$$\begin{aligned} &I_1 = \{i \mid x_i^{(k)} \text{ are floating for all } k \in K \text{ sufficiently large and } x_i^* > 0\}, \\ &I_2 = \{i \mid x_i^{(k)} \text{ are dominated for all } k \in K \text{ sufficiently large}\}, \\ &I_3 = \{i \mid x_i^{(k)} \text{ are floating for all } k \in K \text{ sufficiently large but } x_i^* = 0\}, \text{ and} \\ &I_4 = N \backslash (I_1 \cup I_2 \cup I_3). \end{aligned} \tag{2.12}$$

From time to time we will slightly abuse notation by saying that a variable $x_i$ belongs to (for instance) $I_1$, when strictly we should say that the index of the variable belongs to $I_1$. We will also mention the components of a (given) vector in the set $I_1$ when strictly we mean the components of the vector whose indices lie in $I_1$.

If the iterates are chosen so that $P(x^{(k)}, \nabla_x\Phi^{(k)})$ approaches zero as $k$ increases, we have the following result.

LEMMA 2.1. *Suppose that* $\{x^{(k)}\}$, $k \in K$, *is a convergent subsequence with limit point* $x^*$, *that* $\lambda^{(k)}$, $S^{(k)}$, $\mu^{(k)} I_1$, $I_2$, $I_3$, *and* $I_4$ *are as above, and that* $P(x^{(k)}, \nabla_x\Phi^{(k)})$ *approaches zero as* $k \in K$ *increases. Then*

(i) *The variables in sets* $I_2$, $I_3$, *and* $I_4$ *all converge to their bounds*;

(ii) *The components of* $(\nabla_x\Phi^{(k)})_i$ *in the sets* $I_1$ *and* $I_3$ *converge to zero*; *and*

(iii) *If a component of* $(\nabla_x\Phi^{(k)})_i$ *in the set* $I_4$ *converges to a finite limit, then the limit is zero.*

*Proof.* (i) The result is true for variables in $I_2$ from (2.10), for those in $I_3$ by definition and for those in $I_4$ as, again from (2.10), there must be a subsequence of the $k \in K$ for which $x_i^{(k)}$ converges to zero.

(ii) The result follows for $i$ in $I_1$ and $I_3$ from (2.11).

(iii) This is true for $i$ in $I_4$ as there must be a subsequence of the $k \in K$ for which, from (2.11), $(\nabla_x\Phi^{(k)})_i$ converges to zero. □

It will sometimes be convenient to group the variables in sets $I_3$ and $I_4$ together and call the resulting set

$$I_5 = I_3 \cup I_4. \tag{2.13}$$

As we see from Lemma 2.1, $I_5$ gives variables which are zero at the solution and which may correspond to zero components of the gradient of the augmented Lagrangian function. These variables are potentially (dual) degenerate at the solution of the nonlinear programming problem.

We will let $\hat{g}(x)$ denote the components of $g(x)$ indexed by $I_1$. Similarily, $\hat{A}(x)$ denotes the corresponding columns of the Jacobian matrix; indeed any matrix $\hat{M}$ refers to the columns of the generic matrix $M$ indexed by $I_1$. In addition, we will define the *least-squares Lagrange multiplier estimates* (corresponding to the set $I_1$)

$$\lambda(x) = -(\hat{A}(x)^+)^T \hat{g}(x) \tag{2.14}$$

at all points where the right generalized inverse

$$\hat{A}(x)^+ = \hat{A}(x)^T(\hat{A}(x)\hat{A}(x)^T)^{-1} \tag{2.15}$$

of $\hat{A}(x)$ is well defined. We note that $\lambda(x)$ is differentiable; for completeness the derivative is given in the following lemma.

LEMMA 2.2. *Suppose that* (AS1) *holds. If* $\hat{A}(x)\hat{A}(x)^T$ *is nonsingular,* $\lambda(x)$ *is differentiable and its derivative is given by*

$$\nabla_x\lambda(x) = -(\hat{A}(x)^+)^T \hat{H}_L(x, \lambda(x)) - (\hat{A}(x)\hat{A}(x)^T)^{-1} R(x), \tag{2.16}$$

*where the ith row of* $R(x)$ *is* $(\hat{g}(x) + \hat{A}(x)^T\lambda(x))^T \hat{H}_i(x)$.

*Proof.* The result follows by observing that (2.14) may be rewritten as

$$r(x) - \hat{A}(x)^T\lambda(x) = \hat{g}(x) \quad \text{and} \quad \hat{A}(x)r(x) = 0 \tag{2.17}$$

for some vector $r(x)$. Differentiating (2.17) and eliminating the derivative of $r(x)$ from the resulting equations gives the required result. □

We stress that, as stated, the Lagrange multiplier estimate (2.14) is not a directly calculable quantity as it requires an a priori knowledge of $x^*$. It is merely introduced as an analytical device but we shall show in due course that a variant of this estimate may be calculated and used.

We are now in a position to describe more precisely the algorithms we propose to use.

**3. Statement of the algorithms.** In order to solve problem (1.5)-(1.7), we consider the following algorithmic models. Here $\|\cdot\|$ denotes any vector norm (or its subordinate matrix norm).

ALGORITHM 1.

**Step 0 [Initialization].** An initial vector of Lagrange multiplier estimates $\lambda^{(0)}$ is given. The positive constants $\eta_0$, $\mu_0$, $\omega_0$, $\tau<1$, $\gamma_1<1$, $\omega_* \ll 1$, $\eta_* \ll 1$, $\alpha_\omega$, $\beta_\omega$, $\alpha_\eta$, and $\beta_\eta$ are specified. The diagonal matrices $S_1$ and $S_2$, for which $0<S_1^{-1}\leqq S_2<\infty$, are given (the inequalities are to be understood elementwise for the diagonal elements). Set $\mu^{(0)}=\mu_0$, $\alpha^{(0)}=\min(\mu^{(0)},\gamma_1)$, $\omega^{(0)}=\omega_0(\alpha^{(0)})^{\alpha_\omega}$, $\eta^{(0)}=\eta_0(\alpha^{(0)})^{\alpha_\eta}$, and $k=0$.

**Step 1 [Inner iteration].** Define a diagonal scaling matrix $S^{(k)}$ for which $S_1^{-1}\leqq S^{(k)}\leqq S_2$. Find $x^{(k)}\in B$ such that

$$\|P(x^{(k)},\nabla_x\Phi^{(k)})\|\leqq\omega^{(k)}. \tag{3.1}$$

If

$$\|c(x^{(k)})\|\leqq\eta^{(k)}, \tag{3.2}$$

execute step 2. Otherwise, execute step 3.

**Step 2 [Test for convergence and update Lagrange multiplier estimates].** If $\|P(x^{(k)},\nabla_x\Phi^{(k)})\|\leqq\omega_*$ and $\|c(x^{(k)})\|\leqq\eta_*$, stop. Otherwise, set

$$\begin{aligned}
\lambda^{(k+1)}&=\bar\lambda(x^{(k)},\lambda^{(k)},S^{(k)},\mu^{(k)}),\\
\mu^{(k+1)}&=\mu^{(k)},\\
\alpha^{(k+1)}&=\min(\mu^{(k+1)},\gamma_1),\\
\omega^{(k+1)}&=\omega^{(k)}(\alpha^{(k+1)})^{\beta_\omega},\\
\eta^{(k+1)}&=\eta^{(k)}(\alpha^{(k+1)})^{\beta_\eta},
\end{aligned}\tag{3.3}$$

increment $k$ by one and go to step 1.

**Step 3 [Reduce the penalty parameter].** Set

$$\begin{aligned}
\lambda^{(k+1)}&=\lambda^{(k)},\\
\mu^{(k+1)}&=\tau\mu^{(k)},\\
\alpha^{(k+1)}&=\min(\mu^{(k+1)},\gamma_1),\\
\omega^{(k+1)}&=\omega_0(\alpha^{(k+1)})^{\alpha_\omega},\\
\eta^{(k+1)}&=\eta_0(\alpha^{(k+1)})^{\alpha_\eta},
\end{aligned}\tag{3.4}$$

increment $k$ by one and go to step 1.

ALGORITHM 2.

**Step 0 [Initialization].** An initial vector of Lagrange multiplier estimates, $\lambda^{(0)}$, is given. The nonnegative constant $\alpha_\eta$ and the positive constants $\eta_0$, $\mu_0$, $\tau<1$, $\omega_0$, $\gamma<1$, $\gamma_1<1$, $\omega_*\ll 1$, $\eta_*\ll 1$, $\nu$, $\alpha_\omega$, $\beta_\omega$, and $\beta_\eta$ are specified. The diagonal matrices $S_1$ and $S_2$, for which $0<S_1^{-1}\leqq S_2<\infty$, are given. Set $\mu^{(0)}=\mu_0$, $\alpha^{(0)}=\min(\mu^{(0)},\gamma_1)$, $\omega^{(0)}=\omega_0(\alpha^{(0)})^{\alpha_\omega}$, $\eta^{(0)}=\eta_0(\alpha^{(0)})^{\alpha_\eta}$, and $k=0$.

**Step 1 [Inner iteration].** Define a diagonal scaling matrix $S^{(k)}$ for which $S_1^{-1} \leqq S^{(k)} \leqq S_2$. Find $x^{(k)} \in B$ such that

$$\|P(x^{(k)}, \nabla_x \Phi^{(k)})\| \leqq \omega^{(k)}. \tag{3.5}$$

Compute a new vector of Lagrange multiplier estimates $\hat{\lambda}^{(k+1)}$. If

$$\|c(x^{(k)})\| \leqq \eta^{(k)}, \tag{3.6}$$

execute step 2. Otherwise, execute step 3.

**Step 2 [Test for convergence and update Lagrange multiplier estimates].** If $\|P(x^{(k)}, \nabla_x \Phi^{(k)})\| \leqq \omega_*$ and $\|c(x^{(k)})\| \leqq \eta_*$, stop. Otherwise, set

$$\begin{aligned}
\mu^{(k+1)} &= \mu^{(k)}, \\
\lambda^{(k+1)} &= \begin{cases} \hat{\lambda}^{(k+1)} & \text{if } \|\hat{\lambda}^{(k+1)}\| \leqq \nu(\mu^{(k+1)})^{-\gamma}, \\ \lambda^{(k)} & \text{otherwise,} \end{cases} \\
\alpha^{(k+1)} &= \min(\mu^{(k+1)}, \gamma_1), \\
\omega^{(k+1)} &= \omega^{(k)}(\alpha^{(k+1)})^{\beta_\omega}, \\
\eta^{(k+1)} &= \eta^{(k)}(\alpha^{(k+1)})^{\beta_\eta},
\end{aligned} \tag{3.7}$$

increment $k$ by one and go to step 1.

**Step 3 [Reduce the penalty parameter and update Lagrange multiplier estimates].** Set

$$\begin{aligned}
\mu^{(k+1)} &= \tau\mu^{(k)}, \\
\lambda^{(k+1)} &= \begin{cases} \hat{\lambda}^{(k+1)} & \text{if } \|\hat{\lambda}^{(k+1)}\| \leqq \nu(\mu^{(k+1)})^{-\gamma}, \\ \lambda^{(k)} & \text{otherwise,} \end{cases} \\
\alpha^{(k+1)} &= \min(\mu^{(k+1)}, \gamma_1), \\
\omega^{(k+1)} &= \omega_0(\alpha^{(k+1)})^{\alpha_\omega}, \\
\eta^{(k+1)} &= \eta_0(\alpha^{(k+1)})^{\alpha_\eta},
\end{aligned} \tag{3.8}$$

increment $k$ by one and go to step 1.

The motivation for both algorithms is quite straightforward. Traditional augmented Lagrangian methods are known to be locally convergent if the penalty parameter is sufficiently small and if the augmented Lagrangian is approximately minimized at each stage (see, for instance, Rockafellar (1976), Bertsekas (1982b, § 2.5)). In order to ensure that the method is globally convergent, as a last resort we must drive the penalty parameter to zero and ensure that the Lagrange multiplier estimates do not behave too badly. The convergence of such a scheme is guaranteed, since in this case, the iteration is essentially that used in the quadratic penalty function method (see, for example, Gould (1989)). We consider this further in § 4. In order to try to allow the traditional multiplier iteration to take over, the test on the size of the constraints (3.2)/(3.6) is based upon the size that might be expected if the multiplier iteration is converging. This aspect is considered in § 5.

