Mathematical Programming 49 (1991) 189-211
North-Holland

A trust region algorithm for equality constrained optimization

M.J.D. Powell and Y. Yuan

Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 9EW, England

Received 20 February 1986
Revised manuscript received 14 March 1989

A trust region algorithm for equality constrained optimization is proposed that employs a differentiable exact penalty function. Under certain conditions global convergence and local superlinear convergence results are proved.

Key words: Equality constrained optimization, exact penalty functions, nonlinear programming, superlinear convergence, trust regions.

1. Introduction

We consider the equality constrained problem

$$\text{minimize} \quad f(x), \quad x \in \mathbb{R}^n, \tag{1.1}$$

$$\text{subject to} \quad h_i(x) = 0, \quad i = 1, 2, \ldots, m, \tag{1.2}$$

We suppose that $f(x)$ and $h_i(x)$ $(i = 1, 2, \ldots, m)$ are continuously differentiable and that the constraints gradients are linearly independent (but continuous third derivatives are assumed in the theoretical analysis). We employ the following notation:

$$c(x) = (h_1(x), \ldots, h_m(x))^T, \tag{1.3a}$$

$$A(x) = \nabla(c(x))^T = (\nabla h_1(x), \ldots, \nabla h_m(x)), \tag{1.3b}$$

$$g(x) = \nabla f(x). \tag{1.3c}$$

We also use c_k for $c(x_k)$, A_k for $A(x_k)$, etc.

Given an estimate of the solution x_k, many sequential quadratic programming methods for solving (1.1)–(1.2) obtain a search direction d_k by solving the following subproblem:

$$\text{minimize} \quad g_k^T d + \tfrac{1}{2} d^T B_k d, \quad d \in \mathbb{R}^n, \tag{1.4}$$

$$\text{subject to} \quad c_k + A_k^T d = 0, \tag{1.5}$$

where B_k is an $n \times n$ symmetric matrix. The next iterate has the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.6}$$

where $\alpha_k > 0$ is a step length and α_k depends on a line search technique. For more details see e.g. Bertsekas (1982), Biggs (1978, 1983), Powell (1978, 1983), Powell and Yuan (1986a) and Schittkowski (1981, 1983).

In this paper, we consider a trust region algorithm for solving the constrained optimization problem (1.1)–(1.2). Trust region algorithms, like line search methods, are iterative. At every iteration, a trial step is calculated, and some kind of test is made to decide whether it should be accepted. Trust region algorithms for unconstrained optimization have been discussed by many authors, including Fletcher (1987), Gay (1981), Moré (1983), Powell (1975), Sorensen (1982) and Yuan (1985). They have also been applied recently to equality constrained optimization calculations by Byrd, Schnabel and Shultz (1985) and by Vardi (1985), using a nondifferentiable exact penalty function to force global convergence. However we will employ a differentiable penalty function.

Trust region algorithms require the length of each trial step to be bounded by a positive parameter that is chosen automatically. We let d_k satisfy the inequality

$$\|d\|_2 \leq \Delta_k, \tag{1.7}$$

where $\Delta_k > 0$ is the trust region bound at the kth iteration. One difficulty when using a trust region technique is that the trust region restriction (1.7) and the linearized constraint (1.5) may be inconsistent, that is the equation (1.5) may have no solution within the trust region. Our way of overcoming this difficulty is the one proposed by Celis Dennis and Tapia (1985), namely to replace equation (1.5) by the condition $\|c_k + A_k^T d\|_2 \leq \zeta_k$ where ζ_k is a number between the bounds

$$\min_{\|d\|_2 \leq \Delta_k} \|c_k + A_k^T d\|_2 \leq \zeta_k \leq \|c_k\|_2. \tag{1.8}$$

The lower bound ensures that the constraints on d are consistent, and the upper bound is satisfied as a strict inequality unless $\|c_k\|_2 = 0$ in order to help the correction of constraint violations. Celis et al. choose ζ_k by taking a Cauchy step from $d = 0$ for the function $\{\|c_k + A_k^T d\|_2^2 : d \in \mathbb{R}^n\}$, but we prefer ζ_k to be the least value of $\|c_k + A_k^T d\|_2$ subject to $\|d\|_2 \leq b\Delta_k$ for some $b \in [b_2, b_1]$, where b_1 and b_2 are preassigned constants that satisfy $0 < b_2 \leq b_1 < 1$. Our choice of ζ_k is zero on more iterations when ill-conditioning makes the Cauchy step short, and in both cases the restrictions on d allow some freedom which is used automatically to reduce the quadratic objective function (1.4).

The main aim of this paper is to extend the results in Powell and Yuan (1986a) from line searches to trust regions because there seem to be some advantages in algorithms that use exact differentiable penalty functions to force convergence from poor starting approximations. Specifically, one avoids some difficulties due to first derivative discontinuities, the best known one being the "Maratos effect" (see Powell (1983) for example). Line searches can also cause severe inefficiencies when the

current vector of variables is far from the required solution (Powell (1985)). Further, trust region techniques avoid the need for B_k to be positive definite on the null space of A_k^T which occurs in line search algorithms. There is some discussion of these questions in Section 5.

Our algorithm (except for the updating of B_k) is specified in the next section, some global convergence properties are proved in Section 3, and a local superlinear convergence result is established in Section 4. Some implementation questions are considered in Section 5, in particular the calculation on each iteration of the trial change to the current vector of variables. The merits of trust region methods and differentiable exact penalty functions are considered too. This discussion motivates the given convergence analysis, because it does seem to be worthwhile to try to combine the use of a differentiable exact penalty function with trust region techniques without the calculation of second derivatives.

2. The algorithm

As mentioned in the previous section, trust region algorithms are iterative. In order to begin our calculation, an initial point x_1, an $n \times n$ symmetric B_1 and a trust region bound Δ_1 are required.

At the kth iteration, if x_k does not satisfy the Kuhn–Tucker conditions, we calculate a trial step by solving the subproblem

$$\text{minimize} \quad g_k^T d + \tfrac{1}{2} d^T B_k d, \quad d \in \mathbb{R}^n, \tag{2.1}$$

$$\text{subject to} \quad \|c_k + A_k^T d\|_2 \leq \zeta_k \quad \text{and} \quad \|d\|_2 \leq \Delta_k, \tag{2.2}$$

where ζ_k is any number satisfying the inequalities

$$\min_{\|d\|_2 \leq b_1 \Delta_k} \|c_k + A_k^T d\|_2 \leq \zeta_k \leq \min_{\|d\|_2 \leq b_2 \Delta_k} \|c_k + A_k^T d\|_2, \tag{2.3}$$

and where b_1 and b_2 are two given constants that satisfy $0 < b_2 \leq b_1 < 1$. To test whether we should accept our trial step, d_k say, we use Fletcher's differentiable exact penalty function

$$\phi_k(x) = f(x) - \lambda(x)^T c(x) + \sigma_k \|c(x)\|_2^2, \tag{2.4}$$

where, for each $x \in \mathbb{R}^n$, $\lambda(x) \in \mathbb{R}^m$ minimizes the sum of squares of residuals of the Kuhn–Tucker conditions

$$\|g(x) - A(x)\lambda\|_2^2, \quad \lambda \in \mathbb{R}^m, \tag{2.5}$$

and where $\sigma_k > 0$ is a penalty parameter. We let D_k be the "predicted change"

$$D_k = (g_k - A_k \lambda_k)^T d_k + \tfrac{1}{2} d_k^T B_k d_k$$
$$- [\lambda(x_k + d_k) - \lambda_k]^T (c_k + \tfrac{1}{2} A_k^T d_k) + \sigma_k (\|c_k + A_k^T d_k\|_2^2 - \|c_k\|_2^2), \tag{2.6}$$

in $\phi_k(x)$ where σ_k is chosen so that $D_k < 0$ and where $\hat{d}_k$ is the orthogonal projection of d_k into the null space of A_k^T. Then we calculate the ratio

$$r_k = \frac{\phi_k(x_k + d_k) - \phi_k(x_k)}{D_k} \tag{2.7}$$

of the actual change to the predicted change in $\phi_k(x)$. Our method sets the next iterate x_{k+1} to $x_k + d_k$ if $r_k > 0$; otherwise $x_{k+1} = x_k$. The choice of the next trust region bound Δ_{k+1} depends on Δ_k, $\|d_k\|_2$ and r_k. An $n \times n$ symmetric matrix B_{k+1} is defined and this completes the kth iteration.

The presence of the terms $\hat{d}_k$ and $\frac{1}{2}A_k^T d_k$ in the definition (2.6) deserves some comment. The vector $\hat{d}_k$ occurs because our condition on the matrices $\{B_k: k = 1, 2, 3, \ldots\}$ for superlinear convergence is not that the ratio $\|(B_k - W^*)d_k\|_2 / \|d_k\|_2$ tends to zero as $k \to \infty$, where W^* is the final second derivative matrix of the Lagrangian function: it is the weaker condition that $|v_k^T(B_k - W^*)d_k| / \|d_k\|_2$ tends to zero, where v_k is any normalized vector such that $A_k^T v_k = 0$. Thus, the use of $\hat{d}_k$ can provide Q-superlinear convergence as shown in Section 4. The term $(c_k + \frac{1}{2}A_k^T d_k)$ is just an estimate of the constraint vector at the mid-point of the line segment from x_k to $x_k + d_k$, which is also important to the analysis of Section 4. Both d_k and D_k would be zero only if they were calculated at a Kuhn–Tucker point x_k, but in this case termination would occur first at the beginning of the iteration.

