SIAM J. NUMER. ANAL.
Vol. 25, No. 2, April 1988
© 1988 Society for Industrial and Applied Mathematics
012

GLOBAL CONVERGENCE OF A CLASS OF TRUST REGION ALGORITHMS FOR OPTIMIZATION WITH SIMPLE BOUNDS*

A. R. CONN†, N. I. M. GOULD‡, AND PH. L. TOINT§

Abstract. This paper extends the known excellent global convergence properties of trust region algorithms for unconstrained optimization to the case where bounds on the variables are present. Weak conditions on the accuracy of the Hessian approximations are considered. It is also shown that, when the strict complementarity condition holds, the proposed algorithms reduce to an unconstrained calculation after finitely many iterations, allowing a fast asymptotic rate of convergence.

Key words. trust regions, convergence theory, optimization with bounds

1. Introduction. Minimizing a nonlinear function of several variables subject to satisfying bounds on these variables is probably one of the most common types of constrained optimization problems encountered in practical applications. Some authors (see [9], for instance) even claim that a vast majority of optimization problems should be considered from the point of view that their variables are indeed restricted to certain meaningful intervals, and should therefore be solved in conjunction with bound constraints. Fortunately, it is the simplest of the inequality constrained problems, because of its structure. On the other hand, in a way it is more complex than many equality type problems: indeed it involves a combinatorial part, which is the detection of the set of constraints that are active at the solution. Algorithms that can take advantage of this structure and that are reasonably efficient in the determination of the optimal active set are thus of interest to many practitioners.

This fact has already been observed by many authors, and some special purpose methods have been proposed, as in [1], [2] and [11]. Of particular interest to us is the first of these proposals (on which the third is based), because it provides a rather complete convergence theory to back up a satisfying numerical performance. However, although this theory can easily be applied to convex problems, it is not clear in Bertsekas's presentation in [1] how to extend it to the nonconvex case, and still guarantee global convergence.

This question of ensuring global convergence on nonconvex problems has, on the other hand, been explored extensively in the recent past, in connection with the use of trust region techniques. One of the main reasons behind this development is the combination of a rather intuitive framework with a powerful theoretical foundation ensuring convergence to a stationary point, even from starting points that are far away from the problem's solution. We refer the reader to [12] for an excellent survey of this topic in the context of unconstrained optimization. More recently, many authors have considered extending the trust region concepts and algorithms to the constrained minimization case (see [3], [4], [7], [15], [16], for instance). In most of these papers, quite general equality constraints have been investigated, and a variety of solutions

* Received by the editors April 25, 1986; accepted for publication (in revised form) April 14, 1987.

† Department of Computer Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. The work of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant A8639.

‡ Harwell Laboratories, Harwell, England.

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

have been proposed. Another interesting reference is [8], where the linear inequality constrained case is considered.

It is the purpose of this paper to provide a general global convergence theory for an algorithm that solves nonconvex optimization problems with simple bounds, using a trust region technique. The analysis and algorithm presented merge ideas from [12] and [14] concerning unconstrained problems with those from [1], as far as the treatment of the bounds is concerned. Global convergence and adequate determination of the correct active set are proved. Preliminary numerical results that indicate the viability of the proposed method are reported elsewhere [5].

Section 2 presents the problem and algorithm in more detail, while § 3 is devoted to the convergence theory. Some conclusions are drawn in § 4.

2. An algorithm for bound constrained optimization. We consider solving the problem

$$\min f(x) \tag{1}$$

where $f(x)$ is a function of n real variables which are subject to the constraints

$$l_i \leqq x_i \leqq u_i \qquad (i = 1, \cdots, n). \tag{2}$$

We assume that we can compute the function value $f(x)$ and the gradient $\nabla f(x)$ for any feasible point x. We are also given a feasible starting point x_0, and we wish to start the minimization procedure from that point. If we define $\mathscr{L}$ as the intersection of the set

$$\{x \in \mathbf{R}^n \,|\, f(x) \leqq f(x_0)\}$$

with the feasible region defined by (2), we may formulate our assumptions on the problem as follows.

(AS.1) The set $\mathscr{L}$ is compact and has a nonempty interior.

(AS.2) The objective function $f(\cdot)$ is twice continuously differentiable in an open domain containing $\mathscr{L}$.

The first of these assumptions says that the optimization problem is nontrivial and cannot be reduced to a lower dimension. In particular, infinite (positive and/or negative) bounds are allowed provided the set $\mathscr{L}$ is still bounded.

The algorithm we propose for solving (1) subject to (2) is of trust region type. Indeed, at each iteration, we define a quadratic approximation to the objective function, and a region surrounding the current iterate where we believe this approximation to be adequate. The algorithm then finds, in this region, a candidate for the next iterate that sufficiently reduces the value of the quadratic model to the objective. If the function value calculated at this point matches its predicted value closely enough, the new point is then accepted as the next iterate and the trust region is possibly enlarged; otherwise the point is rejected and the trust region size decreased.

Before describing the details of the method, we need to introduce some notation. We will denote by $I(x)$ the set of all bound constraints that are violated or satisfied as equalities at the point x (we will call them *active*) and

$$C(x) \overset{\text{def}}{=} \text{span}\,\{e_i \,|\, i \notin I(x)\}, \tag{3}$$

where the vectors e_i are the vectors of the canonical basis. This last subspace is nothing but the linear subspace spanned by the variables that are not at their bounds or infeasible at x. We will also need the affine subspace

$$A(x) \overset{\text{def}}{=} \{y \,|\, y = x + z \text{ with } z \in C(x)\}, \tag{4}$$

that contains x and is parallel to $C(x)$. We will also use the "projection" operator defined componentwise by

$$(P[x])_i = \begin{cases} l_i & \text{if } x_i \leqq l_i, \\ u_i & \text{if } x_i \geqq u_i, \\ x_i & \text{otherwise.} \end{cases} \tag{5}$$

This operator "projects" the point x onto the feasible region defined by (2).