The algorithms differ in their use of multiplier updates. Algorithm 1 is designed specifically for the first-order estimate (2.5); the multiplier estimates are encouraged to behave well as a consequence of the test (3.2). For large-scale computations, it is likely that first-order estimates will be used and thus Algorithm 1 is directly applicable. Algorithm 2 allows any multiplier estimate to be used. This extra freedom means that tighter control must be maintained on the acceptance of the estimates to make sure that they do not grow unacceptably fast. In this algorithm, we have in mind using any of the well-known Lagrange multiplier update formulae, including the first-order update (2.5) (used in Algorithm 1), the least-squares update (2.14), and other updates summarized, for instance, by Tapia (1977). We note, however, that some of these updates may require a significant amount of computation and this may prove prohibitively expensive for large-scale problems. Algorithm 2 is identical to Algorithm 1 except for the allowed Lagrange multiplier updates, the fact that these updates may also occur in step 3 and the presence of the scalars $\nu$ and $\gamma$.

Both algorithms use a number of free parameters. To give the reader some feel for what might be typical values, we suggest that for well-scaled problems $\alpha_\omega = \beta_\omega = \gamma = \nu = \eta_0 = \omega_0 = 1$, $\alpha_\eta = \mu_0 = \gamma_1 = 0.1$, $\beta_\eta = 0.9$, and $\tau = 0.01$ are appropriate.

**4. Global convergence analysis.** In this section we shall make use of the following assumptions:

(AS2) The iterates $\{x^{(k)}\}$ considered lie within a closed, bounded domain $\Omega$.

(AS3) The matrix $\hat{A}(x^*)$ has column rank no smaller than $m$ at any limit point, $x^*$, of the sequences $\{x^{(k)}\}$ considered in this paper.

Note that (AS3) excludes the possibility that $I_1$ is empty unless there are no general constraints. In view of Lemma 2.1, this seems reasonable as otherwise we are allowing the possibility that all the constraints and bounds are satisfied as equations at $x^*$. We also observe that (AS3) is equivalent to assuming that the gradients of the general constraints and active bounds at any limit point are linearly independent, an assumption that is commonly made in the analysis of other methods (see Bertsekas (1982b), and Fletcher (1981)).

We shall analyse the convergence of the algorithms of § 3 in the case where the convergence tolerances $\omega_*$ and $\eta_*$ are both zero. We require the following pair of lemmas in the proof of global convergence of our algorithms. Essentially, the results show that the Lagrange multiplier estimates generated by either algorithm cannot behave too badly.

LEMMA 4.1. *Suppose that $\mu^{(k)}$ converges to zero as k increases when Algorithm 1 is executed. Then the product $\mu^{(k)}\|\lambda^{(k)}\|$ converges to zero.*

*Proof.* If $\mu^{(k)}$ converges to zero, step 3 of the algorithm must be executed infinitely often. Let $K = \{k_0, k_1, k_2, \cdots\}$ be the set of the indices of the iterations in which step 3 of the algorithm is executed and for which

$$\mu^{(k)} \leqq \min\left(\left(\tfrac{1}{2}\right)^{1/\beta_\eta}, \gamma_1\right). \tag{4.1}$$

We consider how the Lagrange multiplier estimates change between two successive iterations indexed in the set $K$. At iteration $k_i + j$, for $k_i < k_i + j \leqq k_{i+1}$, we have

$$\lambda^{(k_i+j)} = \lambda^{(k_i)} + \sum_{l=1}^{j-1} S^{(k_i+l)} c(x^{(k_i+l)}) / \mu^{(k_i+l)} \tag{4.2}$$

and

$$\mu^{(k_{i+1})} = \mu^{(k_i+j)} = \mu^{(k_i+1)} = \tau\mu^{(k_i)}, \tag{4.3}$$

where the summation in (4.2) is null if $j = 1$. Now suppose that $j > 1$. Then for the set of iterations $k_i + l$, $1 \leqq l < j$, step 2 of the algorithm must have been executed and hence, from (3.2), (4.3), and the recursive definition of $\eta^{(k)}$, we must also have

$$\|c(x^{(k_i+l)})\| \leqq \eta_0 (\mu^{(k_{i+1})})^{\beta_\eta (l-1) + \alpha_\eta}. \tag{4.4}$$

Combining equations (4.1) to (4.4) and using the imposed upper bound on $S^{(k)}$, we obtain the bound

$$\begin{aligned}\|\lambda^{(k_i+j)}\| &\leqq \|\lambda^{(k_i)}\| + \sum_{l=1}^{j-1} \|S^{(k_i+l)}c(x^{(k_i+l)})\|/\mu^{(k_i+l)} \\ &\leqq \|\lambda^{(k_i)}\| + s_2\eta_0(\mu^{(k_{i+1})})^{\alpha_\eta-1}\sum_{l=1}^{j-1}(\mu^{(k_{i+1})})^{\beta_\eta(l-1)} \qquad (4.5)\\ &\leqq \|\lambda^{(k_i)}\| + s_2\eta_0(\mu^{(k_{i+1})})^{\alpha_\eta-1}/(1-(\mu^{(k_{i+1})})^{\beta_\eta}) \\ &\leqq \|\lambda^{(k_i)}\| + 2s_2\eta_0(\mu^{(k_{i+1})})^{\alpha_\eta-1},\end{aligned}$$

where $s_2$ is the norm of $S_2$. Thus we obtain that

$$(4.6) \qquad \mu^{(k_i+j)}\|\lambda^{(k_i+j)}\| \leqq \tau\mu^{(k_i)}\|\lambda^{(k_i)}\| + 2s_2\eta_0(\mu^{(k_{i+1})})^{\alpha_\eta}.$$

Equation (4.6) is also satisfied when $j=1$ as equations (3.4) and (4.3) give $\mu^{(k_i+1)}\|\lambda^{(k_i+1)}\| = \tau\mu^{(k_i)}\|\lambda^{(k_i)}\|$.

Hence from (4.6),

$$(4.7) \qquad \mu^{(k_{i+1})}\|\lambda^{(k_{i+1})}\| \leqq \tau\mu^{(k_i)}\|\lambda^{(k_i)}\| + 2s_2\eta_0(\mu^{(k_{i+1})})^{\alpha_\eta}.$$

Equation (4.7) then gives that $\mu^{(k_i)}\|\lambda^{(k_i)}\|$ converges to zero as $k$ increases. For, if we define

$$(4.8) \qquad \alpha_i = \mu^{(k_i)}\|\lambda^{(k_i)}\| \quad \text{and} \quad \beta_i = 2s_2\eta_0(\mu^{(k_i)})^{\alpha_\eta},$$

equations (4.3), (4.7), and (4.8) give that

$$(4.9) \qquad \alpha_{i+1} \leqq \tau\alpha_i + \tau^{\alpha_\eta}\beta_i \quad \text{and} \quad \beta_{i+1} = \tau^{\alpha_\eta}\beta_i$$

and hence that

$$(4.10) \qquad 0 \leqq \alpha_i \leqq \tau^i\alpha_0 + (\tau^{\alpha_\eta})^i \sum_{l=0}^{i-1} (\tau^{1-\alpha_\eta})^l\beta_0.$$

If $\alpha_\eta < 1$, the sum in (4.10) can be bounded to give

$$(4.11) \qquad 0 \leqq \alpha_i \leqq \tau^i\alpha_0 + (\tau^{\alpha_\eta})^i\beta_0/(1-\tau^{1-\alpha_\eta}),$$

whereas if $\alpha_\eta > 1$, we obtain the alternative

$$(4.12) \qquad 0 \leqq \alpha_i \leqq \tau^i(\alpha_0 + \tau^{\alpha_\eta-1}\beta_0/(1-\tau^{\alpha_\eta-1})),$$

and if $\alpha_\eta = 1$,

$$(4.13) \qquad 0 \leqq \alpha_i \leqq \tau^i\alpha_0 + i\tau^i\beta_0.$$

But, both $\alpha_0$ and $\beta_0$ are finite. Thus, as $i$ increases, $\alpha_i$ converges to zero; equation (4.9) implies that $\beta_i$ converges to zero. Therefore, as the right-hand side of (4.6) converges to zero, the truth of the lemma is established. □

We note that Lemmas 4.1 may be proved under much weaker conditions on the sequence $\{\eta^{(k)}\}$ than those imposed in Algorithm 1. All that is needed is that, in the proof just given, $\sum_{l=1}^{j-1}\|c(x^{(k_i+l)})\|$ in (4.5) should be bounded by some multiple of a positive power of $\mu^{(k_{i+1})}$.

Turning to Algorithm 2, we have the following easier-to-establish result.

LEMMA 4.2. *Suppose that* $\mu^{(k)}$ *converges to zero as k increases when Algorithm* 2 *is executed. Then the product* $\mu^{(k)}\|\lambda^{(k)}\|$ *converges to zero.*

*Proof.* Let $K=\{k_0, k_1, k_2, \cdots\}$ be the iterations on which

$$\|\hat{\lambda}^{(k+1)}\| \leqq \nu(\mu^{(k+1)})^{-\gamma} \tag{4.14}$$

and consequently on which $\lambda^{(k+1)}=\hat{\lambda}^{(k+1)}$. Then, from (4.14),

$$\mu^{(k_i+1)}\|\lambda^{(k_i+1)}\| \leqq \nu(\mu^{(k_i+1)})^{1-\gamma}. \tag{4.15}$$

If $K$ is finite, $\lambda^{(k)}$ will be fixed for all $k$ sufficiently large and the result is immediate. If $K$ is infinite, for any $k_i < k \leqq k_{i+1}$, $\lambda^{(k)}=\lambda^{(k_i+1)}$, and $\mu^{(k)} \leqq \mu^{(k_i+1)}$. Hence, from (4.15)

$$\mu^{(k)}\|\lambda^{(k)}\| \leqq \nu(\mu^{(k_i+1)})^{1-\gamma}. \tag{4.16}$$

By hypothesis, the right-hand side of (4.16) can be made arbitrarily small by choosing $k_i$ large enough, and so $\mu^{(k)}\|\lambda^{(k)}\|$ converges to zero. □

As a precursor to our main result, we have the following general convergence lemma. The lemma and resulting theorem are in the spirit of Proposition 2.3 of Bertsekas (1982b) but do not require that the Lagrange multiplier estimates stay bounded and allow for our handling of simple bound constraints.

LEMMA 4.3. *Suppose that* (AS1) *holds. Let* $\{x^{(k)}\} \in B$, $k \in K$, *be a sequence satisfying* (AS2) *which converges to the point* $x^*$ *for which* (AS3) *holds and let* $\lambda^* = \lambda(x^*)$, *where* $\lambda$ *satisfies* (2.14). *Assume that* $\{\lambda^{(k)}\}$, $k \in K$, *is any sequence of vectors, that* $\{S^{(k)}\}$, $k \in K$, *is any sequence of diagonal matrices satisfying* $0 < S_1^{-1} \leqq S^{(k)} \leqq S_2 < \infty$, *and that* $\{\mu^{(k)}\}$, $k \in K$, *form a nonincreasing sequence of positive scalars. Suppose further that*

$$\|P(x^{(k)}, \nabla_x\Phi^{(k)})\| \leqq \omega^{(k)} \tag{4.17}$$

*where the* $\omega^{(k)}$ *are positive scalar parameters which converge to zero as* $k \in K$ *increases. Then*

(i) *There are positive constants* $a_1$, $a_2$, $s_1$ *and an integer* $k_0$ *such that*

$$\|\bar{\lambda}(x^{(k)}, \lambda^{(k)}, S^{(k)}, \mu^{(k)}) - \lambda^*\| \leqq a_1\omega^{(k)} + a_2\|x^{(k)}-x^*\|, \tag{4.18}$$

$$\|\lambda(x^{(k)}) - \lambda^*\| \leqq a_2\|x^{(k)}-x^*\|, \tag{4.19}$$

*and*

$$\|c(x^{(k)})\| \leqq s_1(a_1\omega^{(k)}\mu^{(k)} + \mu^{(k)}\|\lambda^{(k)}-\lambda^*\| + a_2\mu^{(k)}\|x^{(k)}-x^*\|) \tag{4.20}$$

*for all* $k \geqq k_0$, $(k \in K)$.