A formal description of our algorithm, including some more details that are important to convergence, is as follows:

Step 0. $x_1 \in \mathbb{R}^n$, $B_1 \in \mathbb{R}^{n \times n}$, $\Delta_1 > 0$, and $0 < b_2 \leq b_1 < 1$ are given. Choose $\sigma_1 > 0$ and small $\varepsilon > 0$. Set $k = 1$.

Step 1. If $\|c_k\|_2 + \|g_k - A_k\lambda_k\|_2 \leq \varepsilon$ then stop. Otherwise solve the problem (2.1)-(2.3) which gives d_k.

Step 2. Calculate D_k by formula (2.6). If the inequality

$$D_k \leq \tfrac{1}{2}\sigma_k(\|c_k + A_k^T d_k\|_2^2 - \|c_k\|_2^2) \tag{2.8}$$

fails then increase σ_k to the value

$$\sigma_k^{\text{new}} = 2\sigma_k^{\text{old}} + \max\left\{0, \frac{2D_k^{\text{old}}}{\|c_k\|_2^2 - \|c_k + A_k^T d_k\|_2^2}\right\} \tag{2.9}$$

which ensures that the new value of expression (2.6) satisfies condition (2.8).

Step 3. Calculate the ratio (2.7). Set the values

$$x_{k+1} = \begin{cases} x_k + d_k & \text{if } r_k > 0, \\ x_k & \text{otherwise,} \end{cases} \tag{2.10}$$

and

$$\Delta_{k+1} = \begin{cases} \max[\Delta_k, 4\|d_k\|_2], & r_k > 0.9, \\ \Delta_k, & 0.1 \leq r_k \leq 0.9, \\ \min[\Delta_k/4, \|d_k\|_2/2], & r_k < 0.1. \end{cases} \tag{2.11}$$

Generate B_{k+1}. Set $\sigma_{k+1} = \sigma_k$. Set $k = k+1$ and go to Step 1.

We note that D_k is always negative at the end of Step 2, due to inequality (2.8) when $c_k \neq 0$, and due to the equivalence of expressions (2.1) and (2.6) when $c_k = A_k^T d_k = 0$. Therefore condition (2.10) accepts a trial step if and only if it reduces the penalty function (2.4). We note also that the penalty parameter σ_k is adjusted automatically. It is quite suitable to choose σ_k to satisfy condition (2.8), because an increase in σ_k is needed only when the step from x_k to $x_k + d_k$ is predicted to increase the Lagrangian part $[f(x) - \lambda(x)^T c(x)]$ of the merit function (2.4). In this case a new value of σ_k makes the new predicted decrease in $\sigma_k \|c(x)\|_2^2$ of the magnitude that is necessary for D_k to be negative. Further, as in Powell and Yuan (1986a), the use of $\lambda(x_k + d_k)$ in expression (2.6) leads to a suitable adjustment of the penalty parameter without the calculation of any second derivatives.

Unfortunately this algorithm is mainly of theoretical interest until an efficient procedure is developed for the calculation of the search direction. This problem is addressed in Yuan (1988). There a Newton-type method for determining d_k in the case when B_k is positive definite is described and analysed, and some numerical results are presented.

3. Global convergence

We call x a stationary point of the problem (1.1)–(1.2) if it satisfies the Kuhn–Tucker conditions

$$\|c(x)\|_2 + \|P(x)g(x)\|_2 = 0, \tag{3.1}$$

where $P(x)$ is the least-squares projection operator from $\mathbb{R}^n$ to the null space of $A(x)^T$, which, in view of the second assumption below, is the matrix

$$P(x) = I - A(x)[A(x)^T A(x)]^{-1} A(x)^T = I - A(x)A(x)^+. \tag{3.2}$$

It follows from the definition of $\lambda(x)$ that $P(x)g(x) = g(x) - A(x)\lambda(x)$, so the convergence test in Step 1 is a test on the Kuhn–Tucker conditions (3.1).

It is proved in this section that the following assumptions imply termination of the algorithm, where ε in Step 1 is any prescribed positive constant. Thus, one can calculate a point that is arbitrarily close to a stationary point of the problem (1.1)–(1.2). Some condition such as Assumption 3.1(a) is inevitable to rule out calculations where the algorithm should generate a divergent sequence $\{x_k: k = 1, 2, 3, \ldots\}$ because the objective function is not bounded below on the feasible region.

Assumptions 3.1. (a) There exists a bounded convex closed set $\Omega \subset \mathbb{R}^n$ such that x_k and $x_k + d_k$ are in Ω for all k.

(b) $A(x)$ has full column rank for all $x \in \Omega$.

(c) The matrices $\{B_k: k = 1, 2, 3, \ldots\}$ are uniformly bounded.

The first three lemmas provide an upper bound on D_k, that is important to the test for increasing σ_k.

Lemma 3.2. *For any positive number Δ, any vector $g \in \mathbb{R}^n$ and any $n \times n$ symmetric matrix B, if $\bar{d}$ is a solution of*

$$\text{minimize} \quad g^T d + \tfrac{1}{2} d^T B d, \quad d \in \mathbb{R}^n, \tag{3.3}$$

$$\text{subject to} \quad \|d\|_2 \leq \Delta \tag{3.4}$$

then the scalar product $g^T \bar{d}$ satisfies the inequality

$$g^T \bar{d} \leq -\frac{\|g\|_2^2 \Delta}{2\|B\|_2 \Delta + \|g\|_2} \leq -\frac{\|g\|_2^2 \|\bar{d}\|_2}{2\|B\|_2 \|\bar{d}\|_2 + \|g\|_2}. \tag{3.5}$$

Proof. Because condition (3.5) is trivial when $g = 0$, we assume that $\|g\|_2 > 0$. If $\|\bar{d}\|_2 < \Delta$, then B is positive definite or positive semidefinite and the equation

$$B\bar{d} + g = 0, \tag{3.6}$$

holds. Hence we have the relation

$$\bar{d} = -B^+ g + \hat{d}, \tag{3.7}$$

where B^+ is the generalized inverse of B and where $\hat{d}$ is a vector in the null space of B. Since g must be in the range space of B, it follows that the inequality

$$g^T \bar{d} = -g^T B^+ g \leq -\frac{1}{\|B\|_2} \|g\|_2^2 \tag{3.8}$$

is satisfied, which implies the first part of condition (3.5), while the second part is always an immediate consequence of $\|\bar{d}\|_2 \leq \Delta$.

If $\|\bar{d}\|_2 = \Delta$, then, by Theorem 5.2.1 of Fletcher (1987), there exists a nonnegative number μ satisfying the equation

$$g + (B + \mu I)\bar{d} = 0, \tag{3.9}$$

where the matrix $B + \mu I$ is positive definite or positive semidefinite. Using the argument above, we have the bound

$$g^T \bar{d} \leq -\|g\|_2^2 / \|B + \mu I\|_2. \tag{3.10}$$

Now equation (3.9) gives the relation

$$\mu = \frac{1}{\|\bar{d}\|_2} (\|B\bar{d} + g\|_2) \leq \|B\|_2 + \|g\|_2 / \Delta, \tag{3.11}$$

so we have the condition

$$\|B + \mu I\|_2 \leq 2\|B\|_2 + \|g\|_2 / \Delta. \tag{3.12}$$

Therefore inequality (3.5) follows from expression (3.10). □

Lemma 3.3. *The inequality*

$$\|c_k\|_2 - \|c_k + A_k^T d_k\|_2 \geq \min\left\{\|c_k\|_2, \frac{b_2\Delta_k}{\|A_k^+\|_2}\right\} \tag{3.13}$$

holds for all k, where b_2 is introduced in (2.3).

Proof. If $b_2\Delta_k \geq \|A_k^+\|_2\|c_k\|_2$, we have that $\zeta_k = 0$, since $c_k + A_k^T[-(A_k^T)^+ c_k] = 0$ and $\|-(A_k^T)^+ c_k\|_2 \leq b_2\Delta_k$. Therefore the equation

$$\|c_k\|_2 - \|c_k + A_k^T d_k\|_2 = \|c_k\|_2, \tag{3.14}$$

is implied by the first of the constraints (2.2).