We are now in position to describe more precisely the strategy we propose in order to choose, at the kth iteration, a candidate for the $(k+1)$th iterate. Our model of the objective function is of the form

$$m_k(x_k + s_k) = f(x_k) + g_k^T s_k + \tfrac{1}{2} s_k^T B_k s_k, \tag{6}$$

where the superscript T stands for the transpose, where $g_k = \nabla f(x_k)$, and where the symmetric matrix B_k is an approximation to the Hessian $\nabla^2 f(x_k)$. (As will be seen below, this approximation may be quite poor.) We also consider the direction w_k defined by the condition

$$D_k^T D_k w_k = g_k \tag{7}$$

for some nonsingular *diagonal scaling matrix* D_k. (Without loss of generality, we assume that the entries of D_k are positive. Of course, $D_k^T = D_k$, but the inclusion of the transpose provides a closer relation with more general changes of variables.) This vector w_k then gives the scaled gradient direction. As in the unconstrained case, we will first ensure a sufficient decrease of our model along this direction, but, because of the bounds, we have to "project" onto the feasible region, yielding the polygonal line $P[x_k - tw_k]$ for $t > 0$. This line satisfies the important *norm increasing* property, that is,

$$\|P[x_k - t_1 w_k] - x_k\| \geqq \|P[x_k - t_2 w_k] - x_k\| \quad \text{whenever } t_1 \geqq t_2. \tag{8}$$

We may then define the continuous piecewise quadratic function

$$q_k(t) = m_k(P[x_k - tw_k]) \tag{9}$$

as a function of $t \geqq 0$, and denote by t_k^C the value of t in (9) corresponding to the first local minimum of $q_k(t)$ subject to the trust region constaint defined by

$$\|D_k(P[x_k - tw_k] - x_k)\| \leqq \nu \Delta_k, \tag{10}$$

where Δ_k is the trust region radius at the kth iteration, ν is a positive constant and $\|\cdot\|$ is the usual l_2 norm on $\mathbf{R}^n$. The point

$$x_k^C = P[x_k - t_k^C w_k] \tag{11}$$

is called the *generalized Cauchy point* at iteration k. This notion is illustrated in Fig. 1, where the circles are the contour lines of the objective function, and the trust region is large and inactive, so that its boundaries do not appear in the figure.

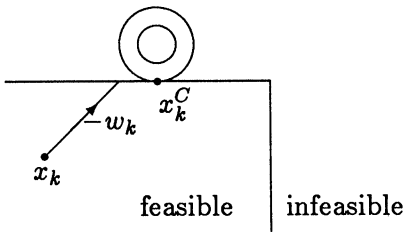

FIG. 1. *The generalized Cauchy point.*

We require that the trial point $x_k + s_k$ is feasible and that the step s_k satisfies the following two conditions:

$$f(x_k) - m_k(x_k + s_k) \geqq \beta_1[f(x_k) - q_k(t_k^C)] \tag{12}$$

and

$$\|D_k s_k\| \leqq \beta_2 \Delta_k. \tag{13}$$

In equations (12) and (13), the constants must satisfy the conditions

$$\beta_1 \in (0, 1] \quad \text{and} \quad \beta_2 \geqq \nu. \tag{14}$$

The first of these conditions requires the model reduction at $x_k + s_k$ to be within a fixed fraction of the model reduction at x_k^C, the second requires it to be inside an extended trust region.

This clearly generalizes the conditions used by Moré in [12] by suitably extending the notion of a Cauchy point to the case where bound constraints are present. Note that, because of the equivalence of norms and the presence of the constants ν and β_2 in (10) and (13), respectively, norms other than the l_2 norm can be chosen to define the trust region. In particular, the l_∞ norm may be of interest, because the shape of the trust region is then that of a box, and its boundaries are aligned with the bound constraints. In this case, β_2 can be set to $\nu\sqrt{n}$.

We also note that we do not assume that s_k be the quasi-Newton step

$$s_k = -B_k^{-1} g_k, \tag{15}$$

whenever B_k is positive definite and $\|D_k s_k\| < \beta_2 \Delta_k$, in contrast to [14]. This assumption may indeed be undesirable when n is large. In this context, the calculation of the direction (15) may be quite costly and is not always justified.

The reason for introducing the scaling matrices D_k is also practical. As discussed in [12], they ensure invariance with respect to diagonal transformations of the problem, which are the only ones that preserve the structure of the constraints. These matrices also allow the use of preconditioned conjugate gradients as a method for deriving a suitable step. Preconditioned conjugate gradients have already proved to be extremely useful in the context of large scale problems [11]. The preconditioner is then defined as the diagonal matrix

$$C_k = D_k^T D_k. \tag{16}$$

Note that, in practical implementations of this flexible method, a more sophisticated preconditioner can be used in the subspace $C(x_k^C)$, in order to improve on efficiency when computing the step s_k. We only require the diagonal preconditioner for the computation of x_k^C, because this is the part of the procedure where the combinatorial treatment of the bounds is performed, and the special structure of the constraints exploited.

Finally, we stress the fact that the computation of t_k^C can be implemented in a rather efficient way, once w_k is known (see [5]).

We are now able to outline our algorithm. It depends on some constants $\mu \in (0, 1)$, $\eta \in (\mu, 1)$, γ_0, γ_1 and γ_2 which must satisfy

$$0 < \gamma_0 \leqq \gamma_1 < 1 \leqq \gamma_2. \tag{17}$$

These constants are used in the trust region radius updating in a way similar to [12].

Step 0. The starting point x_0 and the function value $f(x_0)$ and gradient g_0 are given, as well as an initial trust region radius, Δ_0, and B_0, an initial approximation to the Hessian at the starting point. Set $k=0$.

Step 1. Obtain a step s_k as described above.

Step 2. Compute $f(x_k+s_k)$ and

$$\rho_k = \frac{f(x_k)-f(x_k+s_k)}{f(x_k)-m_k(x_k+s_k)}. \tag{18}$$

Step 3. In the case where

$$\rho_k > \mu, \tag{19}$$

set

$$x_{k+1} = x_k + s_k, \qquad g_{k+1} = \nabla f(x_{k+1}) \tag{20}$$

and

$$\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k] \quad \text{if } \rho_k \geqq \eta \tag{21}$$

or

$$\Delta_{k+1} \in [\gamma_1 \Delta_k, \Delta_k] \quad \text{if } \rho_k < \eta. \tag{22}$$

Otherwise, set

$$x_{k+1} = x_k, \qquad g_{k+1} = g_k \tag{23}$$

and

$$\Delta_{k+1} \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k]. \tag{24}$$

Step 4. Update the matrices B_k and D_k. Increment k by one and go to step 1.

This is obviously a theoretical algorithm. Many details should be added in order to specify a practical numerical procedure. In particular, we have omitted a stopping criterion, and all details on the method to determine the step s_k once the generalized Cauchy point x_k^C has been calculated.

We call an iteration *successful* if the test (19) is satisfied, that is when the achieved function reduction $f(x_k)-f(x_k+s_k)$ is large enough compared to the predicted function reduction $f(x_k)-m_k(x_k+s_k)$. If (19) is not satisfied, the iteration is said to be *unsuccessful.*

We finally observe that this algorithm only generates feasible points, which can be an advantage in the case of "hard" constraints, that is, when the function and/or gradient values are undefined or difficult to compute if some constraints are violated.

3. Convergence analysis. We now turn to the analysis of the behaviour of our algorithm, when applied to problem (1)-(2). It is quite clear that this behaviour will depend on the conditions we impose on the matrices B_k and D_k.

We first state our assumptions of the scaling matrices D_k.

(AS.3) The scaling matrices are diagonal and have bounded inverses, that is,

$$\|D_k^{-1}\| \leqq \sigma_1 \tag{25}$$

for some $\sigma_1 \geqq 1$.

Condition (25) is identical to that imposed in [12]. As in this reference, we observe that this condition does not imply that the scaling matrices have bounded condition numbers. The condition that $\sigma_1 \geqq 1$ can be imposed without loss of generality, and will be useful in the sequel.