*Suppose, in addition, that* $c(x^*)=0$. *Then*

(ii) $x^*$ *is a Kuhn–Tucker point* (*first-order stationary point*) *for the problem* (1.5)–(1.7), $\lambda^*$ *is the corresponding vector of Lagrange multipliers, and the sequences* $\{\bar{\lambda}(x^{(k)}, \lambda^{(k)}, S^{(k)}, \mu^{(k)})\}$ *and* $\{\lambda(x^{(k)})\}$ *converge to* $\lambda^*$ *for* $k \in K$;

(iii) *The gradients* $\nabla_x\Phi^{(k)}$ *converge to* $g_L(x^*, \lambda^*)$ *for* $k \in K$.

*Proof.* As a consequence of (AS1)–(AS3), we have that for $k \in K$ sufficiently large, $\hat{A}(x^{(k)})^+$ exists, is bounded, and converges to $\hat{A}(x^*)^+$. Thus we may write

$$\|(\hat{A}(x^{(k)})^+)^T\| \leqq a_1 \tag{4.21}$$

for some constant $a_1 > 0$. As the variables in the set $I_1$ are floating, equations (2.7), (2.8), (2.11), and the inner iteration termination criterion (4.17) give that

$$\|\hat{g}(x^{(k)}) + \hat{A}(x^{(k)})^T\bar{\lambda}^{(k)}\| \leqq \omega^{(k)}. \tag{4.22}$$

By assumption, $\lambda(x)$ is bounded for all $x$ in a neighbourhood of $x^*$. Thus we may deduce from (2.14), (4.21), and (4.22) that

$$\begin{aligned}\|\bar{\lambda}^{(k)} - \lambda(x^{(k)})\| &= (\hat{A}(x^{(k)})^+)^T\hat{g}(x^{(k)}) + \bar{\lambda}^{(k)}\| \\ &= \|(\hat{A}(x^{(k)})^+)^T(\hat{g}(x^{(k)}) + \hat{A}(x^{(k)})^T\bar{\lambda}^{(k)})\| \\ &\leqq \|(\hat{A}(x^{(k)})^+)^T\|\omega^{(k)} \leqq a_1\omega^{(k)}.\end{aligned} \tag{4.23}$$

Moreover, from the integral mean value theorem and Lemma 2.2 we have that

$$\lambda(x^{(k)})-\lambda(x^*)=\int_0^1 \nabla_x\lambda(x(s))\,ds\cdot(x^{(k)}-x^*), \tag{4.24}$$

where $\nabla_x\lambda(x)$ is given by equation (2.16) and where $x(s)=x^{(k)}+s(x^*-x^{(k)})$. Now the terms within the integral sign are bounded for all $x$ sufficiently close to $x^*$ and hence (4.24) gives

$$\|\lambda(x^{(k)})-\lambda^*\|\leqq a_2\|x^{(k)}-x^*\| \tag{4.25}$$

for some constant $a_2>0$, which is just the inequality (4.19). We then have that $\lambda(x^{(k)})$ converges to $\lambda^*$. Combining (4.23) and (4.25) we obtain

$$\|\bar\lambda^{(k)}-\lambda^*\|\leqq\|\bar\lambda^{(k)}-\lambda(x^{(k)})\|+\|\lambda(x^{(k)})-\lambda^*\|\leqq a_1\omega^{(k)}+a_2\|x^{(k)}-x^*\|, \tag{4.26}$$

the required inequality (4.18). Then, since by construction $\omega^{(k)}$ tends to zero as $k$ increases, (4.18) implies that $\bar\lambda^{(k)}$ converges to $\lambda^*$ and from (4.22) we have that

$$\hat g_L(x^*,\lambda^*)=\hat g(x^*)+\hat A(x^*)^T\lambda^*=0. \tag{4.27}$$

Moreover, from the identity (2.6), $\nabla_x\Phi^{(k)}$ converges to $g_L(x^*,\lambda^*)$. Furthermore, multiplying (2.5) by $\mu^{(k)}$, we obtain

$$c(x^{(k)})=\mu^{(k)}S^{(k)-1}((\bar\lambda^{(k)}-\lambda^*)+(\lambda^*-\lambda^{(k)})). \tag{4.28}$$

Taking norms of (4.28) and using (4.26) we derive (4.20), where $s_1$ is the norm of $S_1$.

Now suppose that

$$c(x^*)=0 \tag{4.29}$$

and consider the status of the variables in the sets $I_1$, $I_2$, and $I_5$. Lemma 2.1 and the convergence of $\nabla_x\Phi^{(k)}$ to $g_L(x^*,\lambda^*)$ show that the complementary slackness condition

$$g_L(x^*,\lambda^*)^Tx^*=0 \tag{4.30}$$

is satisfied. The variables in the set $I_1$ are, by definition, positive at $x^*$. The components of $g_L(x^*,\lambda^*)$ indexed by $I_2$ are all nonnegative from (2.9) as their corresponding variables are dominated. This then gives the conditions

$$\begin{array}{llll} x_i^*>0 & \text{and} \quad (g_L(x^*,\lambda^*))_i=0 & \text{for } i\in I_1, & \\ x_i^*=0 & \text{and} \quad (g_L(x^*,\lambda^*))_i\geqq 0 & \text{for } i\in I_2, & \text{and} \\ x_i^*=0 & \text{and} \quad (g_L(x^*,\lambda^*))_i=0 & \text{for } i\in I_5. & \end{array} \tag{4.31}$$

Equations (4.29) and (4.31) thus show that $x^*$ is a Kuhn–Tucker point and $\lambda^*$ are the corresponding set of Lagrange multipliers. Moreover, (4.18) and (4.19) ensure the convergence of the sequences $\{\bar\lambda(x^{(k)},\lambda^{(k)},S^{(k)},\mu^{(k)})\}$ and $\{\lambda(x^{(k)})\}$ to $\lambda^*$ for $k\in K$. Hence the lemma is proved. □

We now show that both Algorithms 1 and 2 possess a powerful global convergence property under relatively weak conditions.

THEOREM 4.4. *Assume that* (AS1) *holds. Let $x^*$ be any limit point of the sequence $\{x^{(k)}\}$ generated by Algorithm* 1 *or by Algorithm* 2 *of* § 3 *for which* (AS2) *and* (AS3) *hold and let $K$ be the set of indices of an infinite subsequence of the $x^{(k)}$ whose limit is $x^*$. Then conclusions* (i), (ii), *and* (iii) *of Lemma* 4.3 *hold.*

*Proof.* The assumptions given are sufficient to reach the conclusions of part (i) of Lemma 4.3. We now show that (4.29) holds for Algorithms 1 and 2. To see this, we consider two separate cases:

(i) If $\mu^{(k)}$ is bounded away from zero, step 2 must be executed every iteration for $k$ sufficiently large. But this implies that (3.2) is always satisfied ($k$ large enough) and $\eta^{(k)}$ converges to zero.

Hence $c(x^{(k)})$ converges to zero.

(ii) If $\mu^{(k)}$ converges to zero, Lemma 4.1 for Algorithm 1 and Lemma 4.2 for Algorithm 2 show that $\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|$ converges to zero. But then, inequality (4.20) gives the required result.

Hence (4.29) is satisfied and thus conclusions (ii) and (iii) of Lemma 4.3 holds. □

Note that Theorem 4.4 would remain true regardless of the actual choices of $\{\omega^{(k)}\}$ and $\{\eta^{(k)}\}$ provided that both sequences converge to zero.

**5. Asymptotic convergence analysis.** We now give our first rate-of-convergence result. It is inconvenient that the estimates (4.18)-(4.20) depend upon $\|x^{(k)}-x^*\|$. The next lemma removes this dependence and gives a result similar to the classical theory in which the errors in $x$ are bounded by the errors in the multiplier estimates $\|\lambda^{(k)}-\lambda^*\|$ (see Bertsekas (1982b, Prop. 2.4)); however, as an inexact minimization of the augmented Lagrangian function is made, in the spirit of Bertsekas (1982b, Prop. 2.14)), a term reflecting this is also present in the bound. Once again, the result allows for our handling of simple bound constraints. Before giving our result, we need to make two additional assumptions.

We use the notation that, if $J_1$ and $J_2$ are any subsets of $N$, $H_L(x^*,\lambda^*)_{[J_1,J_2]}$ is the matrix formed by taking the *rows and columns* of $H_L(x^*,\lambda^*)$ indexed by $J_1$ and $J_2$, respectively, and $A(x^*)_{[J_1]}$ is the matrix formed by taking the *columns* of $A(x^*)$ indexed by $J_1$. We use the following assumptions:

(AS4) The second derivatives of the functions $f(x)$ and the $c_i(x)$ are Lipschitz continuous at all points within $\Omega$.

(AS5) Suppose that $(x^*,\lambda^*)$ is a Kuhn-Tucker point for problem (1.5)-(1.7) and that

$$\begin{aligned} J_1&=\{i\,|\,(g_L(x^*,\lambda^*))_i=0 \quad\text{and}\quad x_i^*>0\},\\ J_2&=\{i\,|\,(g_L(x^*,\lambda^*))_i=0 \quad\text{and}\quad x_i^*=0\}.\end{aligned} \tag{5.1}$$

Then we assume that the matrix

$$\begin{bmatrix} H_L(x^*,\lambda^*)_{[J,J]} & (A(x^*)_{[J]})^T \\ A(x^*)_{[J]} & 0 \end{bmatrix}$$

is nonsingular for all sets $J$, where $J$ is any set made up from the union of $J_1$ and any subset of $J_2$.

We note that assumption (AS5) implies (AS3). Furthermore, if $J_2$ is empty, any point satisfying the well-known second-order sufficiency condition for a minimizer of (1.5)-(1.7) (see, e.g., Fletcher (1981, Thm. 9.3.2)) automatically satisfies (AS5) (see, e.g., Gould (1985)). When $J_2$ is nonempty, the connection between (AS5) and Fletcher's condition is less clear, although (AS5) is certainly implied by the stronger second-order sufficiency condition given by Luenberger (1973, pp. 234-235). We believe, however, our assumption is reasonable in that small perturbations to the problem can cause some elements of $J_2$ to defect to $J_1$ while others may drop entirely from $J$ as their gradient components become positive. As we do not know which might defect under such perturbations, (AS5) is a form of "insurance" against all possible eventualities.

LEMMA 5.1. *Suppose that* (AS1) *holds. Let* $\{x^{(k)}\}\in B$, $k\in K$, *be a subsequence which converges to the Kuhn-Tucker point* $x^*$ *for which* (AS2), (AS4), *and* (AS5) *hold, and* *let $\lambda^*$ be the corresponding vector of Lagrange multipliers. Assume that $\{\lambda^{(k)}\}$, $k \in K$, is any sequence of vectors, that $\{S^{(k)}\}$, $k \in K$, is any sequence of diagonal matrices satisfying $0 < S_1^{-1} \leqq S^{(k)} \leqq S_2 < \infty$, and that $\{\mu^{(k)}\}$, $k \in K$, form a nonincreasing sequence of positive scalars, so that the product $\mu^{(k)}\|\lambda^{(k)} - \lambda^*\|$ converges to zero as k increases. Now, suppose further that*

$$\|P(x^{(k)}, \nabla_x \Phi^{(k)})\| \leqq \omega^{(k)}, \tag{5.2}$$

*where the $\omega^{(k)}$ are positive scalar parameters which converge to zero as $k \in K$ increases. Then there are positive constants $\bar{\mu}$, $a_3$, $a_4$, $a_5$, $a_6$, and $s_1$ and an integer value $k_0$ so that if $\mu^{(k_0)} \leqq \bar{\mu}$ then*

$$\|x^{(k)} - x^*\| \leqq a_3 \omega^{(k)} + a_4 \mu^{(k)} \|\lambda^{(k)} - \lambda^*\|, \tag{5.3}$$

$$\|\bar{\lambda}(x^{(k)}, \lambda^{(k)}, S^{(k)}, \mu^{(k)}) - \lambda^*\| \leqq a_5 \omega^{(k)} + a_6 \mu^{(k)} \|\lambda^{(k)} - \lambda^*\|, \tag{5.4}$$

*and*

$$\|c(x^{(k)})\| \leqq s_1(a_5 \omega^{(k)} \mu^{(k)} + (\mu^{(k)} + a_6(\mu^{(k)})^2)\|\lambda^{(k)} - \lambda^*\|) \tag{5.5}$$

*for all $k \geqq k_0$, $(k \in K)$.*

*Proof.* We will denote the gradient and Hessian of the Lagrangian function, taken with respect to $x$, at the limit point $(x^*, \lambda^*)$ by $g_L^*$ and $H_L^*$, respectively.