In the case when $b_2\Delta_k < \|A_k^+\|_2\|c_k\|_2$, the second part of condition (2.3) and the relation $\|c_k + A_k^T d_k\|_2 \leq \zeta_k$ give the bound

$$\begin{aligned}\|c_k\|_2 - \|c_k + A_k^T d_k\|_2 &\geq \|c_k\|_2 - \zeta_k \\ &\geq \|c_k\|_2 - \left\| c_k - A_k^T\left\{\frac{b_2\Delta_k}{\|(A_k^T)^+ c_k\|_2}\right\}(A_k^T)^+ c_k \right\|_2 \\ &= \|c_k\|_2 \frac{b_2\Delta_k}{\|(A_k^T)^+ c_k\|_2} \geq \frac{b_2\Delta_k}{\|A_k^+\|_2}.\end{aligned} \tag{3.15}$$

Inequality (3.13) follows from expressions (3.14) and (3.15). □

Lemma 3.4. *There exists a positive constant m_1 such that the inequality*

$$\begin{aligned}&D_k + \tfrac{1}{2}\sigma_k(\|c_k\|_2^2 - \|c_k + A_k^T d_k\|_2^2) \\ &\quad \leq -\tfrac{1}{4}\|P_k\bar{g}_k\|_2^2 \min\left\{\frac{1}{2\|B_k\|_2}, \frac{\bar{\Delta}_k}{\|P_k\bar{g}_k\|_2}\right\} \\ &\qquad + m_1\|d_k\|_2\|c_k\|_2 - \tfrac{1}{2}\sigma_k\|c_k\|_2 \min\left\{\|c_k\|_2, \frac{b_2\Delta_k}{\|A_k^+\|_2}\right\}\end{aligned} \tag{3.16}$$

holds for all k, where we use the notation

$$\bar{g}_k = g_k + B_k\bar{d}_k, \tag{3.17a}$$

$$\bar{d}_k = (I - P_k)d_k, \tag{3.17b}$$

$$\bar{\Delta}_k = (\Delta_k^2 - \|\bar{d}_k\|_2^2)^{1/2}, \tag{3.17c}$$

and $P_k = P(x_k)$.

Proof. The definition of $\bar{d}_k$, equation (3.2) and $\|c_k + A_k^T d_k\|_2 \leq \|c_k\|_2$ imply the bound

$$\|\bar{d}_k\|_2 = \|A_k A_k^+ d_k\|_2 = \|(A_k^+)^T[(c_k + A_k^T d_k) - c_k]\|_2 \leq 2\|A_k^+\|_2\|c_k\|_2. \tag{3.18}$$

We recall from equation (2.6) the notation

$$\hat{d}_k = d_k - \bar{d}_k = P_k d_k. \tag{3.19}$$

Because of the definitions of d_k and $\bar{d}_k$, the vector $\hat{d}_k$ is a solution to the subproblem

$$\text{minimize} \quad g_k^{\mathrm{T}}(\bar{d}_k+d)+\tfrac{1}{2}(\bar{d}_k+d)^{\mathrm{T}}B_k(\bar{d}_k+d), \quad d\in\mathbb{R}^n, \tag{3.20}$$

$$\text{subject to} \quad A_k^{\mathrm{T}}d=0, \quad \|\bar{d}_k+d\|_2\leq\Delta_k. \tag{3.21}$$

Because $A_k^{\mathrm{T}}d=0$ allows d to be replaced by P_kd in this subproblem, it follows that $\hat{d}_k$ also solves the calculation

$$\text{minimize} \quad (P_k\bar{g}_k)^{\mathrm{T}}d+\tfrac{1}{2}d^{\mathrm{T}}P_kB_kP_kd, \quad d\in\mathbb{R}^n, \tag{3.22}$$

$$\text{subject to} \quad A_k^{\mathrm{T}}d=0 \quad \text{and} \quad \|d\|_2\leq\bar{\Delta}_k, \tag{3.23}$$

where the last condition depends on the orthogonality of $\hat{d}_k$ to $\bar{d}_k$. Since the addition to $\hat{d}_k$ of a vector in the column space of A_k would make no difference to the objective function (3.22) but would increase $\|\hat{d}_k\|_2$, it follows that the vector (3.19) solves the problem (3.22)–(3.23) even if the constraint $A_k^{\mathrm{T}}d=0$ is deleted. Therefore Lemma 3.2 gives the relation

$$\begin{aligned}\bar{g}_k^{\mathrm{T}}\hat{d}_k=(P_k\bar{g}_k)^{\mathrm{T}}\hat{d}_k &\leq -\frac{\|P_k\bar{g}_k\|_2^2\bar{\Delta}_k}{2\|B_k\|_2\bar{\Delta}_k+\|P_k\bar{g}_k\|_2}\\ &\leq -\tfrac{1}{2}\|P_k\bar{g}_k\|_2^2\min\left\{\frac{1}{2\|B_k\|_2},\frac{\bar{\Delta}_k}{\|P_k\bar{g}_k\|_2}\right\}.\end{aligned} \tag{3.24}$$

Hence the definitions of λ_k, $\hat{d}_k$ and $\bar{g}_k$, the fact that expression (3.22) increases monotonically between $d=\hat{d}_k$ and $d=0$, and the inequalities (3.18) and $\|\hat{d}_k\|_2\leq\|d_k\|_2$ imply the bound

$$\begin{aligned}&(g_k-A_k\lambda_k)^{\mathrm{T}}d_k+\tfrac{1}{2}d_k^{\mathrm{T}}B_k\hat{d}_k\\ &\quad=(g_k^{\mathrm{T}}+\tfrac{1}{2}d_k^{\mathrm{T}}B_k)\hat{d}_k=\tfrac{1}{2}g_k^{\mathrm{T}}\hat{d}_k+\tfrac{1}{2}\hat{d}_k^{\mathrm{T}}B_k\hat{d}_k+\tfrac{1}{2}\bar{g}_k^{\mathrm{T}}\hat{d}_k\\ &\quad\leq\tfrac{1}{2}g_k^{\mathrm{T}}\hat{d}_k\leq\tfrac{1}{2}\bar{g}_k^{\mathrm{T}}\hat{d}_k+\tfrac{1}{2}\|B_k\bar{d}_k\|_2\|\hat{d}_k\|_2\\ &\quad\leq-\tfrac{1}{4}\|P_k\bar{g}_k\|_2^2\min\left\{\frac{1}{2\|B_k\|_2},\frac{\bar{\Delta}_k}{\|P_k\bar{g}_k\|_2}\right\}+\|A_k^+\|_2\|B_k\|_2\|d_k\|_2\|c_k\|_2.\end{aligned} \tag{3.25}$$

Moreover, due to the definition of $\lambda(x)$ and Assumptions 3.1(a) and (b), there exists a positive constant m_2 such that the condition

$$\|\lambda(x_k)-\lambda(x_k+d_k)\|_2\leq m_2\|d_k\|_2, \tag{3.26}$$

holds for all k. Using elementary properties of norms to deduce from $\|c_k+A_k^{\mathrm{T}}d_k\|_2\leq\|c_k\|_2$ that $\|c_k+\frac{1}{2}A_k^{\mathrm{T}}d_k\|_2\leq\|c_k\|_2$, the inequality (3.16) now follows from the definition (2.6), Lemma 3.3, and the bounds (3.25) and (3.26), if we let $m_1=m_2+\sup_k\{\|B_k\|_2\|A_k^+\|_2\}$, which is finite due to Assumptions 3.1. □

Next the following corollary of Lemma 3.4 is used in Lemma 3.6 to establish that the penalty parameter σ_k remains bounded.

Corollary 3.5. *There exist positive constants m_3 and m_4, such that, on the iterations that satisfy the condition*

$$\|c_k\|_2 \leq m_3\Delta_k, \tag{3.27}$$

we have the inequality

$$D_k+\tfrac{1}{2}\sigma_k(\|c_k\|_2^2-\|c_k+A_k^Td_k\|_2^2)\leq -m_4\Delta_k. \tag{3.28}$$

Proof. Because Assumption 3.1(a) implies that Δ_k is uniformly bounded, we may choose m_3 to be so small that condition (3.27) implies $\|c_k\|_2\leq\frac{1}{3}\varepsilon$. Hence if the convergence test of Step 1 of the algorithm allows the calculation to continue, we have the bound

$$\|g_k-A_k\lambda_k\|_2\geq\tfrac{2}{3}\varepsilon. \tag{3.29}$$

Further, we choose m_3 small enough to yield the inequality

$$\|c_k\|_2\leq\tfrac{1}{6}\varepsilon(\sup_k\|B_k\|_2\|A_k^+\|_2)^{-1}, \tag{3.30}$$

in order that we have the relation

$$\begin{aligned}\|g_k-A_k\lambda_k\|_2&=\|P_kg_k\|_2\\&\leq\|P_k\bar{g}_k\|_2+\|P_kB_k\bar{d}_k\|_2\\&\leq\|P_k\bar{g}_k\|_2+2\|A_k^+\|_2\|B_k\|_2\|c_k\|_2\\&\leq\|P_k\bar{g}_k\|_2+\tfrac{1}{3}\varepsilon,\end{aligned} \tag{3.31}$$

which depends on the definition (3.17a) of $\bar{g}_k$ and on inequalities (3.18) and (3.30). Thus, condition (3.29) gives the bound

$$\|P_k\bar{g}_k\|_2\geq\tfrac{1}{3}\varepsilon \tag{3.32}$$

It follows from Lemma 3.4 that the inequality

$$\begin{aligned}&D_k+\tfrac{1}{2}\sigma_k(\|c_k\|_2^2-\|c_k+A_k^Td_k\|_2^2)\\&\quad\leq-\tfrac{1}{12}\varepsilon\min[\tfrac{1}{6}\varepsilon\|B_k\|_2^{-1},\bar{\Delta}_k]+m_1\|d_k\|_2\|c_k\|_2\end{aligned} \tag{3.33}$$

is satisfied. We impose the restriction $m_3\leq 0.3(\sup_k\|A_k^+\|_2)^{-1}$, in order that condition (3.27) provides $2\|A_k^+\|_2\|c_k\|_2\leq 0.6\Delta_k$, because then expressions (3.17a) and (3.18) give $\bar{\Delta}_k\geq 0.8\Delta_k$. We now see that, for sufficiently small $\|c_k\|_2$, the modulus of the first term on the right-hand side of inequality (3.33) is at least twice the modulus of the second term. Hence, by reducing m_3 again if necessary, we have the condition

$$D_k+\tfrac{1}{2}\sigma_k(\|c_k\|_2^2-\|c_k+A_k^Td_k\|_2^2)\leq-\tfrac{1}{24}\varepsilon\min[\tfrac{1}{6}\varepsilon\|B_k\|_2^{-1},0.8\Delta_k]. \tag{3.34}$$

The corollary now follows from the remark that $\|B_k\|_2^{-1}$ is bounded below by $\Delta_k/\bar{M}$, where $\bar{M}$ is any constant upper bound on the numbers $\{\|B_i\|_2\Delta_i: i=1,2,3,\ldots\}$. □

Now, using the above results, we can easily prove the boundedness of the sequence $\{\sigma_k: k=1,2,3,\ldots\}$, which is important in establishing the convergence properties of our algorithm.