Assumption (AS.3) also allows us to characterize critical points of our problem, that is points at which the first-order Kuhn–Tucker optimality conditions hold. This is expressed in the following statement.

LEMMA 1. *Let x be feasible and let D be a diagonal nonsingular matrix. Then x is a critical point for problem* (1)–(2) *if and only if*

$$P[x - tw] = x \tag{26}$$

for all $t \geqq 0$, where w is given by

$$D^T D w = \nabla f(x). \tag{27}$$

Proof. This lemma results immediately from Proposition 1.35 of [1]. □

Now observe that if we define

$$d_k(t) = P[x_k - tw_k] - x_k, \tag{28}$$

we can rewrite the conditions for x_k to be critical in terms of this new vector: x_k is a critical point if and only if $d_k(t) = 0$ for all $t \geqq 0$, which is equivalent to $d_k(t) = 0$ for any particular $t > 0$. The quantity

$$h_k \overset{\text{def}}{=} \| P[x_k - w_k] - x_k \| = \| d_k(1) \|, \tag{29}$$

can therefore be regarded as a measure of the "criticality" of the kth iterate, provided the scaling matrices D_k stay bounded in norm. The geometric interpretation of this quantity is illustrated in Fig. 2.

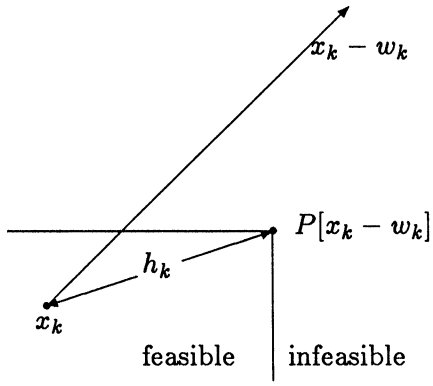

FIG. 2. *The critical length h_k.*

Since we are interested in asymptotic convergence, we will assume below that the sequence of iterates is infinite and $h_k > 0$ for all k.

We now state our condition on the model Hessians, namely,

(AS.4) If we define

$$b_k = 1 + \max_{i=0,\cdots,k} \| D_i^{-T} B_i D_i^{-1} \|, \tag{30}$$

we require that

$$\sum_{k=0}^{\infty} \frac{1}{b_k} = +\infty. \tag{31}$$

Note that we have added 1 to the norm of the scaled Hessian approximation in order to prevent definition problems when this last quantity is identically zero. This condition (31) is the weakest possible involving the whole of B_k for obtaining convergence to a stationary point when $D_k = I$. This was shown by Powell in [14] in the context of unconstrained minimization, where he provides an example showing that, if condition (31) is violated, the algorithm can converge to a noncritical point. It is also worth remembering that the well-known BFGS secant update does satisfy (AS.4) on convex problems (see [13]). This is also the case for a suitably safeguarded Symmetric Rank One update applied to convex and nonconvex problems (see [5]). On the other hand, if second derivatives of the objective function are available and used for B_i, then they are obviously bounded on $\mathscr{L}$, and (AS.4) holds too.

Our convergence analysis can be divided into three parts. In the first part, we examine the consequences of our step strategy and establish the important condition that, for a given successful iteration, iteration k say,

$$f(x_k) - f(x_k + s_k) \geq c_1 \|D_k^{-T} g_k\| \min\left[\Delta_k, \frac{\|D_k^{-T} g_k\|}{\|D_k^{-T} B_k D_k^{-1}\| + 1}\right], \tag{32}$$

for some constant $c_1 > 0$. This condition is crucial in both [12] and [14] for the unconstrained case, and its generalization to the bounded case stays crucial in our framework. In the second part of our theory, we establish global convergence of our algorithm to a critical point of the problem. We also show the important property that, assuming strict complementarity, the algorithm determines the set of bounds that are active at the solution in a finite number of iterations. This means that, asymptotically, the rate of convergence can be regarded as that of a purely unconstrained method. The third part of the analysis is concerned with guaranteeing convergence to a local minimum, not merely to a critical point, under conditions on the step s_k that take second order information into account.

3.1. Obtaining a sufficient decrease in the model. Let us first examine the implications of our step strategy at iteration k. Since the iteration is fixed, so is the scaling matrix D_k, and the complete procedure for determining a step s_k can be viewed as taking place in a scaled space. Indeed, if we denote the scaled quantities with a superscript s, we can define

$$x_k^s = D_k x_k, \quad s_k^s = D_k s_k, \quad d_k^s(t) = D_k\, d_k(t) \tag{33}$$

and

$$g_k^s = D_k^{-T} g_k, \qquad B_k^s = D_k^{-T} B_k D_k^{-1}. \tag{34}$$

Also observe that

$$w_k^s = D_k w_k = g_k^s, \tag{35}$$

so that the scaled w_k direction is nothing but the scaled gradient. In this new space, we can rewrite the piecewise quadratic (9) as

$$\begin{aligned} q_k(t) &= m_k(x_k + d_k(t)) \\ &= f(x_k) + g_k^T d_k(t) + \tfrac{1}{2}[d_k(t)]^T B_k d_k(t) \\ &= f^s(x_k^s) + [g_k^s]^T d_k^s(t) + \tfrac{1}{2}[d_k^s(t)]^T B_k^s d_k^s(t), \end{aligned} \tag{36}$$

where we redefine the objective function on the scaled space by

$$f^s(x^s) = f(D_k^{-1} x^s) = f(x). \tag{37}$$

Similarly, the bounds in (2) are transformed into new bounds on the scaled variables

$$l_i^s \leqq x_i^s \leqq u_i^s \qquad (i = 1, \cdots, n), \tag{38}$$

with

$$l_i^s = (D_k)_{ii} l_i \quad \text{and} \quad u_i^s = (D_k)_{ii} u_i. \tag{39}$$

These bounds, in turn, allow the definition of the set of indices of the active scaled constraints

$$I^s(x^s) = I(x), \tag{40}$$

and of a scaled $P^s[\cdot]$ operator as in (5), corresponding to projection onto the feasible domain defined by (38). Also similarly, a scaled $\mathscr{L}^s$ can be defined using (37) and (38). Then we have that

$$d_k^s(t) = D_k(P[x_k - tw_k] - x_k) = P^s[x_k^s - tg_k^s] - x_k^s \tag{41}$$

for all t, and the constraints (10) and (13) can be rewritten as

$$\|P^s[x_k^s - tg_k^s] - x_k^s\| \leqq \nu\Delta_k \tag{42}$$

and

$$\|s_k^s\| \leqq \beta_2 \Delta_k, \tag{43}$$

respectively. Finally, observe that, for all x,

$$\|x\| \leqq \|D_k^{-1}\| \cdot \|x^s\| \leqq \sigma_1 \|x^s\|. \tag{44}$$

To simplify the notation, we will use this one-to-one correspondence between the original and scaled spaces at iteration k, and therefore assume the scaling $D_k = I$. Hence the scaled and original spaces coincide for this iteration, and the superscript s can be dropped. We will reintroduce the notion of scaled space when we consider several iterations of our algorithm.