We first need to make some observations concerning the status of the variables as the limit point is approached. We pick $k$ sufficiently large that the sets $I_1$ and $I_2$, defined in (2.12), have been determined. Then, for $k \in K$, the remaining variables either float (variables in $I_3$) or oscillate between floating and being dominated (variables in $I_4$). Now recall the definition (2.13) of $I_5$ and pick an infinite subsequence, $\bar{K}$ of $K$ such that:

(i) $I_5 = I_6 \cup I_7$ with $I_6 \cap I_7 = \varnothing$;

(ii) Variables in $I_6$ are floating for all $k = \bar{K}$; and

(iii) Variables in $I_7$ are dominated for all $k \in \bar{K}$.

Note that the set $I_3$ of (2.12) is contained within $I_6$. Note, also, that there are only a finite number $(\leqq 2^{|I_5|})$ of such subsequences $\bar{K}$ and that for $k$ sufficiently large, each $k \in K$ is in one such subsequence. It is thus sufficient to prove the lemma for $k \in \bar{K}$.

Now, for $k \in \bar{K}$, define

$$I_F = I_1 \cup I_6 \quad \text{and} \quad I_D = I_2 \cup I_7. \tag{5.6}$$

So, the variables in $I_F$ are floating while those in $I_D$ are dominated. We may now invoke Lemma 4.3(i) to obtain inequalities (4.18) and (4.20) for some $\lambda^*$. Furthermore, using (4.20) and the current assumption that $\mu^{(k)}\|\lambda^{(k)} - \lambda^*\|$ converges to zero as $k$ increases, we have that $c(x^*) = 0$. Thus Lemma 4.3(ii) implies that $\lambda^*$ is the vector of Lagrange multipliers corresponding to $x^*$. We thus have

$$\|\bar{\lambda}^{(k)} - \lambda^*\| \leqq a_1 \omega^{(k)} + a_2 \|x^{(k)} - x^*\| \tag{5.7}$$

and

$$\|c(x^{(k)})\| \leqq s_1(a_1 \omega^{(k)} \mu^{(k)} + \mu^{(k)} \|\lambda^{(k)} - \lambda^*\| + a_2 \mu^{(k)} \|x^{(k)} - x^*\|) \tag{5.8}$$

for all sufficiently large $k \in \bar{K}$ from inequalities (4.18) and (4.20). Moreover, $\bar{\lambda}^{(k)}$ converges to $\lambda^*$ and hence $\nabla_x \Phi^{(k)}$ converges to $g_L^*$. Therefore, from Lemma 2.1,

$$x_i^* = 0 \quad \text{for all } i \in I_D \quad \text{and} \quad (g_L^*)_i = 0 \quad \text{for all } i \in I_F. \tag{5.9}$$

Using Taylor's theorem,

$$\begin{aligned}\nabla_x\Phi^{(k)} &= g^{(k)}+A^{(k)T}\bar{\lambda}^{(k)}\\ &= g(x^*)+H(x^*)(x^{(k)}-x^*)+A(x^*)^T\bar{\lambda}^{(k)}\\ &\quad+\sum_{j=1}^{m}\bar{\lambda}_j^{(k)}H_j(x^*)(x^{(k)}-x^*)+r_1(x^{(k)},x^*,\bar{\lambda}^{(k)})\\ &= g_L(x^*,\lambda^*)+H_L(x^*,\lambda^*)(x^{(k)}-x^*)+A(x^*)^T(\bar{\lambda}^{(k)}-\lambda^*)\\ &\quad+r_1(x^{(k)},x^*,\bar{\lambda}^{(k)})+r_2(x^{(k)},x^*,\bar{\lambda}^{(k)},\lambda^*),\end{aligned} \tag{5.10}$$

where

$$r_1(x^{(k)},x^*,\bar{\lambda}^{(k)})=\int_0^1(H_L(x^{(k)}+s(x^*-x^{(k)}),\bar{\lambda}^{(k)})-H_L(x^*,\bar{\lambda}^{(k)}))(x^{(k)}-x^*)\,ds \tag{5.11}$$

and

$$r_2(x^{(k)},x^*,\bar{\lambda}^{(k)},\lambda^*)=\sum_{j=1}^{m}(\bar{\lambda}_j^{(k)}-\lambda_j^*)H_j(x^*)(x^{(k)}-x^*). \tag{5.12}$$

The boundedness and Lipschitz continuity of the Hessian matrices of $f$ and $c_i$ in a neighbourhood of $x^*$ along with the convergence of $\bar{\lambda}^{(k)}$ to $\lambda^*$ then give that

$$\begin{aligned}\|r_1(x^{(k)},x^*,\bar{\lambda}^{(k)})\| &\leqq a_7\|x^{(k)}-x^*\|^2,\\ \|r_2(x^{(k)},x^*,\bar{\lambda}^{(k)},\lambda^*)\| &\leqq a_8\|x^{(k)}-x^*\|\,\|\bar{\lambda}^{(k)}-\lambda^*\|\end{aligned} \tag{5.13}$$

for some positive constants $a_7$ and $a_8$. In addition, again using Taylor's theorem and that $c(x^*)=0$,

$$c(x^{(k)})=A(x^*)(x^{(k)}-x^*)+r_3(x^{(k)},x^*), \tag{5.14}$$

where

$$(r_3(x^{(k)},x^*))_i=\int_0^1 s\int_0^1(x^{(k)}-x^*)^TH_i(x^*+ts(x^{(k)}-x^*))(x^{(k)}-x^*)\,dt\,ds \tag{5.15}$$

(see Gruver and Sachs (1980, p. 11)). The boundedness of the Hessian matrices of the $c_i$ in a neighbourhood of $x^*$ then gives that

$$\|r_3(x^{(k)},x^*)\|\leqq a_9\|x^{(k)}-x^*\|^2 \tag{5.16}$$

for some constant $a_9>0$. Combining (5.10) and (5.14), we obtain

$$\begin{pmatrix}H_L(x^*,\lambda^*) & A^T(x^*)\\ A(x^*) & 0\end{pmatrix}\begin{pmatrix}x^{(k)}-x^*\\ \bar{\lambda}^{(k)}-\lambda^*\end{pmatrix}=\begin{pmatrix}\nabla_x\Phi^{(k)}-g_L(x^*,\lambda^*)\\ c(x^{(k)})\end{pmatrix}-\begin{pmatrix}r_1+r_2\\ r_3\end{pmatrix}, \tag{5.17}$$

where we have suppressed the arguments of $r_1$, $r_2$ and $r_3$ for brevity. To proceed further, we introduce the notation that $y_{[J]}$ is the vector formed by taking the components of the vector $y$ indexed by the set $J$. We may then rewrite (5.17) as

$$\begin{aligned}&\begin{pmatrix}H_L(x^*,\lambda^*)_{[I_F,I_F]} & H_L(x^*,\lambda^*)_{[I_F,I_D]} & A^T(x^*)_{[I_F]}\\ H_L(x^*,\lambda^*)_{[I_D,I_F]} & H_L(x^*,\lambda^*)_{[I_D,I_D]} & A^T(x^*)_{[I_D]}\\ A(x^*)_{[I_F]} & A(x^*)_{[I_D]} & 0\end{pmatrix}\begin{pmatrix}(x^{(k)}-x^*)_{[I_F]}\\ (x^{(k)})_{[I_D]}\\ \bar{\lambda}^{(k)}-\lambda^*\end{pmatrix}\\ &\quad=\begin{pmatrix}(\nabla_x\Phi^{(k)})_{[I_F]}\\ (\nabla_x\Phi^{(k)}-g_L(x^*,\lambda^*))_{[I_D]}\\ c(x^{(k)})\end{pmatrix}-\begin{pmatrix}(r_1+r_2)_{[I_F]}\\ (r_1+r_2)_{[I_D]}\\ r_3\end{pmatrix}\end{aligned} \tag{5.18}$$

using (5.9). Then, rearranging (5.18) and removing the middle horizontal block we obtain

$$\begin{aligned}&\begin{pmatrix} H_L(x^*,\lambda^*)_{[I_F,I_F]} & A^T(x^*)_{[I_F]} \\ A(x^*)_{[I_F]} & 0 \end{pmatrix}\begin{pmatrix}(x^{(k)}-x^*)_{[I_F]} \\ \bar\lambda^{(k)}-\lambda^*\end{pmatrix} \\ &\quad=\begin{pmatrix}(\nabla_x\Phi^{(k)})_{[I_F]}-H_L(x^*,\lambda^*)_{[I_F,I_D]}(x^{(k)})_{[I_D]} \\ c(x^{(k)})-A(x^*)_{[I_D]}(x^{(k)})_{[I_D]}\end{pmatrix}-\begin{pmatrix}(r_1+r_2)_{[I_F]} \\ r_3\end{pmatrix}.\end{aligned} \tag{5.19}$$

Roughly, the rest of the proof proceeds by showing that thet the right-hand side of (5.19) is $O(\omega^{(k)})+O(\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|)$. This will then ensure that the vector on the left-hand side is of the same size, which is the result we require. First, observe that

$$\|x^{(k)}_{[I_D]}\|\leqq\omega^{(k)}, \tag{5.20}$$

from (2.10) and (5.2) and

$$\|(\nabla_x\Phi^{(k)})_{[I_F]}\|\leqq\omega^{(k)}, \tag{5.21}$$

from (2.11). Consequently, again using (5.9),

$$\|x^{(k)}-x^*\|\leqq\|(x^{(k)}-x^*)_{[I_F]}\|+\omega^{(k)}. \tag{5.22}$$

Let $\Delta x^{(k)}=\|(x^{(k)}-x^*)_{[I_F]}\|$. Combining (5.7) and (5.22), we obtain

$$\|\bar\lambda^{(k)}-\lambda^*\|\leqq a_{10}\omega^{(k)}+a_2\Delta x^{(k)}, \tag{5.23}$$

where $a_{10}=a_1+a_2$. Furthermore, from (5.13), (5.16), (5.22), and (5.23),

$$\left\|\begin{pmatrix}(r_1+r_2)_{[I_F]} \\ r_3\end{pmatrix}\right\|\leqq a_{11}(\Delta x^{(k)})^2+a_{12}\Delta x^{(k)}\omega^{(k)}+a_{13}(\omega^{(k)})^2, \tag{5.24}$$

where $a_{11}=a_7+a_9+a_8a_2$, $a_{12}=2(a_7+a_9)+a_8(a_{10}+a_2)$, and $a_{13}=a_7+a_9+a_8a_{10}$. Moreover, from (5.8), (5.20), (5.21), and (5.22),

$$\begin{aligned}&\left\|\begin{pmatrix}(\nabla_x\Phi^{(k)})_{[I_F]}-H_L(x^*,\lambda^*)_{[I_F,I_D]}(x^{(k)})_{[I_D]} \\ c(x^{(k)})-A(x^*)_{[I_D]}(x^{(k)})_{[I_D]}\end{pmatrix}\right\| \\ &\quad\leqq a_{14}\omega^{(k)}+s_1(\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|+a_{10}\omega^{(k)}\mu^{(k)}+a_2\mu^{(k)}\Delta x^{(k)}),\end{aligned} \tag{5.25}$$

where

$$a_{14}=1+\left\|\begin{pmatrix}H_L(x^*,\lambda^*)_{[I_F,I_D]} \\ A(x^*)_{[I_D]}\end{pmatrix}\right\|. \tag{5.26}$$

By assumption (AS5), the coefficient matrix on the left-hand side of (5.19) is nonsingular. Let its inverse have norm $M$. Multiplying both sides of the equation by this inverse and taking norms, we obtain

$$\begin{aligned}\left\|\begin{pmatrix}x^{(k)}-x^*)_{[I_F]} \\ \bar\lambda^{(k)}-\lambda^*\end{pmatrix}\right\|&\leqq M[a_{14}\omega^{(k)}+s_1(\mu^{(k)}\|\lambda^{(k)}-\lambda^*\| \\ &\qquad+a_{10}\omega^{(k)}\mu^{(k)}+a_2\mu^{(k)}\Delta x^{(k)}) \\ &\qquad+a_{11}(\Delta x^{(k)})^2+a_{12}\Delta x^{(k)}\omega^{(k)}+a_{13}(\omega^{(k)})^2].\end{aligned} \tag{5.27}$$

Now, suppose that $k$ is sufficiently large that

$$\omega^{(k)}\leqq\min(1,1/(4Ma_{12})). \tag{5.28}$$

Furthermore, let

$$\bar\mu=\min(1,1/(4Ma_2s_1)). \tag{5.29}$$

Then, if $\mu^{(k)} \leqq \bar{\mu}$, (5.27)-(5.29) give

$$\Delta x^{(k)} \leqq \tfrac{1}{2}\Delta x^{(k)} + M(a_{15}\omega^{(k)} + s_1\mu^{(k)}\|\lambda^{(k)} - \lambda^*\| + a_{11}(\Delta x^{(k)})^2), \tag{5.30}$$

where $a_{15} = s_1 a_{10} + a_{13} + a_{14}$. As $\Delta x^{(k)}$ converges to zero, we have that

$$\|\Delta x^{(k)}\| \leqq 1/(4Ma_{11}) \tag{5.31}$$

for all $k$ sufficiently large. Hence inequalities (5.30) and (5.31) give that

$$\Delta x^{(k)} \leqq 4M(a_{15}\omega^{(k)} + s_1\mu^{(k)}\|\lambda^{(k)} - \lambda^*\|). \tag{5.32}$$

Writing $a_3 = 4Ma_{15} + 1$ and $a_4 = 4Ms_1$, we obtain the desired inequality (5.3) from (5.22) and (5.32). Now, using (5.3) and (5.7), we obtain (5.4), where $a_5 = a_1 + a_2 a_3$ and $a_6 = a_2 a_4$. Finally, (5.5) follows from (5.4) by substituting for $\bar{\lambda}^{(k)}$, using (2.5), and multiplying the inequality by $\mu^{(k)}$. □

We can obtain the following simple corollary.