Lemma 3.6. *The sequence* $\{\sigma_k: k=1, 2, 3, \ldots\}$ *remains bounded. In other words, because any increase in* σ_k *is by at least a factor of* 2, *there exists* k_1 *such that*

$$\sigma_k = \sigma_{k_1} \quad \text{for all } k \geq k_1. \tag{3.35}$$

Proof. Corollary 3.5 shows that condition (2.8) fails only if $\|c_k\|_2 > m_3\Delta_k$. In this case, using $\Delta_k \geq \|d_k\|_2$ too, Lemma 3.4 provides the bound

$$\begin{aligned} &D_k + \tfrac{1}{2}\sigma_k(\|c_k\|_2^2 - \|c_k + A_k^{\mathrm{T}} d_k\|_2^2) \\ &\quad \leq \|d_k\|_2 \|c_k\|_2 [m_1 - \tfrac{1}{2}\sigma_k \min(m_3, b_2/m_5)], \end{aligned} \tag{3.36}$$

where m_5 is an upper bound on $\{\|A_k^+\|_2: k=1, 2, 3, \ldots\}$. Hence condition (2.8) holds if σ_k is not less than the number $2m_1 \max(1/m_3, m_5/b_2)$. Therefore the number of increases in σ_k is finite. □

We now assume without loss of generality that σ_k is independent of k. Our next two lemmas show that both the trust region bound and the constraints converge to zero, if the algorithm does not terminate after finitely many iterations.

Lemma 3.7. *If the algorithm does not terminate, we have the limit*

$$\lim_{k\to\infty} \Delta_k = 0. \tag{3.37}$$

Proof. To prove the lemma, we assume that the number

$$\eta = \limsup_{k\to\infty} \Delta_k \tag{3.38}$$

is positive and deduce a contradiction. If the assumption were true there would exist an infinite subsequence $\{k(i): i=1, 2, 3, \ldots\}$ such that $\Delta_{k(i)} < 2\eta$ and $\Delta_{k(i)+1} > \frac{1}{2}\eta$ for all i. In this case condition (2.11) implies the bounds

$$r_{k(i)} \geq 0.1 \quad \text{and} \quad \Delta_{k(i)} > \eta/8. \tag{3.39}$$

Since the monotonically decreasing sequence $\{\phi(x_k): k=1, 2, 3, \ldots\}$ is convergent, the condition $r_{k(i)} \geq 0.1$ on expression (2.7) implies the limit

$$\lim_{i\to\infty} D_{k(i)} = 0. \tag{3.40}$$

We combine this remark with Corollary 3.5. If condition (3.28) held for $k(i)$ we would have $D_{k(i)} \leq -m_4\Delta_{k(i)} < -m_4\eta/8$ which is not possible for sufficiently large i. We may therefore assume that the condition of Corollary 3.5 fails, which gives $\|c_{k(i)}\|_2 > m_3\Delta_{k(i)}$. Thus, inequalities (2.8) and (3.13) imply the relation

$$\begin{aligned} D_{k(i)} &\leq \tfrac{1}{2}\sigma\|c_{k(i)}\|_2(\|c_{k(i)} + A_{k(i)}^{\mathrm{T}} d_{k(i)}\|_2 - \|c_{k(i)}\|_2) \\ &\leq -\tfrac{1}{2}\sigma m_3 \Delta_{k(i)}^2 \min[m_3, b_2/m_5], \end{aligned} \tag{3.41}$$

where m_5 is defined after expression (3.36). Now, however, inequality (3.39) contradicts the limit (3.40). Therefore the lemma is true. □

Lemma 3.8. *If the algorithm does not terminate, we have the limit*

$$\lim_{k\to\infty} \|c_k\|_2 = 0. \tag{3.42}$$

Proof. In this proof we deduce a contradiction from the assumption

$$\xi = \limsup_{k\to\infty} \|c_k\|_2 > 0. \tag{3.43}$$

Because we can restrict attention to large values of k, it follows from Lemma 3.7 that we can assume without loss of generality that the inequality

$$0.1\xi \geq b_2\Delta_k / m_5 \tag{3.44}$$

holds for all k, where m_5 is still a constant upper bound on the norms $\{\|A_k^T\|_2: k = 1, 2, 3, \ldots\}$. We extend this assumption in a way that is related to the remark that, due to the definitions (2.4) and (2.6) and the continuity of third derivatives, we have the bound

$$|D_k - [\phi_k(x_k + d_k) - \phi_k(x_k)]| \leq m_6\|d_k\|_2^2, \tag{3.45}$$

where m_6 is another positive constant. Specifically, Lemma 3.7 allows us to presume that the inequality

$$\tfrac{1}{2}\sigma(0.1\xi)b_2/m_5 \geq 2m_6\Delta_k \tag{3.46}$$

also holds for all k.

We now consider the value of D_k when k is in the infinite set $K_\xi \equiv \{k: \|c_k\|_2 \geq 0.1\xi\}$. Inequalities (3.13) and (3.44) imply the relation

$$\|c_k\|_2 - \|c_k + A_k^T d_k\|_2 \geq b_2\Delta_k / m_5, \quad k \in K_\xi, \tag{3.47}$$

so condition (2.8) yields the bound

$$D_k \leq -\tfrac{1}{2}\sigma(0.1\xi)b_2\Delta_k / m_5, \quad k \in K_\xi. \tag{3.48}$$

Hence, remembering $\|d_k\|_2 \leq \Delta_k$, it follows from expressions (3.45) and (3.46) that the inequality $\{r_k \geq 0.5: k \in K_\xi\}$ is satisfied, where r_k has the value (2.7). Thus we have the relation

$$\phi(x_k + d_k) - \phi(x_k) \leq -0.025\sigma\xi b_2\Delta_k / m_5, \quad k \in K_\xi, \tag{3.49}$$

and the definition (2.11) implies $\{\Delta_{k+1} \geq \Delta_k: k \in K_\xi\}$.

In view of Lemma 3.7, this last condition shows that there are infinitely many positive integers that are not in K_ξ. Therefore we may let $\{k(i): i = 1, 2, 3, \ldots\}$ be an infinite subsequence such that $\{\|c_{k(i)}\|_2 \geq 0.9\xi: i = 1, 2, 3, \ldots\}$ and such that any two adjacent members of the subsequence are separated by an integer that is not in K_ξ. For each i, we let $l(i)$ be the least integer such that $l(i) > k(i)$ and $l(i) \notin K_\xi$, and we consider the difference $[\phi(x_{l(i)}) - \phi(x_{k(i)})]$. Conditions (1.7) and (3.49) and the triangle inequality imply the bound

$$\phi(x_{l(i)}) - \phi(x_{k(i)}) \leq -0.025\sigma\xi b_2\|x_{l(i)} - x_{k(i)}\|_2 / m_5, \tag{3.50}$$

and the conditions $\|c(x_{k(i)})\|_2 \geq 0.9\xi$ and $\|c(x_{l(i)})\| \leq 0.1\xi$ imply that $\|x_{l(i)} - x_{k(i)}\|_2$ is bounded away from zero. These remarks, however, contradict the fact that the sequence $\{\phi(x_k)\colon k = 1, 2, 3, \ldots\}$ is monotonic and convergent. Therefore the lemma is true. □

Now we can prove our global convergence result:

Theorem 3.9. *Under Assumptions 3.1, our algorithm will terminate after finitely many iterations. In other words, if we remove the convergence test from Step 1, then $d_k = 0$ for some k or the limit*

$$\liminf_{k\to\infty}[\|c_k\|_2 + \|P_k g_k\|_2] = 0 \tag{3.51}$$

is obtained, which ensures that $\{x_k\colon k = 1, 2, 3, \ldots\}$ is not bounded away from stationary points of the problem (1.1)–(1.2).

Proof. The termination condition of the algorithm and the limit (3.42) allow the bound

$$\|P_k g_k\|_2 \geq 3\epsilon/4 \tag{3.52}$$

to be assumed without loss of generality. In view of Lemma 3.7 we may also assume $\|B_k \bar{d}_k\| \leq \epsilon/4$, in order that expressions (3.17a) and (3.52) imply the inequality

$$\|P_k \bar{g}_k\|_2 \geq \tfrac{1}{2}\epsilon. \tag{3.53}$$

We suppose that the algorithm fails to terminate and deduce a contradiction.