LEMMA 2. *For all* $0 < t_1 \leqq t_2$ *and all* k, *we have that*

$$I(x_k + d_k(t_1)) \subseteq I(x_k + d_k(t_2)). \tag{45}$$

Proof. The proof of this lemma is trivial once we observe that the set of active bounds can only be increased or remain the same as one follows the polygonal line defined by $x_k + d_k(t)$, and no satisfied bounds can become violated. □

We now define the reduced gradient with respect to a given set of active bounds as follows:

$$[z_k(t)]_i \overset{\text{def}}{=} \begin{cases} [w_k]_i & \text{if } i \notin I(x_k + d_k(t)), \\ 0 & \text{otherwise,} \end{cases} \tag{46}$$

where the subscript i denotes the ith component of the vector, and where $t \geqq 0$.

LEMMA 3. *Assume that* (AS.1)-(AS.2) *hold. Also assume that* $h_k > 0$. *Then, if*

$$c_2 \overset{\text{def}}{=} \max[1, \max_{z \in \mathscr{L}} \|\nabla f(x)\|], \tag{47}$$

we obtain that

$$\|z_k(t_k^{(1)})\| \geqq \tfrac{1}{2} h_k \tag{48}$$

where

$$t_k^{(1)} = \frac{h_k}{2c_2}. \tag{49}$$

Proof. To prove this lemma, we first note that c_2 is well defined because of the continuity of the gradient and the compactness of $\mathscr{L}$. We also deduce that

$$\|d_k(t_k^{(1)})\| \leqq \|t_k^{(1)} g_k\| = \frac{h_k}{2c_2}\|g_k\| \leqq \tfrac{1}{2}h_k. \tag{50}$$

Let us now denote by $t_k^{(2)}$ the smallest t such that

$$\|d_k(t)\| = h_k. \tag{51}$$

Then, using (50), we obtain

$$0 < t_k^{(1)} < t_k^{(2)} \leqq 1. \tag{52}$$

Hence, because of (50) and (51),

$$\begin{aligned} \|z_k(t_k^{(1)})\| &\geqq (t_k^{(2)} - t_k^{(1)})\|z_k(t_k^{(1)})\| \\ &\geqq \|d_k(t_k^{(2)}) - d_k(t_k^{(1)})\| \\ &\geqq \|d_k(t_k^{(2)})\| - \|d_k(t_k^{(1)})\| \\ &\geqq \tfrac{1}{2}h_k, \end{aligned} \tag{53}$$

which proves the lemma. □

It is worth noting that the line coordinate $t_k^{(1)}$ depends only on h_k and the problem.

With this tool at our disposal, we may now examine a crucial part of our development: the guaranteed decrease in the quadratic model starting from a noncritical point.

LEMMA 4. *Assume* (AS.1)-(AS.2) *hold. Also assume that, for some* $t_k^{(3)} > 0$,

$$\alpha_k \overset{\text{def}}{=} \|z_k(t_k^{(3)})\| > 0. \tag{54}$$

Then, if T *is the set of points in* $[0, t_k^{(3)}]$ *at which the piecewise quadratic* (9) *is differentiable,*

$$\frac{d}{dt} q_k(t) \leqq -\alpha_k^2 + 2t_k^{(3)} c_2^2 \|B_k\| \tag{55}$$

for all $t \in T$. *Furthermore,*

$$\frac{d}{dt} q_k(t) \leqq -\tfrac{1}{2}\alpha_k^2 \tag{56}$$

for all $t \in T \cap [0, t_k^{(4)}]$, *and*

$$q_k(t) \leqq f(x_k) - \tfrac{1}{2}\alpha_k^2 t, \tag{57}$$

for all $t \in [0, t_k^{(4)}]$, *where*

$$t_k^{(4)} \overset{\text{def}}{=} \min\left[t_k^{(3)}, \frac{\alpha_k^2}{4c_2^2(\|B_k\|+1)}\right]. \tag{58}$$

Proof. To prove this result, we first examine the behaviour of the slope of the function $q_k(t)$ in the interval $[0, t_k^{(3)}]$. As t increases from 0 to $t_k^{(3)}$, the polygonal line $x_k + d_k(t)$ may hit several bounds. Let us label

$$0 = t_0 < t_1 < \cdots < t_m = t_k^{(3)} \tag{59}$$

the "bends" or breakpoints points (if any) of this polygonal line. Now consider the set T as defined above, that is, the complete interval $[0, t_k^{(3)}]$ minus the breakpoints (59). Then for any $t \in T$,

$$\frac{d}{dt} q_k(t) = g_k^T d_k'(t) + d_k^T B_k d_k'(t), \tag{60}$$

where $d_k'(t)$ is defined componentwise by

$$[d_k'(t)]_i = \frac{d}{dt}[d_k(t)]_i \tag{61}$$

for $i = 1, \cdots, n$. We now assume, without loss of generality, that if bounds are active at x_k or become active as t increases from 0 to $t_k^{(3)}$, they do so in order of successive indices, that is, the bounds on the first variables become active first. It is then possible to write that

$$[d_k(t)]_i = \begin{cases} -t[g_k]_i & \text{for } t \in [0, \omega_i], \\ -\omega_i[g_k]_i & \text{for } t \in (\omega_i, t_k^{(3)}], \end{cases} \tag{62}$$

where ω_i is the t value at which the ith bound becomes active, if applicable. Equivalently,

$$[d_k(t)]_i = -[g_k]_i[t\delta(t \in [0, \omega_i]) + \omega_i \delta(t \in (\omega_i, t_k^{(3)}])], \tag{63}$$

where the function $\delta(\textit{condition})$ is equal to 1 if *condition* is true and zero otherwise. Then

$$[d_k'(t)]_i = -[g_k]_i \delta(t \in [0, \omega_i]). \tag{64}$$

We now examine the quadratic term in (60):

$$\begin{aligned}[d_k(t)]^T B_k d_k'(t) = \sum_{i,j=1}^{n} [B_k]_{ij}[g_k]_i[g_k]_j [t\delta(t \in [0, \omega_i]) \\ + \omega_i \delta(t \in (\omega_i, t_k^{(3)}])]\delta(t \in [0, \omega_j]).\end{aligned} \tag{65}$$

Now let us consider a given $t \in T$. Then there is an integer s such that $t \in (t_s, t_{s+1})$. Define r by

$$I(x_k + d_k(t_s)) = \{1, \cdots, r-1\}. \tag{66}$$

Therefore we obtain the following equivalences:

$$\begin{aligned} t \in [0, \omega_i] &\Leftrightarrow i \geqq r, \\ t \in (\omega_i, t_k^{(3)}] &\Leftrightarrow i \leqq r-1. \end{aligned} \tag{67}$$