COROLLARY 5.2. *Suppose that the conditions of Lemma* 5.1 *hold and that* $\hat{\lambda}^{(k+1)}$ *is any Lagrange multiplier estimate for which*

$$\|\hat{\lambda}^{(k+1)} - \lambda^*\| \leqq a_{16}\|x^{(k)} - x^*\| + a_{17}\omega^{(k)} \tag{5.33}$$

*for some positive constants* $a_{16}$ *and* $a_{17}$ *and all* $k \in K$ *sufficiently large. Then there are positive constants* $\bar{\mu}$, $a_3$, $a_4$, $a_5$, $a_6$, $s_1$ *and an integer value* $k_0$ *so that if* $\mu^{(k_0)} \leqq \bar{\mu}$ *then* (5.3),

$$\|\hat{\lambda}^{(k+1)} - \lambda^*\| \leqq a_5\omega^{(k)} + a_6\mu^{(k)}\|\lambda^{(k)} - \lambda^*\|, \tag{5.34}$$

*and* (5.5) *hold for all* $k \geqq k_0$, $(k \in K)$.

*Proof.* Inequality (5.34) follows immediately from (5.33) and (5.3). □

We now show that the penalty parameter will normally be bounded away from zero in both Algorithms 1 and 2. This is important as many methods for solving the inner iteration subproblem will encounter difficulties if the parameter converges to zero since this causes the Hessian of the augmented Lagrangian to become increasingly ill-conditioned.

THEOREM 5.3. *Suppose that the iterates* $\{x^{(k)}\}$ *of Algorithm* 1 *or* 2 *of* § 3 *converges to the single limit point* $x^*$, *that* (AS1), (AS2), (AS4), *and* (AS5) *hold, that* $\alpha_\eta$ *and* $\beta_\eta$ *satisfy*

$$\alpha_\eta < \alpha \equiv \min(1, \alpha_\omega), \tag{5.35}$$

$$\beta_\eta < \min(1, \beta_\omega), \tag{5.36}$$

*and that* (5.33) *holds for all k sufficiently large when Algorithm* 2 *is used. Then there is a constant* $\mu > 0$ *such that* $\mu^{(k)} \geqq \mu$ *for all k.*

*Proof.* Suppose, otherwise, that $\mu^{(k)}$ tends to zero. Then, step 3 of the algorithm must be executed infinitely often. We aim to obtain a contradiction to this statement by showing that step 2 is always executed for $k$ sufficiently large. We note that our assumptions are sufficient for the conclusions of Theorem 4.4 to hold.

First, we show that the sequence of Lagrange multipliers $\{\lambda^{(k)}\}$ converges to $\lambda^*$.

Consider Algorithm 1. The result is clear if step 2 is executed infinitely often as each time the step is executed, $\lambda^{(k+1)} = \bar{\lambda}^{(k)}$ and the inequality (4.18) guarantees that $\bar{\lambda}^{(k)}$ converges to $\lambda^*$. Suppose that step 2 is not executed infinitely often. Then $\|\lambda^{(k)} - \lambda^*\|$ will remain fixed for all $k \geqq k_1$ for some $k_1$, as step 3 is executed for each remaining iteration. But then (4.20) implies that $\|c(x^{(k)})\| \leqq a_{17}\mu^{(k)}$ for some constant $a_{17}$ for all $k \geqq k_2 \geqq k_1$. As $\mu^{(k)}$ converges to zero as $k$ increases, $a_{17}\mu^{(k)} \leqq \eta_0(\mu^{(k)})^{\alpha_\eta} = \eta^{(k)}$ for all $k$ sufficiently large. But then inequality (3.2) must be satisfied for some $k \geqq k_1$, which is impossible, as this would imply that step 2 is again executed. Hence, step 2 must be executed infinitely often.

Now consider Algorithm 2. The result is clear if the multiplier updates are accepted infinitely often, as each time the update is performed $\lambda^{(k+1)}=\hat{\lambda}^{(k+1)}$ and assumption (5.33) guarantees that $\hat{\lambda}^{(k+1)}$ converges to $\lambda^*$. Suppose that the update is not accepted infinitely often. Then for all $k$ sufficiently large, $\|\hat{\lambda}^{(k+1)}\|>\nu(\mu^{(k+1)})^{-\gamma}$ which implies that $\|\hat{\lambda}^{(k+1)}\|$ diverges. But this contradicts assumption (5.33) and hence $\lambda^{(k)}$ converges to $\lambda^*$.

Therefore $\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|$ tends to zero as $k$ increases for both algorithms.

Let $k_1$ be the smallest integer for which

$$\mu^{(k)}\leqq\gamma_1<1 \tag{5.37}$$

for all $k\geqq k_1$. Now let $\omega^{(k)}$ be as generated by either algorithm. Note that, by construction and inequality (5.37),

$$\omega^{(k)}\leqq\omega_0(\mu^{(k)})^{\alpha_\omega} \tag{5.38}$$

for all $k\geqq k_1$. (This follows by definition if step 2 of either algorithm occurs and because the penalty parameter is unchanged while $\omega^{(k)}$ is reduced when step 3 occurs.) We shall apply Lemma 5.1 or Corollary 5.2 to the iterates generated by the algorithm; we identify the set $K$ with the complete set of integers larger than $k_1$ and the scalars $\mu^{(k)}$ with the set of penalty parameters computed in steps 2 and 3 of either algorithm. Therefore we can ensure that $\mu^{(k)}$ is sufficiently small so that Lemma 5.1 applies to step 1 of Algorithm 1 (or Corollary 5.2 to step 1 of Algorithm 2) and thus that there is an integer $k_2$ and constants $a_5$, $a_6$, and $s_1$ so that (5.4)/(5.34) and (5.5) hold for all $k\geqq k_2$. Let $k_3$ be the smallest integer such that

$$(\mu^{(k)})^{1-\alpha_\eta}\leqq\frac{\eta_0}{\omega_0 s_1(a_5+2)}, \tag{5.39}$$

$$(\mu^{(k)})^{1-\beta_\eta}\leqq\min\left(\frac{1}{a_{18}},\frac{\eta_0}{\omega_0 s_1(a_5+2a_{18})}\right), \tag{5.40}$$

and, if Algorithm 2 is used,

$$(\mu^{(k)})^{\gamma}\leqq\frac{\nu}{\|\lambda^*\|+\omega_0 a_{18}}, \tag{5.41}$$

where $a_{18}=a_5+a_6$. Note that (5.37) and (5.40) imply that

$$\mu^{(k)}\leqq(\mu^{(k)})^{1-\beta_\eta}\leqq\frac{1}{a_{18}}\leqq\frac{1}{a_6} \tag{5.42}$$

for all $k\geqq k_3$. Furthermore, let $k_4$ be such that

$$\|\lambda^{(k)}-\lambda^*\|\leqq\omega_0 \tag{5.43}$$

for all $k\geqq k_4$. Now define $k_5=\max(k_1,k_2,k_3,k_4)$, let $\Gamma$ be the set $\{k|$ Step 3 is executed at iteration $k-1$ and $k\geqq k_5\}$ and let $k_0$ be the smallest element of $\Gamma$. By assumption, $\Gamma$ has an infinite number of elements.

If Algorithm 2 is used, inequality (5.34) gives that

$$\begin{aligned}\|\hat{\lambda}^{(k+1)}\|&\leqq\|\lambda^*\|+a_5\omega^{(k)}+a_6\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|\\&\leqq\|\lambda^*\|+a_5\omega_0(\mu^{(k)})^{\alpha_\omega}+a_6\mu^{(k)}\|\lambda^{(k)}-\lambda^*\|\quad\text{(from (5.38))}\\&\leqq\|\lambda^*\|+\omega_0(a_5(\mu^{(k)})^{\alpha_\omega}+a_6\mu^{(k)})\quad\text{(from (5.43))}\\&\leqq\|\lambda^*\|+\omega_0 a_{18}(\mu^{(k)})^{\alpha}\quad\text{(from (5.35))}\\&\leqq\|\lambda^*\|+\omega_0 a_{18}\quad\text{(from (5.37))}\\&\leqq\nu(\mu^{(k+1)})^{-\gamma}\end{aligned} \tag{5.44}$$

for all $k > k_5$, the last inequality following from (5.41) and because $\mu^{(k+1)} \leqq \mu^{(k)}$. Hence, the multiplier update $\lambda^{(k+1)} = \hat{\lambda}^{(k+1)}$ in Algorithm 2 will always take place when $k \geqq k_0$.

For iteration $k_0$, $\omega^{(k_0)} = \omega_0(\mu^{(k_0)})^{\alpha_\omega}$, and $\eta^{(k_0)} = \eta_0(\mu^{(k_0)})^{\alpha_\eta}$. Then (5.5) gives

$$\begin{aligned}
\|c(x^{(k_0)})\| &\leqq s_1((\mu^{(k_0)} + a_6(\mu^{(k_0)})^2)\|\lambda^{(k_0)} - \lambda^*\| + a_5\omega^{(k_0)}\mu^{(k_0)}) \\
&\leqq s_1(2\mu^{(k_0)}\|\lambda^{(k_0)} - \lambda^*\| + a_5\omega^{(k_0)}\mu^{(k_0)}) \quad \text{(from (5.42))} \\
&\leqq s_1(2\omega_0\mu^{(k_0)} + a_5\omega_0(\mu^{(k_0)})^{1+\alpha_\omega}) \quad \text{(from (5.43))} \\
&\leqq \omega_0 s_1(a_5+2)\mu^{(k_0)} \quad \text{(from (5.37))} \\
&\leqq \eta_0(\mu^{(k_0)})^{\alpha_\eta} = \eta^{(k_0)} \quad \text{(from (5.39))}.
\end{aligned} \tag{5.45}$$

Thus, from (5.45), Step 2 of Algorithm 1 or the same step of Algorithm 2 will be executed with $\lambda^{(k_0+1)} = \bar{\lambda}(x^{(k_0)}, \lambda^{(k_0)}, S^{(k_0)}, \mu^{(k_0)})$ or $\lambda^{(k_0+1)} = \hat{\lambda}^{(k_0+1)}$ respectively. Inequality (5.4)/(5.34) in conjunction with (5.35), (5.38), and (5.43) guarantee that

$$\|\lambda^{(k_0+1)} - \lambda^*\| \leqq a_5\omega^{(k_0)} + a_6\mu^{(k_0)}\|\lambda^{(k_0)} - \lambda^*\| \leqq \omega_0 a_{18}(\mu^{(k_0)})^{\alpha}. \tag{5.46}$$

We shall now suppose that step 2 is executed for iterations $k_0 + i$, $(0 \leqq i \leqq j)$, and that

$$\|\lambda^{(k_0+i+1)} - \lambda^*\| \leqq \omega_0 a_{18}(\mu^{(k_0)})^{\alpha+\beta_\eta i}. \tag{5.47}$$