The first part of the proof will show that the ratios $\{\|P_k d_k\|_2/\Delta_k\colon k = 1, 2, 3, \ldots\}$ cannot stay bounded away from zero. Then we pick an infinite subsequence $\{k(i)\colon i = 1, 2, 3, \ldots\}$ for which these ratios tend to zero. Hence we find a particular sequence $\{\bar{d}_{k(i)}\colon i = 1, 2, 3, \ldots\}$ of vectors in $\mathbb{R}^n$ that satisfies the constraints (2.2) on d for sufficiently large i. The contradiction is that, for large i, the value of the objective function (2.1) when $d = \bar{d}_{k(i)}$ is less than the value that occurs when $d = d_{k(i)}$. This construction depends on $b_1 < 1$ in the left-hand inequality of expression (2.3).

Lemma 3.4 and condition (3.53) imply the bound

$$D_k \leq m_1\|d_k\|_2\|c_k\|_2 - \tfrac{1}{8}\epsilon \min\left\{\frac{\epsilon}{4\|B_k\|_2}, \bar{\Delta}_k\right\}. \tag{3.54}$$

Hence, because $\|d_k\|_2 \leq \Delta_k$ and $\|c_k\|_2 \to 0$, if $\bar{\Delta}_k/\Delta_k$ were bounded away from zero, then, for sufficiently large k, D_k would be bounded above by a negative multiple of Δ_k. It would follow from expression (3.45), however, that the ratio (2.7) would tend to one, and then we would have $\Delta_{k+1} \geq \Delta_k$ for all sufficiently large k, contradicting Lemma 3.7. Therefore the algorithm gives the limit

$$\liminf_{k\to\infty}(\bar{\Delta}_k/\Delta_k) = 0. \tag{3.55}$$

Now the relation $\bar{\Delta}_k \geq \|P_k d_k\|_2$ is an elementary consequence of expressions (1.7), (3.17b) and (3.17c). Therefore, as asserted earlier, there is a subsequence $\{k(i): i = 1, 2, 3, \ldots\}$ of positive integers that provides the equation

$$\lim_{i\to\infty} \|P_{k(i)} d_{k(i)}\|_2 / \Delta_{k(i)} = 0. \tag{3.56}$$

The vector $\tilde{d}_{k(i)}$ has the form

$$\tilde{d}_{k(i)} = \tau d^0_{k(i)} + (1-\tau) d_{k(i)} - m_7 \tau \Delta_{k(i)} P_{k(i)} g_{k(i)}, \tag{3.57}$$

where τ is a constant from the interval $(0, 1)$ that will be chosen later, where $d^0_{k(i)}$ is any vector that satisfies the conditions

$$\|d^0_{k(i)}\|_2 \leq b_1 \Delta_{k(i)}, \tag{3.58a}$$

$$\|c_{k(i)} + A^{\mathrm{T}}_{k(i)} d^0_{k(i)}\|_2 \leq \zeta_{k(i)}, \tag{3.58b}$$

whose existence is a consequence of inequality (2.3), and where m_7 is the constant

$$m_7 = 4 \sup \|g_k\|_2 / \varepsilon^2. \tag{3.59}$$

Because $0 < \tau < 1$ and because $A^{\mathrm{T}}_{k(i)} P_{k(i)} = 0$, it follows from expressions (2.2), (3.57) and (3.58b) that the constraint $\|c_{k(i)} + A^{\mathrm{T}}_{k(i)} d\|_2 \leq \zeta_{k(i)}$ is achieved by $d = \tilde{d}_{k(i)}$. We see that $\|\tilde{d}_{k(i)}\|_2 \leq \Delta_{k(i)}$ is satisfied too if we have the inequality

$$\|(1-\tau) d_{k(i)} - m_7 \tau \Delta_{k(i)} P_{k(i)} g_{k(i)}\|_2 \leq (1 - b_1 \tau) \Delta_{k(i)}. \tag{3.60}$$

We square both sides of this expression, we rearrange terms and we employ the bounds $\|d_{k(i)}\|_2 \leq \Delta_{k(i)}$ and $\|P_{k(i)} g_{k(i)}\|_2 \leq \|g_{k(i)}\|_2$. Thus, we find that the condition

$$\begin{aligned} &-2 m_7 (1-\tau) d^{\mathrm{T}}_{k(i)} P_{k(i)} g_{k(i)} + m_7^2 \tau \Delta_{k(i)} \|g_{k(i)}\|_2^2 \\ &\quad \leq (1 - b_1)(2 - \tau - b_1 \tau) \Delta_{k(i)} \end{aligned} \tag{3.61}$$

is sufficient for $\|\tilde{d}_{k(i)}\|_2 \leq \Delta_{k(i)}$. Now the Cauchy–Schwarz inequality

$$|d^{\mathrm{T}}_{k(i)} P_{k(i)} g_{k(i)}| \leq \|P_{k(i)} d_{k(i)}\|_2 \|g_{k(i)}\|_2 \tag{3.62}$$

and the limit (3.56) show that the first term of expression (3.61) tends to be much smaller than the other two terms. Therefore we obtain the required conditions on $\tilde{d}_{k(i)}$ for sufficiently large i by choosing τ to be any constant from the open interval $(0, 1)$ that satisfies the relation

$$m_7^2 \tau \sup_k \|g_k\|_2^2 < (1 - b_1)(2 - \tau - b_1 \tau), \tag{3.63}$$

so it is sufficient to let τ be small and positive.

Finally we recall that, because of the definition of d_k in the algorithm, the condition

$$g^{\mathrm{T}}_{k(i)} d_{k(i)} + \tfrac{1}{2} d^{\mathrm{T}}_{k(i)} B_{k(i)} d_{k(i)} \leq g^{\mathrm{T}}_{k(i)} \tilde{d}_{k(i)} + \tfrac{1}{2} \tilde{d}^{\mathrm{T}}_{k(i)} B_{k(i)} \tilde{d}_{k(i)} \tag{3.64}$$

should hold when i is so large that $\|\tilde{d}_{k(i)}\|_2 \leq \Delta_{k(i)}$. Equation (3.57) gives the value

$$g^{\mathrm{T}}_{k(i)} (d_{k(i)} - \tilde{d}_{k(i)}) = \tau g^{\mathrm{T}}_{k(i)} (d_{k(i)} - d^0_{k(i)}) + m_7 \tau \Delta_{k(i)} \|P_{k(i)} g_{k(i)}\|_2^2. \tag{3.65}$$

Hence, using the elementary bound

$$g_{k(i)}^{\mathrm{T}}(d_{k(i)}-d_{k(i)}^{0}) \geq -2\|g_{k(i)}\|_2\Delta_{k(i)}, \tag{3.66}$$

the definition (3.59) of m_7 and inequality (3.52), we deduce the relation

$$\begin{aligned} g_{k(i)}^{\mathrm{T}}(d_{k(i)}-\tilde{d}_{k(i)}) &\geq [-2\|g_{k(i)}\|_2+9\sup\|g_k\|_2/4]\tau\Delta_{k(i)} \\ &\geq \sup\|g_k\|_2\tau\Delta_{k(i)}/4. \end{aligned} \tag{3.67}$$

Since the second-order terms of expression (3.64) are bounded by $\Delta_{k(i)}^2\sup\|B_k\|$, and since $\Delta_{k(i)}$ tends to zero as $i\to\infty$, it follows that condition (3.64) fails for sufficiently large i. Therefore the theorem is true. □

4. Local superlinear convergence

In this section we analyse the rate of convergence of the algorithm when $\epsilon = 0$ and when the sequence $\{x_k\}$ converges to a point x^*. Theorem 3.9 shows that x^* is a Kuhn–Tucker point. We require both the second-order sufficiency condition and the assumption on the accuracy of the matrices $\{B_i: k = 1, 2, 3, \ldots\}$ that are given below.

Assumptions 4.1. (a) $x_k \to x^*$;

(b) There exists a constant $m_8 > 0$ such that the inequality

$$d^{\mathrm{T}}W^*d \geq m_8\|d\|_2^2 \tag{4.1}$$

holds for all d satisfying $A(x^*)^{\mathrm{T}}d = 0$, where W^* is the matrix

$$W^* = \nabla^2 f(x^*) - \sum_{i=1}^{m}\lambda_i^*\nabla^2 c_i(x^*), \tag{4.2}$$

and where the Lagrange multipliers $\{\lambda_i^*: i = 1, 2, \ldots, m\}$ are defined by the equation

$$\nabla f(x^*) = \sum_{i=1}^{m}\lambda_i^*\nabla c_i(x^*). \tag{4.3}$$

Assumption 4.2.

$$\lim_{k\to\infty}\max_{A_k^{\mathrm{T}}d=0,\ \|d\|_2=1}|d^{\mathrm{T}}(B_k - W^*)d_k|/\|d_k\|_2 = 0. \tag{4.4}$$

Assumption 4.1(a) is not very restrictive because, when Assumption 4.1(b) holds, the exact differentiable penalty function (2.4) has a local minimum at x^* for sufficiently large σ_{k_0} and it is usual for this local minimum to trap the sequence $\{x_k: k = 1, 2, 3, \ldots\}$. In this case, assuming that the final value of σ_k is large enough, Theorem 3.9 and the monotonicity of the sequence $\{\phi_k(x_k): k = 1, 2, 3, \ldots\}$ after σ_k achieves its final value imply that $\{x_k: k = 1, 2, 3, \ldots\}$ does converge to x^*.