Hence, for $t \in (t_s, t_{s+1})$,

$$[d_k(t)]^T B_k d_k'(t) = \sum_{i,j=1}^{n} [B_k]_{ij}[g_k]_i[g_k]_j [t\delta(i \geqq r)\delta(j \geqq r) + \omega_i \delta(i < r)\delta(j \geqq r)], \tag{68}$$

while

$$[g_k]^T d_k'(t) = -\sum_{i=1}^{n} [g_k]_i^2 \delta(i \geqq r) = -\sum_{i=r}^{n} [g_k]_i^2. \tag{69}$$

Gathering (68) and (69), we obtain from (60)

$$\begin{aligned}\frac{d}{dt} q_k(t) &= -\sum_{i=r}^{n} [g_k]_i^2 + t \sum_{i,j=r}^{n} [B_k]_{ij}[g_k]_i[g_k]_j + \sum_{i,j=1}^{n} [B_k^*]_{ij}[g_k]_i[g_k]_j \omega_i \\ &\leqq -\sum_{i=r}^{n} [g_k]_i^2 + t \sum_{i,j=r}^{n} [B_k]_{ij}[g_k]_i[g_k]_j + t_k^{(3)} \|g_k\|^2 \|B_k^*\|,\end{aligned} \tag{70}$$

where B_k^* is defined by

$$[B_k^*]_{ij} = \begin{cases} [B_k]_{ij} & \text{if } i < r \text{ and } j \geqq r, \\ 0 & \text{otherwise.} \end{cases} \tag{71}$$

Also define

$$[B_k^+]_{ij} = \begin{cases} [B_k]_{ij} & \text{if } i < r \quad (\text{all } j), \\ 0 & \text{if } i \geqq r \end{cases} \tag{72}$$

and now observe that

$$\|B_k^*\| \leqq \|B_k^+\| \leqq \|B_k\|. \tag{73}$$

If we put $v_k = z_k(t_s)$, then (70) implies that, for all $t \in (t_s, t_{s+1})$,

$$\frac{d}{dt} q_k(t) \leqq -\|v_k\|^2 + t[v_k]^T B_k v_k + t_k^{(3)} \|g_k\|^2 \|B_k\|. \tag{74}$$

Hence, for all $t \in T$,

$$\frac{d}{dt} q_k(t) \leqq -\alpha_k^2 + 2t_k^{(3)} c_2^2 \|B_k\|, \tag{75}$$

where we have used (47), the Cauchy-Schwarz inequality and the fact that Lemma 2 implies the inequality $\|v_k\| \geqq \alpha_k$. This proves (55). In particular, we may consider $t_k^{(4)}$ as defined by (58). Since $t_k^{(4)} \leqq t_k^{(3)}$, we obtain that $\|z_k(t_k^{(4)})\| \geqq \alpha_k > 0$, and we can apply the above argument with $t_k^{(3)}$ replaced by $t_k^{(4)}$, and yield

$$\frac{d}{dt} q_k(t) \leqq -\alpha_k^2 + 2t_k^{(4)} c_2^2 \|B_k\| \leqq -\tfrac{1}{2}\alpha_k^2 \tag{76}$$

(when the derivative exists), and, consequently,

$$q_k(t) \leqq f(x_k) - \tfrac{1}{2}\alpha_k^2 t \tag{77}$$

for all $t \in [0, t_k^{(4)}]$. □

Observe that this lemma does not take the trust region constraint (13) into account: only the model and the problem bounds are considered. Also observe that the term $(\|B_k\| + 1)$ can be replaced by $\|B_k\|$ when this is nonzero, but the given term simplifies the later analysis.

We can now state the equivalent of condition (32) in the case of bounded problems. In order to avoid confusion when applying this property, we formulate our result without assuming that $D_k = I$.

LEMMA 5. *Assume* (AS.1)-(AS.3) *hold. Also assume that*

$$h_k^s \overset{\text{def}}{=} \|D_k(P[x_k - w_k] - x_k)\| > 0. \tag{78}$$

Then

$$f(x_k) - m_k(x_k + s_k) \geqq \frac{1}{2} c_3 [h_k^s]^2 \min\left[\frac{[h_k^s]^2}{b_k}, \Delta_k\right], \tag{79}$$

where

$$c_3 \overset{\text{def}}{=} \min\left[\frac{\beta_2^2}{1-\eta}, \frac{\min(1, \nu)\beta_1}{64 c_2^2 \sigma_1^2}\right] \leqq 1. \tag{80}$$

If the kth iteration is successful, we also have that

$$f(x_k)-f(x_k+s_k)\geqq\frac{1}{2}\mu c_3[h_k^s]^2\min\left[\frac{[h_k^s]^2}{b_k},\Delta_k\right]. \tag{81}$$

Proof. We prove the result in the scaled space first, and then reformulate it in the original space. We observe that, if $t_k^{(1)}$ is used as $t_k^{(3)}$ in Lemma 4, then, by Lemma 3, $\alpha_k\geqq\frac{1}{2}h_k>0$, and relation (57) implies that

$$q_k(t)\leqq f(x_k)-\tfrac{1}{8}h_k^2 t, \tag{82}$$

where t is in an interval whose upper bound is

$$t_k^{(5)}\overset{\text{def}}{=}\min\left[\frac{h_k}{2c_2},\frac{h_k^2}{16c_2^2(\|B_k\|+1)}\right]=\frac{h_k^2}{16c_2^2(\|B_k\|+1)}, \tag{83}$$

because $h_k\leqq c_2$ by definition. Assume first that $\|d_k(t_k^{(5)})\|\leqq\nu\Delta_k$. Then, recalling (12), we obtain that

$$f(x_k)-m_k(x_k+s_k)\geqq\tfrac{1}{8}\beta_1 h_k^2 t_k^{(5)}. \tag{84}$$

On the other hand, if $\|d_k(t_k^{(5)})\|>\nu\Delta_k$, we know that $\|d_k(t_k^C)\|=\nu\Delta_k$, and, since

$$\left\|d_k\left(\frac{\nu\Delta_k}{c_2}\right)\right\|\leqq\frac{\nu\Delta_k}{c_2}\|g_k\|\leqq\nu\Delta_k, \tag{85}$$

we can deduce that

$$t_k^C\geqq\frac{\nu\Delta_k}{c_2}. \tag{86}$$

Therefore, because $t_k^C<t_k^{(5)}$, (12) and (82) imply that

$$f(x_k)-m_k(x_k+s_k)\geqq\frac{\nu\beta_1}{8c_2}h_k^2\Delta_k. \tag{87}$$

Gathering (83), (84) and (87), we obtain that

$$f(x_k)-m_k(x_k+s_k)\geqq\frac{\min(1,\nu)\beta_1}{128c_2^2}h_k^2\min\left[\frac{h_k^2}{\|B_k\|+1},\Delta_k\right]. \tag{88}$$

In this last inequality, as in the whole development in the scaled space, we have assumed that the scaled gradients are bounded above by the constant c_2. Returning to the unscaled space, we have to replace this bound by $\sigma_1 c_2$, which reintroduces the scaling and uses (25). We also replace $\|B_k\|+1$ by the scalar b_k, which already incorporates the scaling, and h_k by h_k^s. The relation (88) is thus nothing but the scaled form of (79), except that the constant c_3 is possibly further reduced, in order to simplify another development below. The inequality (81) then results from (79), (18) and (19). □

We now turn our attention to a useful variant of this result.