Inequalities (5.45) and (5.46) show that this is true for $j = 0$. We aim to show that the same is true for $i = j+1$. Under our supposition, we have, for iteration $k_0+j+1$, that $\mu^{(k_0+j+1)} = \mu^{(k_0)}$, $\omega^{(k_0+j+1)} = \omega_0(\mu^{(k_0)})^{\beta_\omega(j+1)+\alpha_\omega}$, and $\eta^{(k_0+j+1)} = \eta_0(\mu^{(k_0)})^{\beta_\eta(j+1)+\alpha_\eta}$. Then (5.5) gives

$$\begin{aligned}
\|c(x^{(k_0+j+1)})\| &\leqq s_1((\mu^{(k_0+j+1)} + a_6(\mu^{(k_0+j+1)})^2)\|\lambda^{(k_0+j+1)} - \lambda^*\| + a_5\omega^{(k_0+j+1)}\mu^{(k_0+j+1)}) \\
&\leqq s_1(2\mu^{(k_0+j+1)}\|\lambda^{(k_0+j+1)} - \lambda^*\| + a_5\omega^{(k_0+j+1)}\mu^{(k_0+j+1)}) \quad \text{(from (5.42))} \\
&\leqq s_1(2\omega_0 a_{18}\mu^{(k_0)}(\mu^{(k_0)})^{\alpha+\beta_\eta j} + a_5\omega_0(\mu^{(k_0)})^{\alpha_\omega+\beta_\omega(j+1)+1}) \quad \text{(from 5.47))} \\
&\leqq s_1(2\omega_0 a_{18}\mu^{(k_0)}(\mu^{(k_0)})^{\alpha_\eta+\beta_\eta j} + a_5\omega_0(\mu^{(k_0)})^{\alpha_\eta+\beta_\eta(j+1)+1}) \\
&\qquad \text{(from (5.35)-(5.37))} \\
&\leqq \omega_0 s_1(a_5+2a_{18})(\mu^{(k_0)})^{1-\beta_\eta}(\mu^{(k_0)})^{\beta_\eta(j+1)+\alpha_\eta} \quad \text{(from (5.37))} \\
&\leqq \eta_0(\mu^{(k_0)})^{\beta_\eta(j+1)+\alpha_\eta} = \eta^{(k_0+j+1)} \quad \text{(from (5.40))}.
\end{aligned} \tag{5.48}$$

Thus, from (5.48), step 2 of Algorithm 1 or the same step of Algorithm 2 will be executed with $\lambda^{(k_0+j+2)} = \bar{\lambda}(x^{(k_0+j+1)}, \lambda^{(k_0+j+1)}, S^{(k_0+j+1)}, \mu^{(k_0+j+1)})$ or $\lambda^{(k_0+j+2)} = \hat{\lambda}^{(k_0+j+2)}$ respectively. Inequality (5.4)/(5.34) then guarantees that

$$\begin{aligned}
\|\lambda^{(k_0+j+2)} - \lambda^*\| &\leqq a_5\omega^{(k_0+j+1)} + a_6\mu^{(k_0+j+1)}\|\lambda^{(k_0+j+1)} - \lambda^*\| \\
&\leqq \omega_0 a_5(\mu^{(k_0)})^{\alpha_\omega+\beta_\omega(j+1)} + \omega_0 a_6 a_{18}\mu^{(k_0)}(\mu^{(k_0)})^{\alpha+\beta_\eta j} \quad \text{(from (5.47))} \\
&\leqq \omega_0 a_5(\mu^{(k_0)})^{\alpha+\beta_\eta(j+1)} + \omega_0 a_6 a_{18}\mu^{(k_0)}(\mu^{(k_0)})^{\alpha+\beta_\eta j} \quad \text{(from (5.35)-(5.37))} \\
&= \omega_0(a_5 + a_6 a_{18}(\mu^{(k_0)})^{1-\beta_\eta})(\mu^{(k_0)})^{\alpha+\beta_\eta(j+1)} \\
&\leqq \omega_0(a_5 + a_6)(\mu^{(k_0)})^{\alpha+\beta_\eta(j+1)} \quad \text{(from (5.40))} \\
&= \omega_0 a_{18}(\mu^{(k_0)})^{\alpha+\beta_\eta(j+1)},
\end{aligned} \tag{5.49}$$

which establishes (5.47) for $i = j+1$. Hence, step 2 of the appropriate algorithm is executed for all iterations $k \geqq k_0$. But this implies that $\Gamma$ is finite, which contradicts the assumption that step 3 is executed infinitely often. Hence the theorem is proved. □

Note, in particular, that if Algorithm 2 is used with $\hat{\lambda}^{(k+1)}$ chosen as either the first-order or least-squares multiplier estimates, the penalty parameter $\mu^{(k)}$ will stay bounded away from zero. This follows directly from Theorem 5.3 because of the inequalities (4.18) and (4.20).

Our definition of floating variables has a further desirable consequence if we make the following additional assumption.

(AS6) (Strict complementary slackness condition) If the iterates $x^{(k)}$, $k \in K$, converge to the limit point $x^*$ with corresponding Lagrange multipliers $\lambda^*$, we assume that the set

$$J_2 = \{i \mid (g_L(x^*, \lambda^*))_i = 0 \quad \text{and} \quad x_i^* = 0\} \tag{5.50}$$

is empty.

Note that if inequality constraints $c_i(x) \geqq 0$ have been converted to equations by the subtraction of slack variables (i.e., rewritten as $c_i(x) - x_{n+i} = 0$, $x_{n+i} \geqq 0$), this statement of strict complementary slackness is equivalent to the more usual one which says that no inequality constraint shall be both active (the constraint function vanishing) and have a corresponding zero Lagrange parameter (see, e.g., Fletcher (1981, p. 51)). For it is easy to show that the Lagrange parameter for such a constraint is precisely the corresponding component of the gradient of the Lagrangian function. A constraint being active and having a corresponding zero Lagrange parameter is thus the same as the slack variable having the value zero, and its corresponding element in the gradient of the Lagrangian function vanishing so the latter is excluded under (AS6).

THEOREM 5.4. *Suppose that the iterates $x^{(k)}$, $k \in K$, converge to the limit point $x^*$ with corresponding Lagrange multipliers $\lambda^*$, and that* (AS1)-(AS3) *and* (AS6) *hold. Then for k sufficiently large, the set of floating variables are precisely those which lie away from their bounds at $x^*$.*

*Proof.* From Theorem 4.4, $\nabla_x \Phi^{(k)}$ converges to $g_L(x^*, \lambda^*)$ and from Lemma 2.1, the variables in the set $I_5$ then converge to zero and the corresponding components of $g_L(x^*, \lambda^*)$ are zero. Hence, under (AS6), $I_5$ is null. Therefore, each variable ultimately remains tied to one of the sets $I_1$ or $I_2$ for all $k$ sufficiently large; a variable in $I_1$ is, by definition, floating and converges to a value away from its bound. Conversely, a variable in $I_2$ is dominated and converges to its bound. □

As a consequence of Theorem 5.4, the least-squares multiplier estimates (2.14) are implementable. By this we mean that if $\bar{A}^{(k)}$ and $\bar{g}^{(k)}$ are the columns of $A(x^{(k)})$ and components of $g(x^{(k)})$ corresponding to the floating variables at $x^{(k)}$, respectively, the estimates

$$\hat{\lambda}^{(k)} = -(\bar{A}^{(k)+})^T \bar{g}^{(k)} \tag{5.51}$$

are identical to those given by (2.14) for all $k$ sufficiently large. The estimates (5.51), unlike (2.14), are well defined when $x^*$ is unknown.

We conclude the section by giving a rate-of-convergence result for our algorithms. For a comprehensive discussion of convergence, the reader is referred to Ortega and Rheinboldt (1970).

THEOREM 5.5. *Under the assumptions of Theorem* 5.3, *the iterates $x^{(k)}$, the Lagrange multiplier estimates $\bar{\lambda}^{(k)}$ of Algorithm* 1 *and any $\hat{\lambda}^{(k)}$ satisfying* (5.33) *for Algorithm* 2 *are at least R-linearly convergent with R-factor at most $\hat{\mu}^{\min(\beta_\omega, \beta_\eta)}$, where $\hat{\mu} = \min[\gamma_1, \mu]$ and where $\mu$ is the smallest value of the penalty parameter generated by the algorithm in question.*

*Proof.* The proof parallels that of Lemma 5.1. First, for $k$ sufficiently large, Theorem 5.3 shows that the penalty parameter $\mu^{(k)}$ remains fixed at some value $\mu$, say, and, for all subsequent iterations, inequalities (3.2)/(3.6) and

$$\omega^{(k+1)} = \hat{\mu})^{\beta_\omega}\omega^{(k)} \quad \text{and} \quad \eta^{(k+1)} = (\hat{\mu})^{\beta_\eta}\eta^{(k)} \tag{5.52}$$

hold. Then, from (3.2)/(3.6), (5.20), and (5.21), the bound on the right-hand side of (5.25) may be replaced by $a_{14}\omega^{(k)} + \eta^{(k)}$, and consequently,

$$\Delta x^{(k)} \leqq M(a_{14}\omega^{(k)} + \eta^{(k)} + a_{11}(\Delta x^{(k)})^2 + a_{12}\Delta x^{(k)}\omega^{(k)} + a_{13}(\omega^{(k)})^2). \tag{5.53}$$

Hence, if $k$ is sufficiently large that

$$\omega^{(k)} \leqq \min(1, 1/(2Ma_{12})) \tag{5.54}$$

and

$$\Delta x^{(k)} \leqq 1/(4Ma_{11}), \tag{5.55}$$

inequalities (5.53)-(5.55) can be rearranged to give

$$\Delta x^{(k)} \leqq 4M(a_{19}\omega^{(k)} + \eta^{(k)}) \tag{5.56}$$

where $a_{19} = a_{13} + a_{14}$. But then (5.22) and (5.56) give

$$\|x^{(k)} - x^*\| \leqq a_{20}\omega^{(k)} + a_{21}\eta^{(k)}, \tag{5.57}$$

where $a_{20} = 1 + 4Ma_{19}$ and $a_{21} = 4M$. As, by assumption, $\beta_\eta < 1$, (5.52) and (5.57) show that $x^{(k)}$ converges at least $R$-linearly, with $R$-factor $\hat{\mu}^{\min(\beta_\omega, \beta_\eta)}$, to $x^*$. That the same is true for $\bar{\lambda}^{(k)}$ and $\hat{\lambda}^{(k)}$ follows directly from (5.7)/(5.33) and (5.57). □

**6. An example.** In Theorem 5.3, we showed that, if there is a unique limit point for the iterates generated by the algorithms, the penalty parameter $\mu^{(k)}$ is necessarily bounded away from zero. We now show that, if there is more than a single limit point, but all the other assumptions of Theorem 5.3 are satisfied, it is indeed possible for the penalty parameter to become arbitrarily close to zero.

We consider the problem

$$\underset{x}{\text{minimize}}\ \sigma x \tag{6.1}$$

subject to the single constraint

$$x^2 - 1 = 0, \tag{6.2}$$

for some $\sigma > 0$. This problem has two stationary points, namely,

$$(x_1^*, \lambda_1^*) = \left(-1, \frac{\sigma}{2}\right) \quad \text{and} \quad (x_2^*, \lambda_2^*) = \left(1, -\frac{\sigma}{2}\right). \tag{6.3}$$

No bounds appear in the problem, and hence $P(x^{(k)}, \nabla_x\Phi^{(k)}) = \nabla_x\Phi^{(k)}$ for all $k$. (Of course, strictly we have not yet defined our algorithms for such a case—this case is covered in § 8; however, we might think of (6.1)-(6.2) as resulting from a transformation of variables where the nonnegativity constraint has been shifted so as to play no role here.) For simplicity, we choose $S^{(k)} = I$ for all $k$, and it can be verified that

$$\nabla_x\Phi(x, \lambda, I, \mu) = \frac{2}{\mu}x(x^2 - 1) + 2x\lambda + \sigma. \tag{6.4}$$

We wish to show that Algorithm 1 can generate a sequence of points that oscillate between neighbourhoods of $x_1^*$ and $x_2^*$, and such that the penalty parameter $\mu^{(k)}$ tends to zero. The idea is to consider an infinite sequence of iteration cycles, each of length $j+1$, where $j$ is the smallest integer such that

$$\eta_0(\min[\mu_0, \gamma_1])^{\alpha_\eta + j\beta_\eta} < \frac{\sigma}{2} \min[\mu_0, \gamma_1]. \tag{6.5}$$

For the first $j$ iterations of each cycle, $x^{(k)}$ lies in a neighbourhood of $x_2^*$ and step 2 is executed; for the iteration that remains, $x^{(k)}$ has a value less than $x_1^*$ and the penalty parameter is reduced as step 3 is executed. The process is started with $\lambda^{(0)} = \lambda_2^*$.