It is proved in Boggs, Tolle and Wang (1982) and Powell (1983) that, if d_k minimizes the function (1.4) subject to the constraints (1.5), and if $x_{k+1} = x_k + d_k$ for all sufficiently large k, then the rate of convergence of the sequence $\{x_k: k = 1, 2, 3, \ldots\}$ is superlinear if and only if condition (4.4) holds. Therefore, making the given assumptions, the purpose of this section is to show that our use of trust regions does not invalidate this superlinear convergence property. First we establish the limit

$$\lim_{k\to\infty} \|d_k\|_2 = 0. \tag{4.5}$$

and then, using four lemmas, it is shown that $r_k \to 1$, so formula (2.11) keeps Δ_k bounded away from zero. Thus, the limit $\|c_k\|_2 \to 0$ and expression (2.3) imply $\zeta_k = 0$ for all sufficiently large k. It follows from the limit (4.5) that eventually every d_k is the search direction of the analysis of Boggs, Tolle and Wang (1982) and Powell (1983), which gives the required superlinear rate of convergence. One complication, addressed in Lemma 4.4, is that, although Lemmas 3.2–3.4 still hold, we have to re-establish Lemma 3.6, because the method of proof in Section 3 depends on $\varepsilon > 0$ in Step 1 of the algorithm, but now we are investigating the case when $\varepsilon = 0$.

Expression (4.5) can be deduced from Lemma 4.3 below, but the following proof is more elementary. Because this limit follows from $\|d_k\|_2 \leq \Delta_k$ if $\Delta_k \to 0$, we assume $\limsup \Delta_k = \eta > 0$. For all sufficiently large k we have $\|x_{k+1} - x_k\| \leq \eta/8$ by Assumption 4.1(a), which implies $\Delta_{k+1} \leq \max[\Delta_k, \frac{1}{2}\eta]$, but when $x_{k+1} = x_k$ formula (2.11) gives $\Delta_{k+1} \leq \Delta_k/4$. Therefore $\limsup \Delta_k = \eta > 0$ is possible only if the number of iterations that set $x_{k+1} = x_k$ is finite. Hence $x_{k+1} = x_k + d_k$ for all sufficiently large k, so the limit (4.5) is a consequence of Assumption 4.1(a).

Lemma 4.3. *The algorithm gives the bound*

$$\|d_k\|_2 = O(\|c_k\|_2 + \|P_k g_k\|_2) \tag{4.6}$$

on the trial steps.

Proof. Because the vector (3.19) satisfies $A_k^T \hat{d}_k = 0$, Assumption 4.2 provides the equation

$$|\hat{d}_k^T B_k d_k - \hat{d}_k^T W^* d_k| = o(\|d_k\|_2 \|\hat{d}_k\|_2). \tag{4.7}$$

Therefore by substituting $d_k = \hat{d}_k + \bar{d}_k$ we find the relation

$$|\hat{d}_k^T B_k \hat{d}_k - \hat{d}_k^T W^* \hat{d}_k| = o(\|\hat{d}_k\|_2^2) + O(\|\bar{d}_k\|_2 \|\hat{d}_k\|_2). \tag{4.8}$$

Thus, from condition (4.1), from the non-positivity of expression (3.22) when $d = \hat{d}_k$, and from the definition (3.17a) of $\bar{g}_k$, we deduce the inequality

$$\begin{aligned} \tfrac{1}{2} m_8 \|\hat{d}_k\|_2^2 &\leq \hat{d}_k^T B_k \hat{d}_k + O(\|\bar{d}_k\|_2 \|\hat{d}_k\|_2) \\ &\leq 2\|P_k \bar{g}_k\|_2 \|\hat{d}_k\|_2 + O(\|\bar{d}_k\|_2 \|\hat{d}_k\|_2) \\ &= 2\|P_k g_k\|_2 \|\hat{d}_k\|_2 + O(\|\bar{d}_k\|_2 \|\hat{d}_k\|_2) \end{aligned} \tag{4.9}$$

for sufficiently large k. It follows from $\|d_k\|_2 \leq \|\hat{d}_k\|_2 + \|\bar{d}_k\|_2$ that we have the bound

$$\tfrac{1}{2}m_8\|d_k\|_2 \leq 2\|P_k g_k\|_2 + O(\|\bar{d}_k\|_2). \tag{4.10}$$

Condition (4.6) is now a consequence of inequality (3.18), so the lemma is true. □

Next it is shown that the sequence $\{\sigma_k: k = 1, 2, 3, \ldots\}$ is bounded above under the second order sufficiency condition. In other words Lemma 3.6 remains valid if we remove the convergence test in Step 1 of the algorithm.

Lemma 4.4. *There exists an integer k_2 such that*

$$\sigma_k = \sigma_{k_2} \quad \text{for all } k \geq k_2. \tag{4.11}$$

Proof. As in the proof of Lemma 3.6, it is sufficient to show that the right-hand side of expression (3.16) is negative if k and σ_k are sufficiently large. First, we establish the existence of positive constants m_9 and k_3 such that the conditions

$$\|c_k\|_2 \leq m_9\|d_k\|_2, \quad k \geq k_3, \tag{4.12}$$

imply that the sum of the first two of the three terms on the right-hand side of expression (3.16) is non-positive. No details of the remainder of the proof are given, as the alternative case when $\|c_k\|_2 > m_9\|d_k\|_2$ can be treated by replacing m_3 by m_9 throughout the proof of Lemma 3.6.

We let k_3 be any constant integer such that the middle line of inequality (4.9) provides the bound

$$\begin{aligned}\|P_k\bar{g}_k\|_2 &\geq \tfrac{1}{4}m_8\|\hat{d}_k\|_2 - m_{10}\|\bar{d}_k\|_2 \\ &\geq \tfrac{1}{4}m_8\|d_k\|_2 - (\tfrac{1}{4}m_8 + m_{10})\|\bar{d}_k\|_2, \quad k \geq k_3.\end{aligned} \tag{4.13}$$

where m_{10} is another positive constant, and where the last line is derived from the triangle inequality. Therefore, letting m_9 satisfy the constraint,

$$m_9 \leq m_8/[(4m_8 + 16m_{10})\sup\|A_k^+\|_2], \tag{4.14}$$

expressions (3.18), (4.12) and (4.13) give the relation

$$\|P_k\bar{g}_k\|_2 \geq \tfrac{1}{8}m_8\|d_k\|_2. \tag{4.15}$$

Further, following the argument that comes immediately after inequality (3.33), we impose $m_9 \leq 0.3(\sup\|A_k^+\|_2)^{-1}$, in order that $\bar{\Delta}_k \geq 0.8\Delta_k$. Thus, when the conditions (4.12) are satisfied, the first term on the right-hand side of expression (3.16) is bounded above by the negative number

$$-\|d_k\|_2^2 \min\left\{\frac{m_8^2}{512\sup\|B_k\|_2}, \frac{m_8}{40}\right\} = -m_{11}\|d_k\|_2^2, \tag{4.16}$$

say. Therefore the right-hand side of expression (3.16) is negative as required if we also impose the constraint $m_9 \leq m_{11}/m_1$. The proof is completed in the way that is mentioned after inequality (4.12). □

Our next lemma gives a lower bound for the predicted reduction in the penalty function.

Lemma 4.5. *There exists a positive constant m_{12} such that the inequality*

$$D_k \leq -m_{12}\|d_k\|_2(\|d_k\|_2+\|c_k\|_2) \tag{4.17}$$

holds for all sufficiently large k.

Proof. Let m_{13} be a positive constant that satisfies the condition

$$m_{13} \leq \min[m_3, m_{11}/(2m_1+m_{11})], \tag{4.18}$$

where m_3 and $\hat{m}_{11}$ occur in the proof of Lemma 4.4. Thus the inequalities

$$\|c_k\|_2 \leq m_{13}\|d_k\|_2, \quad k \geq k_3, \tag{4.19}$$

imply the relation

$$\begin{aligned} D_k &\leq -m_{11}\|d_k\|_2^2+m_1\|d_k\|_2\|c_k\|_2 \\ &\leq -\tfrac{1}{2}m_{11}\|d_k\|_2[\|d_k\|_2+\|c_k\|_2], \end{aligned} \tag{4.20}$$

where the first line depends on expressions (3.16) and (4.16), and where the second line is an elementary consequence of the constraint $\|c_k\|_2 \leq m_{11}\|d_k\|_2/(2m_1+m_{11})$ which is given by the bounds (4.18) and (4.19). Therefore, in the case (4.19), we achieve the required condition (4.17) by choosing $m_{12} \leq \frac{1}{2}m_{11}$.