LEMMA 6. *Assume* (AS.1)-(AS.3) *hold. Also assume that*

$$a_k^s\overset{\text{def}}{=}\|D_k z_k(t_k^C)\|>0. \tag{89}$$

Then

$$f(x_k)-m_k(x_k+s_k)\geqq\frac{1}{2}c_3[a_k^s]^2\min\left[\frac{[a_k^s]^2}{b_k},\Delta_k\right], \tag{90}$$

where c_3 is defined in (80). *If the kth iteration is successful, we also have that*

$$f(x_k)-f(x_k+s_k)\geqq\frac{1}{2}\mu c_3[a_k^s]^2\min\left[\frac{[a_k^s]^2}{b_k},\Delta_k\right]. \tag{91}$$

Proof. The proof of this lemma is similar to that of Lemma 5. In the scaled space, (89) is equivalent to

$$a_k\overset{\text{def}}{=}\|z_k(t_k^C)\|>0. \tag{92}$$

Moreover, the inequality $\|z_k(0)\|\geqq\|z_k(t_k^C)\|>0$ implies that $t_k^C>0$. Hence we can apply Lemma 4 with $t_k^{(3)}$ replaced by t_k^C. We then deduce that

$$q_k(t)\leqq f(x_k)-\tfrac{1}{2}a_k^2 t \tag{93}$$

for all nonnegative $t\in[0,\min(t_k^C,t_k^{(6)})]$, where

$$t_k^{(6)}\overset{\text{def}}{=}\frac{a_k^2}{4c_2^2(\|B_k\|+1)}. \tag{94}$$

Consider first the case where $t_k^C<t_k^{(6)}$. If $\|d_k(t_k^C)\|<\nu\Delta_k$, then there exists a $t_k^{(7)}\geqq t_k^C$ such that the derivative of $q_k(t)$ is defined and nonnegative at this point, and such that

$$I(x_k+d_k(t_k^{(7)}))=I(x_k^C). \tag{95}$$

This last condition implies that $z_k(t_k^{(7)})=z_k(t_k^C)$, and we obtain that

$$0\leqq-a_k^2+2t_k^{(7)}c_2^2\|B_k\| \tag{96}$$

because of Lemma 4. This ensures that

$$t_k^{(7)}\geqq\frac{a_k^2}{2c_2^2(\|B_k\|+1)}>t_k^{(6)}. \tag{97}$$

But $t_k^{(7)}$ is arbitrarily close to t_k^C, and hence $t_k^C\geqq t_k^{(6)}$, which is impossible. Therefore, we deduce that $\|d_k(t_k^C)\|=\nu\Delta_k$. We can then apply an argument identical to that used in the previous lemma, and we obtain that

$$f(x_k)-m_k(x_k+s_k)\geqq\frac{\nu\beta_1}{2c_2}a_k^2\Delta_k. \tag{98}$$

If we now consider the case where $t_k^C\geqq t_k^{(6)}$, we can first deduce that $\|d_k(t_k^{(6)})\|\leqq\nu\Delta_k$ because of the definition of t_k^C and the norm increasing property (8). Therefore, the inequality (93), (94) and (12) give

$$f(x_k)-m_k(z_k+s_k)\geqq\frac{1}{2}\beta_1 a_k^2\left(\frac{a_k^2}{4c_2^2(\|B_k\|+1)}\right). \tag{99}$$

Gathering now (98) and (99), we obtain

$$f(x_k)-m_k(x_k+s_k)\geqq\frac{1}{2}c_3 a_k^2\min\left[\frac{a_k^2}{\|B_k\|+1},\Delta k\right]. \tag{100}$$

This implies (90) in the unscaled space, as explained above. The inequality (91) then again results from (90), (18) and (19). □

This closes the first part of our convergence theory which was devoted to finding a suitable generalization of condition (32) to the case where bound constraints are present.

3.2. Global convergence to critical points. We now wish to use the guarantee of a sufficient decrease in the model to prove global convergence to critical points for the algorithm. This will be accomplished very much in the spirit of [14]. The global convergence itself indeed will be separated into two propositions, with statements close to those appearing in Powell's paper. It is interesting to note that, although the proof of the second statement follows Powell's argument closely, the proof of the first of these statements is quite different from that in [14].

LEMMA 7. *Assume that* (AS.1)-(AS.3) *hold. Consider a sequence* $\{x_k\}$ *of points generated by the algorithm, and assume that there is a constant* $\varepsilon > 0$ *such that, for all* k,

$$h_k^s \geqq \varepsilon, \tag{101}$$

where h_k^s *is defined by* (78). *Then there exist a constant* $c_4 > 0$ *such that, for all* $k \geqq 1$,

$$\Delta_k \geqq \frac{c_4}{b_k}, \tag{102}$$

where b_k *is defined by* (30).

Proof. We first define

$$c_5 \overset{\text{def}}{=} \max_{x \in \mathscr{L}} \|\nabla^2 f(x)\|, \tag{103}$$

and assume, without loss of generality, that $c_5 \sigma_1^2 \geqq 1$, and that

$$\varepsilon < \min\left(1, \beta_2 \sqrt{\frac{(c_5\sigma_1^2 + b_0)\Delta_0}{\gamma_0 c_3 (1-\eta)}}\right). \tag{104}$$

We also have that

$$|f(x_k + s_k) - m_k(x_k + s_k)| = \tfrac{1}{2}|s_k^T (G_k - B_k) s_k| \tag{105}$$

because of (6), where

$$G_k = 2\int_0^1 (1-t)\nabla^2 f(x_k + t s_k)\, dt. \tag{106}$$

This yields

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leqq \tfrac{1}{2}\beta_2^2 \Delta_k^2 \|D_k^{-T}(G_k - B_k) D_k^{-1}\| \leqq \tfrac{1}{2}\beta_2^2 \Delta_k^2 [c_5\sigma_1^2 + b_k], \tag{107}$$

where we have used (AS.3), (13), (30), (33) and the inequality $\|G_k\| \leqq c_5$. For further reference, we note that this inequality holds independently of (101). Assume now that there is a k such that

$$(c_5\sigma_1^2 + b_k)\Delta_k \leqq \gamma_0 (1-\eta) c_3 \frac{\varepsilon^2}{\beta_2^2}, \tag{108}$$

and define r as the first iteration number such that (108) holds. (Note that (104) implies that (108) does not hold for $k = 0$, and hence $r \geqq 1$.) Then the mechanism of the algorithm ensures that

$$b_{r-1}\Delta_{r-1} \leqq (c_5\sigma_1^2 + b_{r-1})\Delta_{r-1} \leqq (c_5\sigma_1^2 + b_r)\frac{\Delta_r}{\gamma_0} \leqq (1-\eta) c_3 \frac{\varepsilon^2}{\beta_2^2} \leqq \varepsilon^2, \tag{109}$$

where we used (80) to derive the last inequality. This last inequality, (30), (101), (104) and Lemma 5 then imply that

$$f(x_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1}) \geqq \frac{1}{2} c_3 \varepsilon^2 \min\left[\frac{\varepsilon^2}{b_{r-1}}, \Delta_{r-1}\right] \geqq \frac{1}{2} c_3 \varepsilon^2 \Delta_{r-1}. \tag{110}$$