It remains to show that such a sequence can be constructed. For simplicity, we shall consider a single cycle. We will denote the sequence of generated iterates and corresponding Lagrange multiplier estimates by $\{x^{(k)}\}$ and $\{\lambda^{(k)}\}$, $1 \leqq k \leqq j+1$, respectively, and will let $\mu = \mu^{(k)}$ denote the constant penalty parameter throughout the cycle.

Now define $\beta = \max(1, \alpha_\omega, \beta_\omega)$. Note that, under conditions (5.35) and (5.36), $\alpha_\eta$ and $\beta_\eta$ are smaller than $\beta$. Therefore, because our cycle consists of $j$ iterations in which step 2 is executed followed by a single iteration with step 3, the convergence tolerances satisfy

$$\omega^{(k)} = \omega_0 \alpha^{\alpha_\omega + (k-1)\beta_\omega} \geqq \omega_0 \alpha^{k\beta} \tag{6.6}$$

and

$$\eta^{(k)} = \eta_0 \alpha^{\alpha_\eta + (k-1)\beta_\eta} \geqq \eta_0 \alpha^{k\beta}, \tag{6.7}$$

where $\alpha = \min(\mu, \gamma_1) < 1$. Furthermore, pick

$$\sigma \leqq 4/\mu_0 \tag{6.8}$$

and define

$$\xi = \min\left(\frac{\min(1, \eta_0)}{\mu_0}, \sigma, \eta_0, \frac{\omega_0}{6}\right). \tag{6.9}$$

The cycle involves two types of iterate:

(i) For the first $j$ iterations,

$$\lambda^{(k)} = \begin{cases} -\dfrac{\sigma}{2} & \text{for } k = 1, \\ -\dfrac{\sigma}{2} + \xi\alpha^{(k+1)\beta} & \text{for } 1 < k \leqq j, \end{cases} \tag{6.10}$$

$x^{(k)}$ lies in a neighbourhood of $x_2^*$ and step 2 is executed. (Strictly, for this demonstration, the power of $\alpha$ in (6.10) need only be $k\beta$; however, the extra power of $\beta$ will be important when we discuss Algorithm 2.)

(ii) For the last iteration, $\lambda^{(j+1)} = -\sigma/2$, $x^{(j+1)} < x_1^*$ and step 3 is executed.

Turning to details, consider case (i). We first show that equation (6.10) determines $x^{(k)}$. For (6.10) gives that

$$\lambda^{(k+1)} = \lambda^{(k)} + \begin{cases} \xi\alpha^{3\beta} & \text{for } k = 1, \\ -\xi(1-\alpha^\beta)\alpha^{(k+1)\beta} & \text{for } 1 < k < j, \\ -\xi\alpha^{(j+1)\beta} & \text{for } k = j. \end{cases} \tag{6.11}$$

But then equations (2.5), (3.3), and (6.11) imply that

$$c(x^{(k)}) = \begin{cases} \xi\mu\alpha^{3\beta} & \text{for } k = 1, \\ -\xi(1-\alpha^\beta)\mu\alpha^{(k+1)\beta} & \text{for } 1 < k < j, \\ -\xi\mu\alpha^{(j+1)\beta} & \text{for } k = j, \end{cases} \tag{6.12}$$

and therefore

$$x^{(k)}=\begin{cases}\sqrt{1+\xi\mu\alpha^{3\beta}} & \text{for } k=1,\\ \sqrt{1-\xi(1-a^{\beta})\mu\alpha^{(k+1)\beta}} & \text{for } 1<k<j,\\ \sqrt{1-\xi\mu\alpha^{(j+1)\beta}} & \text{for } k=j\end{cases}\tag{6.13}$$

as $c(x)=x^2-1$. We now show that such values of $x^{(k)}$ and $\lambda^{(k)}$ pass the acceptance tests (3.1) and (3.2). The constraint test (3.2) is satisfied for all $1\leqq k\leqq j$ because of (6.7), (6.9), (6.12) and because $\mu\leqq\mu_0$ and $\alpha<1$. Furthermore, it follows from (6.4) and (6.10) that the gradient of the augmented Lagrangian function at $(x^{(k)}, \lambda^{(k)})$ is

$$\nabla_x\Phi(x^{(k)}, \lambda^{(k)}, I, \mu)=\begin{cases}\sigma(1-x^{(k)})+2\xi x^{(k)}\alpha^{(k+2)\beta} & \text{for } 1\leqq k<j,\\ \sigma(1-x^{(k)}) & \text{for } k=j.\end{cases}\tag{6.14}$$

It remains to show that this gradient is acceptably small. First,

$$1\leqq x^{(1)}\leqq 1+\tfrac{1}{2}\xi\mu\alpha^{3\beta}\tag{6.15}$$

by dint of (6.13). Thus, from (6.8), (6.14), and (6.15),

$$0\leqq\xi(2-\tfrac{1}{2}\sigma\mu)\alpha^{3\beta}\leqq\nabla_x\Phi(x^{(1)}, \lambda^{(1)}, I, \mu)\leqq\xi(2+\xi\mu\alpha^{3\beta})\alpha^{3\beta}.\tag{6.16}$$

But as $\mu\leqq\mu_0$ and $\alpha<1$, definition (6.9) gives

$$\xi(2+\xi\mu\alpha^{3\beta})\alpha^{2\beta}\leqq\xi(2+\xi\mu_0)\leqq 3\xi\leqq 6\xi\leqq\omega_0.\tag{6.17}$$

Thus (6.6), (6.16), and (6.17) imply (3.1). Next, consider any $k$ for which $1<k<j$. Then

$$0\leqq 1-\xi(1-\alpha^{\beta})\mu\alpha^{(k+1)\beta}\leqq x^{(k)}\leqq 1\tag{6.18}$$

because of (6.9) and (6.13). Hence, from (6.14) and (6.18)

$$0\leqq\nabla_x\Phi(x^{(k)}, \lambda^{(k)}, I, \mu)\leqq\xi(2\alpha^{\beta}+\sigma(1-\alpha^{\beta})\mu)\alpha^{(k+1)\beta}.\tag{6.19}$$

Once again, as $\mu\leqq\mu_0$ and $\alpha<1$, (6.8) and (6.9) give

$$\xi(2\alpha^{\beta}+\sigma(1-\alpha^{\beta})\mu)\alpha^{\beta}\leqq\xi(2+\sigma\mu_0)\leqq 6\xi\leqq\omega_0.\tag{6.20}$$

Equations (6.6), (6.19), and (6.20) then imply (3.1). Finally, (6.13) gives

$$0\leqq 1-\xi\mu\alpha^{(j+1)\beta}\leqq x^{(j)}\leqq 1.\tag{6.21}$$

Hence, from (6.14) and (6.21)

$$0\leqq\nabla_x\Phi(x^{(j)}, \lambda^{(j)}, I, \mu)\leqq\xi\sigma\mu\alpha^{(j+1)\beta}.\tag{6.22}$$

Then (3.1) follows from (6.6), (6.8), (6.9), and (6.22). Moreover, it follows from (6.15), (6.18), and (6.21) that $x_2^*$ is the only possible limit point of the first $j$ iterates of the cycle. Thus we have shown that the first $j$ iterates in our cycle have the required properties.

We now consider case (ii). We have to show that it is possible to have $x^{(j+1)}<x_1^*$ with $\|c(x^{(j+1)})\|>\eta^{(j+1)}$. Equivalently, we show that inequality (3.1) (but not (3.2)) of step 1 is satisfied for some $x^{(j+1)}$ of the form

$$x^{(j+1)}<-1.\tag{6.23}$$

We note that (6.5) and (6.7) imply that

$$\frac{\eta^{(j+1)}}{\mu}<\frac{\sigma}{2}.\tag{6.24}$$

Recalling that $\lambda^{(j+1)} = \lambda_2^* = -\frac{1}{2}\sigma$, we thus require that the inequalities (6.23),

$$|\psi(x)| \equiv \left| \frac{2}{\mu} x(x^2-1) - \sigma x + \sigma \right| \leqq \omega^{(j+1)}, \tag{6.25}$$

and

$$|x^2-1| > \eta^{(j+1)} \tag{6.26}$$

are satisfied at $x = x^{(j+1)}$. Now observe that any $x \in (-\sqrt{1+\sigma\mu}, -\sqrt{1+\eta^{(j+1)}})$ satisfies (6.26). At the endpoints of the interval, we have that

$$\psi(-\sqrt{1+\eta^{(j+1)}}) = \sqrt{1+\eta^{(j+1)}} \left[ \frac{-2\eta^{(j+1)}}{\mu} + \sigma + \frac{\sigma}{\sqrt{1+\eta^{(j+1)}}} \right] > 0 \tag{6.27}$$

and

$$\psi(-\sqrt{1+\sigma\mu}) = \sigma\sqrt{1+\sigma\mu} \left[ -1 + \frac{1}{\sqrt{1+\sigma\mu}} \right] < 0. \tag{6.28}$$

The continuity of the function $\psi$ along with (6.27) and (6.28) implies the existence of a root inside the interval. Any $x$ sufficiently close to this root will therefore satisfy the required inequalities (6.23), (6.25), and (6.26) and we select such a point to define $x^{(j+1)}$. Because of (6.26), step 3 is executed and $\lambda^{(j+2)}$ remains equal to $\lambda_2^*$.

Furthermore, since the interval $(-\sqrt{1+\sigma\mu}, -\sqrt{1+\eta^{(j+1)}})$ of case (ii) shrinks to the single point $x_1^* = -1$ as $\mu$ tends to zero, this point is the only possible limit point of the sequence of iterates besides $x_2^*$.

This completes our example for Algorithm 1.

We now show that a slightly modified form of this example applies to Algorithm 2. Given $\mu_0$, pick $\sigma$ sufficiently small such that

$$\sigma < \min\left( \frac{2\nu}{\mu_0^\gamma}, \frac{1}{2\mu_0} \right). \tag{6.29}$$

Note that such a $\sigma$ satisfies (6.8). We construct an infinite sequence of iteration cycles, each of length $j+2$ with $j$ defined as before to be the smallest integer such that inequality (6.5) is satisfied. The first $j$ iterations are identical to those already described in case (i) above. Iteration $j+1$ is identical to case (ii) above, except that the Lagrange multiplier estimate is set to $\lambda_1^* = \sigma/2$ in step 3. For the remaining iteration, a point $x^{(j+2)}$ close to $x_2^*$ is selected and the Lagrange multiplier estimate is reset to $\lambda_2^* = -\sigma/2$. Note that each Lagrange multiplier estimate, (6.10) or $\pm\sigma/2$, satisfies

$$|\hat{\lambda}^{(k)}| \leqq \nu\mu^{-\gamma} \tag{6.30}$$

because $\alpha < 1$, $\xi \leqq \sigma$, and by choice of $\sigma$ in (6.29). Moreover, (5.33) is satisfied because the errors in the multiplier estimates (6.10) are bounded by $\omega^{(k)}$ from (6.7) and (6.9).