Alternatively, when we have the relation

$$\|c_k\|_2 > m_{13}\|d_k\|_2, \tag{4.21}$$

we make use of the fact that, due to Lemma 4.4, inequality (2.8) holds for all large k. Thus, using inequality (3.13) also, we deduce the bound

$$\begin{aligned} D_k &\leq -\tfrac{1}{2}\sigma_k(\|c_k\|_2^2-\|c_k+A_k^{\mathrm{T}}d_k\|_2^2) \\ &\leq -\tfrac{1}{2}\sigma_k\|c_k\|_2 \min\left\{\|c_k\|_2, \frac{b_2\Delta_k}{\|A_k^+\|_2}\right\} \\ &\leq -\tfrac{1}{2}\sigma_k \min\left\{m_{13}, \frac{b_2}{\|A_k^+\|_2}\right\}\|c_k\|_2\|d_k\|_2 \\ &\leq -\tfrac{1}{2}\sigma_k \min\left\{m_{13}, \frac{b_2}{\|A_k^+\|_2}\right\}\frac{m_{13}}{1+m_{13}}\|d_k\|_2(\|d_k\|_2+\|c_k\|_2), \end{aligned} \tag{4.22}$$

where the last line is an elementary consequence of condition (4.21). Therefore the lemma is true. □

Lemma 4.6. *Under our assumptions, we have the limit*

$$\lim_{k\to\infty} r_k = 1. \tag{4.23}$$

Proof. The definitions (2.4) and (2.6) and Assumption 4.1(a) imply the relation

$$\begin{aligned}
&\phi_k(x_k+d_k)-\phi_k(x_k)-D_k\\
&\quad=(g_k-A_k\lambda_k)^{\mathrm{T}}d_k+\tfrac{1}{2}d_k^{\mathrm{T}}W^*d_k\\
&\qquad-[\lambda(x_k+d_k)-\lambda_k]^{\mathrm{T}}(c_k+A_k^{\mathrm{T}}d_k)-\sigma_k(\|c_k\|_2^2-\|c_k+A_k^{\mathrm{T}}d_k\|_2^2)\\
&\qquad-(g_k-A_k\lambda_k)^{\mathrm{T}}d_k-\tfrac{1}{2}d_k^{\mathrm{T}}B_k\hat{d}_k+[\lambda(x_k+d_k)-\lambda_k]^{\mathrm{T}}(c_k+\tfrac{1}{2}A_k^{\mathrm{T}}d_k)\\
&\qquad+\sigma_k(\|c_k\|_2^2-\|c_k+A_k^{\mathrm{T}}d_k\|_2^2)+o(\|d_k\|_2[\|d_k\|_2+\|c_k\|_2])\\
&\quad=\tfrac{1}{2}d_k^{\mathrm{T}}W^*d_k-\tfrac{1}{2}d_k^{\mathrm{T}}B_k\hat{d}_k\\
&\qquad-\tfrac{1}{2}[\lambda(x_k+d_k)-\lambda_k]^{\mathrm{T}}A_k^{\mathrm{T}}d_k+o(\|d_k\|_2[\|d_k\|_2+\|c_k\|_2]).
\end{aligned}\tag{4.24}$$

By differentiating the normal equation

$$A(x)^{\mathrm{T}}(g(x)-A(x)\lambda(x))=0 \tag{4.25}$$

at $x=x^*$, we obtain the identity

$$\nabla(\lambda(x^*))^{\mathrm{T}}=(A(x^*))^{+}W^*. \tag{4.26}$$

which gives the equation

$$\begin{aligned}
[\lambda(x_k+d_k)-\lambda(x_k)]^{\mathrm{T}}A_k^{\mathrm{T}}d_k&=d_k^{\mathrm{T}}W^*[(A(x^*))^{+}]^{\mathrm{T}}A_k^{\mathrm{T}}d_k+o(\|d_k\|_2^2)\\
&=d_k^{\mathrm{T}}W^*\bar{d}_k+o(\|d_k\|_2^2).
\end{aligned}\tag{4.27}$$

We use expressions (4.24), (4.27) and (4.4) to deduce the bound

$$\begin{aligned}
\phi_k(x_k+d_k)-\phi_k(x_k)-D_k&=\tfrac{1}{2}d_k^{\mathrm{T}}(W^*-B_k)\hat{d}_k+o(\|d_k\|_2[\|d_k\|_2+\|c_k\|_2])\\
&=o(\|d_k\|_2[\|d_k\|_2+\|c_k\|_2]).
\end{aligned}\tag{4.28}$$

Thus, the limit (4.23) follows from Lemma 4.5. □

Lemmas 4.3 and 4.6 provide our superlinear convergence result.

Theorem 4.7. *If Assumptions* 3.1, 4.1 *and* 4.2 *are satisfied, then* $\{x_k\}$ *generated by the algorithm converges to* x^* *superlinearly.*

Proof. Due to Lemma 4.6 and equation (2.11), the sequence $\{\Delta_k: k=1,2,3,\ldots\}$ is bounded away from zero. It follows from equation (4.5) that the trust region bound becomes inactive for sufficiently large k. Thus, for all large k, d_k is the solution of the problem

$$\text{minimize}\quad g_k^{\mathrm{T}}d+\tfrac{1}{2}d^{\mathrm{T}}B_kd,\quad d\in\mathbb{R}^n, \tag{4.29}$$

$$\text{subject to}\quad c_k+A_k^{\mathrm{T}}d=0. \tag{4.30}$$

Therefore superlinear convergence is a consequence of the theory of Boggs, Tolle and Wang (1982) and Powell (1983). □

5. Discussion

The given analysis suggests that, when solving optimization problems with nonlinear equality constraints, it is possible to combine the differentiable exact penalty function (2.4) with the use of trust regions in a way that gives good theoretical convergence properties. This conclusion may be important, because trust regions have some advantages over line searches when there are nonlinear constraints. In particular, if nonlinearities make it necessary for $\|x_{k+1}-x_k\|$ to be very small when x_k is far from a Kuhn–Tucker point, then a line search algorithm would use a step length that is much less than one, so, even if linear approximations to constraints are quite accurate, any constraint violations tend to zero only at a slow linear rate, at least until it becomes possible to take unit steps along search directions. With trust regions, however, when $x_{k+1} \neq x_k$, the step from x_k to x_{k+1} can include a component of length at least $b_2\Delta_k$ that is used primarily to reduce constraint violations. In other words because a trust region step is the full d_k that is calculated even when Δ_k is small, one gains from all of the linear approximation to a constraint that occurs in the definition of d_k, but a line search algorithm introduces the step length as a scaling factor on the predicted reductions in constraint violations.

The advantages of preferring a penalty function of the form (2.4) over a non-differentiable one are well known, for example see Bertsekas (1982). The need to calculate $\lambda(x)$ for any x, however, introduces severe difficulties if the constraint gradients become linearly dependent. This is unlikely to happen if there are fewer constraints than variables, but a possible remedy is to change the definition of $\lambda(x)$ to the vector that minimizes the expression

$$\|g(x)-A(x)\lambda\|_2^2+\theta\|c(x)\|_2^2\|\lambda\|_2^2, \quad \lambda\in\mathbb{R}^m, \tag{5.1}$$

where θ is a prescribed positive constant. Thus the calculation breaks down only if the constraint gradients are linearly dependent and all the constraints are satisfied. Our theory does not apply to this new choice of $\lambda(x)$, but the main results may still hold. Indeed $\{\lambda(x): x\in\mathbb{R}^n\}$ remains differentiable, the new value of $\|\lambda(x)\|_2$ is no greater than before, and the change to $\lambda(x)$ due to $\theta>0$ in expression (5.1) is of magnitude $\|x-x^*\|_2^2$ when x is close to a feasible point x^*, provided that constraint gradients are linearly independent at all feasible points.

In order to find a value of ζ_k in the interval (2.3), one may begin by calculating the shortest vector $\bar{d}$ that satisfies $c_k+A_k^T\bar{d}=0$, which is straightforward if one preserves the QR factorization of A_k from the computation of λ_k. It is suitable to set $\zeta_k=0$ if $\|\bar{d}\|_2\leq b_1\Delta_k$, but otherwise ζ_k must have the value

$$\zeta_k=\|c_k+A_k^Td(\theta)\|_2, \tag{5.2}$$

where θ is any positive parameter that provides the conditions

$$b_2\Delta_k\leq\|d(\theta)\|_2\leq b_1\Delta_k, \tag{5.3}$$

and where $d(\theta)$ minimizes the expression

$$\|c_k+A_k^Td\|_2^2+\theta\|d\|_2^2, \quad d\in\mathbb{R}^n. \tag{5.4}$$

Because $d(\theta)$ has to lie in the column space of A_k, the adjustment of θ is similar to the trust region calculation that is considered on pages 104–106 of Fletcher (1987). Of course it is helpful to choose $b_2 < b_1$.

When $\zeta_k = 0$, one can complete the computation of d_k by another similar trust region calculation, because the required trial step has the form

$$d_k = \bar{d}_k + \hat{d}_k, \tag{5.5}$$

where $\bar{d}_k = \bar{d}$ is defined in the previous paragraph (so its value is known) and where $\hat{d}_k$ is from the null space of A_k^T. Specifically, one adjust $\hat{d}_k$ so that the vector (5.5) minimizes the quadratic function (2.1), the trust region constraint being the inequality

$$\|\hat{d}_k\|_2 \leq (\Delta_k^2 - \|\bar{d}_k\|_2^2)^{1/2}. \tag{5.6}$$

This calculation is equivalent to the generation of trial steps in trust region algorithms for unconstrained optimization, because one can express $\hat{d}_k$ in terms of an orthogonal basis of the null space of A_k^T.

We could also apply this method to the $\zeta_k > 0$ case if we satisfied the constraint $\|c_k + A_k^T d\|_2 \leq \zeta_k$ by imposing the condition $A_k^T d_k = A_k^T d(\theta)$, where $d(\theta)$ occurs in equation (5.2) and is known. We would replace $\bar{d}_k$ by $d(\theta)$ throughout the previous paragraph. In general, however, one can achieve a smaller value of the objective function (2.1) by not making this simplification. It is therefore desirable, both in our algorithm and in the procedure of Celis et al. (1985), to find a suitable technique for minimizing the function (2.1) subject to both of the constraints (2.2).