Combining (107) and (110), we obtain that

$$|\rho_{r-1}-1| = \frac{|f(x_{r-1}+s_{r-1}) - m_{r-1}(x_{r-1}+s_{r-1})|}{|f(x_{r-1}) - m_{r-1}(x_{r-1}+s_{r-1})|} \leqq \frac{\beta_2^2(c_5\sigma_1^2 + b_{r-1})\Delta_{r-1}}{c_3\varepsilon^2} \leqq 1-\eta, \tag{111}$$

where we also used (18) and the middle part of (109). This imposes $\rho_{r-1} \geqq \eta$, and therefore (21) implies that $\Delta_r \geqq \Delta_{r-1}$. This, in turn, gives

$$(c_5\sigma_1^2 + b_{r-1})\Delta_{r-1} \leqq (c_5\sigma_1^2 + b_r)\Delta_r \leqq \gamma_0(1-\eta)c_3\frac{\varepsilon^2}{\beta_2^2}, \tag{112}$$

which contradicts the assumption that r was the first index with (108) satisfied. Hence, (108) never holds, and we obtain that

$$(c_5\sigma_1^2 + b_k)\Delta_k > \gamma_0(1-\eta)c_3\frac{\varepsilon^2}{\beta_2^2} \tag{113}$$

for all $k \geqq 1$. But, since

$$c_5\sigma_1^2 + b_k \leqq c_5\sigma_1^2(b_k+1) \leqq 2c_5\sigma_1^2 b_k \tag{114}$$

for all such k, we have proved (102) with

$$c_4 = \frac{\gamma_0(1-\eta)c_3\varepsilon^2}{2c_5\sigma_1^2\beta_2^2}. \tag{115}$$

We are now ready for our main theorem in this section.

THEOREM 8. *Assume that* (AS.1)-(AS.4) *hold. Also assume that* $\{x_k\}$ *is a sequence of iterates generated by the algorithm. Then*

$$\liminf_{k\to\infty} h_k = 0, \tag{116}$$

where h_k *is defined in* (29).

Proof. The proof of this statement is by contradiction. Assume therefore that there is an $\varepsilon > 0$ such that

$$h_k \geqq \varepsilon\sigma_1 \tag{117}$$

for all k, implying that

$$h_k^s \geqq \varepsilon \tag{118}$$

because of (44). Now Lemma 5 and the fact that the objective function is bounded below on the set $\mathscr{L}$ imply that

$$\frac{1}{2}\mu c_3\varepsilon^2 \sum_{k\in S} \min\left[\frac{\varepsilon^2}{b_k}, \Delta_k\right] \leqq \sum_{k\in S}[f(x_k) - f(x_{k+1})] < +\infty, \tag{119}$$

where S is the set of successful iterations. Using this inequality and Lemma 7, we may deduce that the sum

$$\min(\varepsilon^2, c_4)\sum_{k\in S}\frac{1}{b_k} \leqq \sum_{k\in S}\min\left[\frac{\varepsilon^2}{b_k}, \frac{c_4}{b_k}\right] \leqq \sum_{k\in S}\min\left[\frac{\varepsilon^2}{b_k}, \Delta_k\right] \tag{120}$$

converges to a finite limit. Now let p be defined as an integer such that

$$\gamma_2\gamma_1^{p-1} < 1 \tag{121}$$

and also define

$$S(k) = |S \cap \{0, \cdots, k\}| \tag{122}$$

the number of successful iterations up to iteration k. Then define

$$J_1 = \{k \mid k \leqq pS(k)\} \quad \text{and} \quad J_2 = \{k \mid k > pS(k)\}. \tag{123}$$

We now want to show that both sums

$$\sum_{k \in J_1} \frac{1}{b_k} \quad \text{and} \quad \sum_{k \in J_2} \frac{1}{b_k} \tag{124}$$

are finite. Consider the first. If it has only finitely many terms, its convergence is obvious. Otherwise, we may assume that J_1 has an infinite number of elements, and we then construct two subsequences. The first one consists of the indices of J_1 in ascending order and the second one, J_3 say, of the set of indices in S (in ascending order) with each index repeated p times. Hence the jth element of J_3 is no greater than the jth element of J_1. This gives

$$\sum_{k \in J_1} \frac{1}{b_k} \leqq \sum_{k \in J_3} \frac{1}{b_k} = p \sum_{k \in S} \frac{1}{b_k} \leqq +\infty \tag{125}$$

because of the nondecreasing nature of the sequence $\{b_k\}$ and the convergence of the last sum. We now turn our attention to the second sum in (124). Observe that, for $k \in J_2$,

$$\frac{c_4}{b_k} \leqq \Delta_k \leqq \gamma_2^{S(k)} \gamma_1^{k-S(k)} \Delta_0 \leqq \gamma_2^{k/p} \gamma_1^{k-k/p} \Delta_0 \leqq (\gamma_2 \gamma_1^{p-1})^{k/p} \Delta_0, \tag{126}$$

where we have used Lemma 7 and the definition of J_2 in (123). This yields

$$\sum_{k \in J_2} \frac{1}{b_k} \leqq \frac{\Delta_0}{c_4} \sum_{k \in J_2} (\gamma_2 \gamma_1^{p-1})^{k/p} < +\infty \tag{127}$$

and the second sum is convergent. Therefore the sum

$$\sum_{k=0}^{\infty} \frac{1}{b_k} = \sum_{k \in J_1} \frac{1}{b_k} + \sum_{k \in J_2} \frac{1}{b_k} \tag{128}$$

is finite, which contradicts (AS.4). Hence condition (117) is impossible and (116) is true. □

We also obtain the following important corollary directly from this proof.

COROLLARY 9. *Assume that* (AS.1)-(AS.4) *hold, and that* $\{x_k\}$ *is a sequence of iterates generated by the algorithm. Then*

$$\liminf_{k \to \infty} h_k^s = 0, \tag{129}$$

where h_k^s *is defined in* (78).

We note that (129) gives a scaled equivalent of (116). We often prefer (129) as a convergence result, because we believe that, in most cases, the scaled quantities are more meaningful when they are used to assert convergence (see also [6] on this subject). Also observe that one has to be cautious in interpreting (129) or (116) as "convergence." In fact, these equations mean that at least a subsequence of the iterates approaches a critical point provided the norms of the scaling matrices D_k stay bounded, as has already been pointed out.