It remains to show that we can construct a suitable iterate $x^{(j+2)}$ at the last iteration of each cycle. We thus require that

$$\|\nabla_x \Phi(x^{(j+2)}, \lambda_1^*, I, \mu)\| = \left| \frac{2}{\mu} x^{(j+2)}(x^{(j+2)2} - 1) + \sigma x^{(j+2)} + \sigma \right| \leqq \omega^{(j+2)}. \tag{6.31}$$

This may be achieved by choosing $x^{(j+2)}$ as the zero

$$\frac{1+\sqrt{1-2\sigma\mu}}{2} \tag{6.32}$$

of that function. The Lagrange multiplier estimate is then reset to $\lambda_2^*$, which is allowed because of (6.30) and the fact that the root (6.32) converges to $x_2^* = 1$ as $\mu$ tends to zero. Whether or not (3.6) holds is mainly irrelevant, since its failure only causes a further reduction of the penalty parameter, which suits our purpose. However, if (3.6) is violated, we note that (6.6) and (6.7) must be replaced on the next cycle by

$$\omega^{(k)} = \omega_0 \alpha^{\alpha_\omega + k\beta_\omega} \geqq \omega_0 \alpha^{(k+1)\beta}, \tag{6.33}$$

and

$$\eta^{(k)} = \eta_0 \alpha^{\alpha_\eta + k\beta_\eta} \geqq \eta_0 \alpha^{(k+1)\beta}, \tag{6.34}$$

where $\alpha = \min(\mu, \gamma_1) < 1$. These slightly more stringent tolerances are still acceptable in the tests (3.5) and (3.6) on the first $j$ iterations of the next cycle because of the presence of the extra $\alpha^\beta$ term in equations (6.12) and (6.14).

**7. Second-order conditions.** It is useful to know how our algorithms behave if we impose further conditions on the iterates generated by the inner iteration. In particular, suppose that $x^{(k)}$ satisfies the following second-order sufficiency condition:

(AS7) Suppose that $x^{(k)}$ satisfies (3.1)/(3.5), converges to $x^*$ for $k \in K$, and that $J_1$ and $J_2$ are as defined by (5.1). Then we assume that $\nabla_{xx}\Phi^{(k)}_{[J_1 \cup J_2, J_1 \cup J_2]}$ is uniformly positive definite (that is, its smallest eigenvalue is uniformly bounded away from zero) for all $k \in K$ sufficiently large.

With such a condition we have the following result.

THEOREM 7.1. *Under* (AS1)–(AS3) *and* (AS7), *the iterates* $x^{(k)}$, $k \in K$, *generated by either Algorithm* 1 *or* 2 *converge to an isolated local solution of* (1.5)–(1.7).

*Proof.* By definition of $\Phi$,

$$\nabla_{xx}\Phi^{(k)} = H_L(x^{(k)}, \bar{\lambda}^{(k)}) + A^{(k)T} S^{(k)} A^{(k)} / \mu^{(k)}. \tag{7.1}$$

Let $s_{[J]}$ be any nonzero vector satisfying

$$A^{(k)}_{[J]} s_{[J]} = 0, \tag{7.2}$$

where $J$ is any set made up from the union of $J_1$ and any subset of $J_2$. Then for any such vector,

$$s^T_{[J]} \nabla_{xx} \Phi^{(k)}_{[J,J]} s_{[J]} \geqq \varepsilon, \tag{7.3}$$

for some $\varepsilon > 0$, under (AS7) as $J$ is a subset of $J_1 \cup J_2$. It follows from (7.1)–(7.3) that

$$s^T_{[J]} H_L(x^{(k)}, \bar{\lambda}^{(k)})_{[J,J]} s_{[J]} \geqq \varepsilon. \tag{7.4}$$

By continuity of $H_L$ as $x^{(k)}$ and $\bar{\lambda}^{(k)}$ approach their limits, this gives that

$$s^T_{[J]} H_L(x^*, \lambda^*)_{[J,J]} s_{[J]} \geqq \varepsilon \tag{7.5}$$

for all nonzero $s_{[J]}$ satisfying (7.2), which implies that $x^*$ is an isolated local solution to (1.5)–(1.7) (see, for example, Avriel, (1976, Thm. 3.11)). □

The importance of (AS7) is that the inner iteration termination test (step 1 of either algorithm) might be tightened so that $\nabla_{xx}\Phi^{(k)}_{[J^{(k)}, J^{(k)}]}$ is required to be uniformly positive definite, for all floating variables $J^{(k)}$ and all $k$ sufficiently large, in addition to (3.1)/(3.5). If the strict complementary slackness condition (AS6) holds at $x^*$, Theorem 5.4 ensures that the set $J_2$ is empty and $J_1$ identical to the set of floating variables after a finite number of iterations and thus, under this tighter termination test, (AS7) and Theorem 7.1 hold.

There is a weaker version of this result, proved in the same way, that if the assumption of uniform positive definiteness in (AS7) is replaced by an assumption of positive semidefiniteness, the limit point then satisfies second-order necessary conditions (Avriel (1976, Thm. 3.10)) for a minimizer. This weaker version of (AS7) is easier to ensure in practice as certain methods for solving the inner iteration subproblem, for instance, that of Conn, Gould, and Toint (1988a), guarantee that the second derivative matrix at the limit point of a sequence of generated inner iterates will be positive semidefinite.

**8. Further comments.** We now briefly turn to the more general problem (1.1)-(1.3). As we indicated in our Introduction, the presence of the more general constraints (1.3) does not significantly alter the conclusions that we have drawn so far. If we define the appropriate generalization of the projection (2.1) by

$$(P[x])_i = \begin{cases} l_i & \text{if } x_i \leqq l_i, \\ u_i & \text{if } x_i \geqq u_i, \\ x_i & \text{otherwise,} \end{cases} \tag{8.1}$$

and let $B = \{x \mid l \leqq x \leqq u\}$, we may then use the algorithms of § 3 without further significant modification. Our concept of floating and dominated variables stays essentially the same; for any iterate $x^{(k)}$ in $B$ we have three mutually exclusive possibilities for each component $x_i^{(k)}$, namely,

$$\begin{aligned} &\text{(i)} && 0 \leqq x_i^{(k)} - l_i \leqq (\nabla_x \Phi^{(k)})_i, \\ &\text{(ii)} && (\nabla_x \Phi^{(k)})_i \leqq x_i^{(k)} - u_i \leqq 0, \\ &\text{(iii)} && x_i^{(k)} - u_i < (\nabla_x \Phi^{(k)})_i < x_i^{(k)} - l_i. \end{aligned} \tag{8.2}$$

In case (i) we then have

$$(P(x^{(k)}, \nabla_x \Phi^{(k)}))_i = x_i^{(k)} - l_i, \tag{8.3}$$

whereas in case (ii) we have

$$(P(x^{(k)}, \nabla_x \Phi^{(k)}))_i = x_i^{(k)} - u_i, \tag{8.4}$$

and in case (iii)

$$(P(x^{(k)}, \nabla_x \Phi^{(k)}))_i = (\nabla_x \Phi^{(k)})_i. \tag{8.5}$$

The $x_i^{(k)}$ which satisfy (i) or (ii) are now the dominated variables (the ones satisfying (i) are said to be *dominated above* and those satisfying (ii) *dominated below*); those which satisfy (iii) are the floating variables. As a consequence, the sets corresponding to those given in (2.12) are straightforward to define. $I_1$ now contains variables that float for all $k \in K$ sufficiently large and converge to the interior of $B$. $I_2$ is now the union of the two sets $I_{2l}$, made up of variables that are dominated above for all $k \in K$ sufficiently large, and $I_{2u}$, made up of variables that are dominated below for all $k \in K$ sufficiently large. Likewise, $I_3$ is the union of the two sets $I_{3l}$, made up of variables that are floating for all sufficiently large $k \in K$ but converge to their lower bounds, and $I_{3u}$, made up of variables that are floating for all sufficiently large $k \in K$ but converge to their upper bounds. With such definitions, we may reprove all of the results of §§ 3-7, assumptions (AS5) and (AS6) being extended in the obvious way and Theorem 5.4 being strengthened to say that, for all $k \in K$ sufficiently large, $I_{2l}$ and $I_{2u}$ are precisely the variables that lie at their lower and upper bounds (respectively) at $x^*$.

We have not made any statement here about how the scaling matrices $S^{(k)}$ should be constructed, merely that they may be used. We consider that constraint scaling is essential for any realistic algorithm and believe that it is important that the scaling can be changed (albeit not too drastically) as the computation proceeds. We defer a discussion of the issues of how to choose such scalings until we have performed significant numerical testing of our algorithms. We also note that the results given here are unaltered if the convergence tolerance (3.1)/(3.5) is replaced by

$$\|D^{(k)}P(x^{(k)}, \nabla_x\Phi^{(k)})\| \leqq \omega^{(k)} \tag{8.6}$$

for any sequence of positive diagonal matrices $\{D^{(k)}\}$ with uniformly bounded condition number. This is important as the method of Conn, Gould, and Toint (1988a), which we would consider using to solve the inner iteration problem, allows for different scalings for the components of the gradients to cope with variables of differing magnitudes.

Finally, although the rules for how the convergence tolerances $\eta^{(k)}$ and $\omega^{(k)}$ are updated have been made rather rigid in this paper and although the results contained here may be proved under more general updating rules, we have refrained from doing so here as the resulting conditions on the updates seemed rather complicated and are unlikely to provide more practical updates.

**Acknowledgments.** The authors thank Annick Sartenaer for her careful reading of this paper. We are also grateful to two anonymous referees whose comments have helped us improve the original draft.

## REFERENCES

M. AVRIEL, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

M. C. BARTHOLOMEW-BIGGS, *Recursive quadratic programming methods based on the augmented Lagrangian function*, Math. Programming Stud., 31 (1987), pp. 21-42.

D. P. BERTSEKAS, *Augmented Lagrangian and exact penalty methods*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, London, New York, 1982a, pp. 223-234.

———, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, London, New York, 1982b.

A. R. CONN, N. I. M. GOULD, M. LESCRENIER, AND PH. L. TOINT, *Performance of a multifrontal scheme for partially separable optimization*, Report CSS 218, AERE, Harewell Laboratory, Harwell, U.K., 1987.

A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 25 (1988a), pp. 433-460, see also SIAM J. Numer. Anal., 26 (1989), pp. 764-767.

——, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988b), 399-430.

R. S. DEMBO, S. C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400-408.

R. S. DEMBO AND U. TULOWITZKI, *Local convergence analysis for successive inexact quadratic programming methods*, School of Organization and Management Working paper series B no. 78, Yale University, New Haven, CT, 1984.

L. C. W. DIXON, P. DOLAN, AND R. PRICE, *Finite element optimization: the use of structured automatic differentiation*, in Stimulation and Optimization of Large Systems, A. Osiadacz, ed., Oxford University Press, Oxford, 1988, pp. 117-141.

A. DRUD, CONOPT: *a* GRG *code for large sparse dynamic nonlinear optimization problems*, Math. Programming, 31 (1985), pp. 153-191.

R. FLETCHER, *Practical Methods of Optimization, Vol.* 2, John Wiley, London, New York, 1981.

P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *Some theoretical properties of an augmented Lagrangian merit function*, Report SOL 86-6, Department of Operations Research, Stanford University, Stanford, CA, 1986.

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.

———, *On the convergence of a sequential penalty function method for constrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 107-128.

A. Griewank and Ph. L. Toint, *Partitioned variable metric updates for large structured optimization problems*, Numer. Math., 39 (1982), pp. 119-137.

W. A. Gruver and E. Sachs, *Algorithmic Methods in Optimal Control*, Research Notes in Math., 47, Pitman, Boston, 1980.

W. W. Hager, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., 55 (1987), pp. 37-71.

M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303-320.

L. S. Lasdon, *Reduced gradient methods*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, London, New York, 1982, pp. 235-242.

D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, London, 1973.

J. J. Moré, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming: The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 258-287.

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.

B. A. Murtagh and M. A. Saunders, MINOS/*Augmented user's manual*, Report SOL 80-14, Department of Operations Research, Stanford University, Stanford, CA, 1980.

J. M. Ortega and W. C. Rheinbolt, *Iterative solution of nonlinear equations in several variables*, Academic Press, London, New York, 1970.

E. Polak and A. L. Tits, *A globally convergent, implementable multiplier method with automatic penalty limitation*, Appl. Math. Optim., 6 (1980), pp. 335-360.

M. J. D. Powell, *A method for nonlinear constraints in minimization problems* in Optimization, R. Fletcher, ed., Academic Press, London, New York, 1969.

R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97-116.

K. Schittkowski, *The nonlinear programming method of Wilson, Han and Powell with an augmented Lagrangian type line search function*, Numer. Math., 38 (1981), pp. 83-114.

T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626-637.

R. A. Tapia, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135-194.

Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, New York, 1981.

———, *Nonlinear optimization in a large number of variables*, in Simulation and Optimization of Large Systems, A. Osiadacz, ed., Oxford University Press, Oxford, 1988.

H. Yamashita, *A globally convergent constrained quasi-Newton method with an augmented Lagrangian type penalty function*, Math. Programming, 23 (1982), pp. 75-86.