In practice one requires tolerances on this subproblem in order that it can be solved in a finite number of computer operations, but one should ensure that the tolerances preserve the convergence properties of Sections 3 and 4. The following technique is shown to be suitable in Powell and Yuan (1986b), which is the original version of the present paper. We let b_0 and b_3 be any constants that satisfy $0 < b_0 \leq 1$ and $b_3 \geq 1$. Then it is sufficient if, instead of the conditions (2.2), the vector d_k minimizes the function (2.1) subject to the constraints

$$\|c_k + A_k^T d\|_2 \leq \tilde{\zeta}_k \quad \text{and} \quad \|d\|_2 \leq \tilde{\Delta}_k, \tag{5.7}$$

where $\tilde{\zeta}_k$ and $\tilde{\Delta}_k$ are *any* numbers from the intervals $[\zeta_k, b_0\zeta_k + (1-b_0)\|c_k\|_2]$ and $[\Delta_k, b_3\Delta_k]$, respectively, except that we require $\tilde{\zeta}_k = \zeta_k$ when $\zeta_k = 0$, which is the case that has been addressed already in the paragraph that includes equation (5.5).

The form (5.7) of the constraints is convenient in practice, because the Lagrangian function of the calculation of the search direction is the expression

$$L_k(d) \equiv g_k^T d + \tfrac{1}{2} d^T B_k d + \tfrac{1}{2}\psi_1 \|c_k + A_k^T d\|_2^2 + \tfrac{1}{2}\psi_2 \|d\|_2^2, \quad d \in \mathbb{R}^n. \tag{5.8}$$

Therefore, if for any choice of $\psi_1 \geq 0$ and $\psi_2 \geq 0$ we solve the linear system of equations $\nabla L_k(d) = 0$, then we find a stationary point of the function (2.1) subject to the constraints (5.7), where $\tilde{\zeta}_k$ and $\tilde{\Delta}_k$ are the values of $\|c_k + A_k^T d\|_2$ and $\|d\|_2$ that occur for the d that has just been calculated. Further, there is usually a range of values of ψ_1 and ψ_2 that yields a d_k that satisfies the conditions of the previous

paragraph. Thus, it should be possible to develop finite procedures for generating search directions that give the convergence properties that have been presented. It has been mentioned already that a suitable algorithm is proposed in Yuan (1988) for the case when B_k is positive definite.

Partly because we have not yet decided how to adjust ψ_1 and ψ_2, our algorithm has not been used yet for any numerical computations. However, it was developed from an earlier procedure that has been tested. The main difference from the present method is that, on every iteration of the earlier procedure, d_k is calculated to minimize $\{\|c_k + A_k^T d\|_2 : d \in \mathbb{R}^n\}$ subject to $\|d\|_2 \leq \Delta_k$, so the objective function (2.1) is relevant only if some freedom remains in d after minimizing $\|c_k + A_k^T d\|_2$. Numerical results from the earlier procedure were presented by Yuan in an unpublished paper at the 1985 Dundee Numerical Analysis conference. They compare favourably with the results of the algorithm of Powell and Yuan (1986a) that uses line searches instead of trust regions.

Another question that has to be answered in order to provide numerical results is the choice of the matrices $\{B_k : k = 1, 2, 3, \ldots\}$. Powell and Yuan (1986a) obtained B_{k+1} by applying the BFGS formula to the change in gradient of the Lagrangian function along d_k, using the Lagrange multiplier estimates λ_k and λ_{k+1}, and including a device that preserves positive definiteness. An advantage of trust regions, however, is that one can give up positive definiteness, which is particularly helpful when each B_k is to be sparse, because sparseness conditions on B_k make it more difficult to preserve positive definiteness, even in unconstrained calculations with exact line searches (Sorensen, 1981). Therefore the generality of B_k in our theoretical analysis seems to be useful.

Section 3 is entirely suitable to the stopping condition of the algorithm, and Section 4 is of practical value, because it shows that convergence to a Kuhn–Tucker point cannot be impaired by the Maratos effect when the given conditions hold. It is possible, however, that the sequence $\{x_k : k = 1, 2, 3, \ldots\}$ would not be convergent if ε were set to zero. We doubt the value of modifications to rule out this behaviour, because the stopping condition in Step 1 does not need them. A usual remedy is to accept a trial step d_k only if it gives a "sufficient decrease" in the merit function (2.4), the least acceptable decrease being proportional to Δ_k times the Kuhn–Tucker residual term $[\|c_k\|_2 + \|g_k - A_k\lambda_k\|_2]$. Convergence analysis becomes much easier in this case, and one may be able to prove under our assumptions that all limit points of $\{x_k : k = 1, 2, 3, \ldots\}$ would be Kuhn–Tucker points if ε were set to zero. We believe, however, that d_k should be accepted whenever $\phi_k(x_k + d_k) < \phi_k(x_k)$, because the main purpose of merit functions is to compare estimates of the solution x^*. A striking example of the disadvantage of the "sufficient decrease" criterion is that it can prevent the step to the true solution in an unconstrained calculation when there is only one variable.

Practical experience would be needed to decide on suitable values of the constants b_0, b_1, b_2 and b_3, that help to make the calculation of the trial step finite, and that balance the corrections to constraint violations with the reduction in the objective

function. It is encouraging, however, that our theory seems to allow many suitable choices of these parameters.

Acknowledgment

The method of proof of Lemma 3.7 was suggested by a referee. We are very grateful for this improvement to the original analysis.

References

D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (Academic Press, New York, 1982).

M.C. Biggs, "On the convergence of some constrained minimization algorithms based on recursive quadratic programming," *Journal of the Institute of Mathematics and its Applications* 21 (1978) 67–82.

M.C. Biggs, "A recursive quadratic programming algorithm based on the augmented Lagrangian function," Technical Report 139, Numerical Optimisation Centre, Hatfield Polytechnic (Hatfield, England, 1983).

P.T. Boggs, J.W. Tolle and P. Wang, "On the local convergence of quasi-Newton methods for constrained optimization," *SIAM Journal on Control and Optimization* 20 (1982) 161–171.

R.H. Byrd, R.B. Schnabel and G.A. Shultz, "A trust region algorithm for nonlinearly constrained optimization," Technical Report CU-CS-313-85, University of Colorado (Boulder, CO, 1985).

M.R. Celis, J.E. Dennis and R.A. Tapia, "A trust region strategy for nonlinear equality constrained optimization," in: P.T. Boggs, R.H. Byrd and R.B. Schnabel, eds., *Numerical Optimization 1984* (SIAM, Philadelphia, 1985) pp. 71–82.

R. Fletcher, *Practical Methods of Optimization* (John Wiley and Sons, Chichester, 1987, 2nd ed.).

D.M. Gay, "Computing optimal locally constrained steps," *SIAM Journal on Scientific and Statistical Computing* 2 (1981) 186–197.

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

M.J.D. Powell, "Convergence properties of a class of minimization algorithms," in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds. *Nonlinear Programming, Vol. 2* (Academic Press, New York, 1975) pp. 1–27.

M.J.D. Powell, "The convergence of variable metric methods for nonlinearly constrained optimization," in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., *Nonlinear Programming, Vol. 3* (Academic Press, New York, 1978) pp. 27–63.

M.J.D. Powell, "Variable metric methods for constrained optimization," in: A. Bachem, M. Grötschel and B. Korte, eds., *Mathematical Programming, The State of the Art* (Springer-Verlag, Berlin, 1983) pp. 288–311.

M.J.D. Powell, "The performance of two subroutines for constrained optimization on some difficult test problems," in: P.T. Boggs, R.H. Byrd and R.B. Schnabel, eds., *Numerical Optimization 1984* (SIAM, Philadelphia, 1985) pp. 160–177.

M.J.D. Powell and Y. Yuan, "A recursive quadratic programming algorithm for equality constrained optimization," *Mathematical Programming* 35 (1986a) 265–278.

M.J.D. Powell and Y. Yuan, "A trust region algorithm for equality constrained optimization," Report DAMTP 1986/NA2, University of Cambridge, 1986b.

K. Schittkowski, "The nonlinear programming method of Wilson, Han, and Powell with an augmented Lagrangian type line search function, Part I: convergence analysis," *Numerische Mathematik* 38 (1981) 83–114.

K. Schittkowski, "On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function," *Mathematische Operationsforschung und Statistik, Series Optimization* 14 (1983) 197–216.

D.C. Sorensen, "An example concerning quasi-Newton estimates of a sparse Hessian," *SIGNUM Newsletter* 16, No. 2 (1981) 8–10.

D.C. Sorensen, "Trust region methods for unconstrained optimization," in: M.J.D. Powell, ed., *Nonlinear Optimization 1981* (Academic Press, London, 1982) pp. 29–38.

A. Vardi, "A trust region algorithm for equality constrained minimization: convergence properties and implementation," *SIAM Journal on Numerical Analysis* 22 (1985) 575–591.

Y. Yuan, "Conditions for convergence of trust region algorithms for nonsmooth optimization," *Mathematical Programming* 31 (1985) 220–228.

Y. Yuan, "A dual algorithm for minimizing a quadratic function with two quadratic constraints," Report DAMTP 1988/NA3, University of Cambridge (Cambridge, 1988).