We also note here that Lemma 6 could be used instead of Lemma 5 in the proofs of Lemma 7 and Theorem 8, in order to prove that

$$\liminf_{k\to\infty} a_k = \liminf_{k\to\infty} a_k^s = 0, \tag{130}$$

where a_k and a_k^s are defined in (92) and (89), respectively.

We are now able to prove that the "lim inf" in (116) can be replaced by a true limit, if we strengthen our assumptions somewhat, as is shown in the next theorem.

We first complete our assumptions on the scaling matrices, and ensure that h_k is a suitable measure of criticality.

(AS.5) There is a positive constant $\sigma_2 \geqq 1$ such that, for all k,

$$\|D_k\| \leqq \sigma_2. \tag{131}$$

This and (AS.3) clearly implies that the scaling matrices have bounded condition numbers, and we obtain the following result.

LEMMA 10. *Assume that* (AS.3) *and* (AS.5) *hold. Then*

$$\frac{1}{\sigma_1^2} h_k \leqq \|P[x_k - g_k] - x_k\| \leqq \sigma_2^2 h_k \tag{132}$$

for all k.

Proof. This lemma is not difficult to prove. We first observe that (AS.3), (AS.5) and (7) imply that

$$\frac{1}{\sigma_2^2} |[g_k]_j| \leqq |[w_k]_j| \leqq \sigma_1^2 |[g_k]_j| \tag{133}$$

for all $j = 1, \cdots, n$, and that the components of w_k and g_k have the same sign. But, if we set

$$v_j = \begin{cases} u_j & \text{if } [g_k]_j < 0, \\ l_j & \text{if } [g_k]_j > 0, \end{cases} \tag{134}$$

we also have that, for all j,

$$\begin{aligned} |[P[x_k - w_k] - x_k]_j| &= \min\left(|[w_k]_j|, |[x_k]_j - v_j|\right) \\ &\leqq \min\left(\sigma_1^2 |[g_k]_j|, |[x_k]_j - v_j|\right) \\ &\leqq \sigma_1^2 |[P[x_k - g_k] - x_k]_j|. \end{aligned} \tag{135}$$

Similarly, for all j,

$$\begin{aligned} |[P[x_k - w_k] - x_k]_j| &\geqq \min\left(\frac{1}{\sigma_2^2} |[g_k]_j|, |[x_k]_j - v_j|\right) \\ &\geqq \frac{1}{\sigma_2^2} |[P[x_k - g_k] - x_k]_j|. \end{aligned} \tag{136}$$

Formula (132) follows immediately. □

We also require the following condition:

$$\lim_{k\to\infty} b_k [f(x_k) - f(x_{k+1})] = 0. \tag{137}$$

(AS.6)

It says that the norm of the approximating Hessians should not increase too fast compared with the speed of convergence of the function values. This condition clearly holds if the sequence $\{b_k\}$ is bounded, as assumed in [12]. As noted above, this is

obviously the case when exact Hessians are used to form the quadratic model, yielding a variant of Newton's method. Observe also that the fact that $f(\cdot)$ is bounded below and (AS.4) already imply that

$$\liminf_{k\to\infty} b_k[f(x_k)-f(x_{k+1})]=0. \tag{138}$$

Indeed, assume that

$$b_k[f(x_k)-f(x_{k+1})]\geqq\varepsilon \tag{139}$$

for some $\varepsilon>0$. Then

$$\sum_{k=0}^{\infty}\frac{1}{b_k}\leqq\frac{1}{\varepsilon}\sum_{k=0}^{\infty}[f(x_k)-f(x_{k+1})]<+\infty \tag{140}$$

if $f(\cdot)$ is bounded below, which is impossible because of (AS.4).

THEOREM 11. *Assume* (AS.1)-(AS.6) *hold. Also assume that* $\{x_k\}$ *is a sequence of iterates generated by the algorithm. Then,*

$$\lim_{k\to\infty} h_k=0. \tag{141}$$

Proof. We prove this theorem by contradiction. Assume therefore that there is an $\varepsilon_1\in(0,1)$ and a subsequence $\{m_i\}$ of successful iterates such that, for all m_i in this subsequence

$$h_{m_i}\geqq\sigma_1\varepsilon_1, \tag{142}$$

and thus, by (44),

$$h^s_{m_i}\geqq\varepsilon_1. \tag{143}$$

Corollary 9 guarantees the existence of another subsequence $\{l_i\}$ such that

$$h^s_k\geqq\varepsilon_2 \quad \text{for } m_i\leqq k<l_i \quad \text{and} \quad h^s_{l_i}<\varepsilon_2, \tag{144}$$

where we have set

$$\varepsilon_2=\frac{\varepsilon_1}{4\sigma_1^2\sigma_2^2}<\varepsilon_1. \tag{145}$$

Note that the last inequality in (144), (44) and Lemma 10 imply that

$$\|P[x_{l_i}-g_{l_i}]-x_{l_i}\|\leqq\sigma_2^2 h_{l_i}\leqq\sigma_1\sigma_2^2\varepsilon_2. \tag{146}$$

We may now restrict our attention to the subsequence of successful iterations whose index is in the set

$$K=\{k\mid k\in S \text{ and } m_i\leqq k<l_i\}, \tag{147}$$

where m_i and l_i belong, respectively, to the two subsequences defined above. Now applying Lemma 5, for $k\in K$, we obtain that

$$f(x_k)-f(x_{k+1})\geqq\frac{1}{2}\mu c_3\varepsilon_2^2\min\left[\frac{\varepsilon_2^2}{b_k},\Delta_k\right]. \tag{148}$$

Because of (137), we then have that

$$\lim_{k\in K} b_k\Delta_k=0. \tag{149}$$

Therefore, we can deduce that, for i sufficiently large,

$$\begin{aligned} \|x_{m_i} - x_{l_i}\| &\leqq \sum_{k=m_i}^{l_i-1} \|x_{k+1} - x_k\| \\ &\leqq \beta_2 \sigma_1 \sum_{k=m_i}^{l_i-1}{}^{(K)} \Delta_k \\ &\leqq c_6 \sum_{k=m_i}^{l_i-1}{}^{(K)} [f(x_k) - f(x_{k+1})] \\ &\leqq c_6 [f(x_{m_i}) - f(x_{l_i})], \end{aligned} \tag{150}$$

where the sums with superscript (K) are restricted to the indices in K, and where we have set

$$c_6 \overset{\text{def}}{=} \frac{2\beta_2 \sigma_1}{\mu c_3 \varepsilon_2^2}. \tag{151}$$

But the last right-hand side of relation (150) tends to zero as i tends to infinity, and thus continuity and (146) imply that

$$\|P[x_{m_i} - g_{m_i}] - x_{m_i}\| \leqq 2\sigma_1 \sigma_2^2 \varepsilon_2 \tag{152}$$

for i sufficiently large. We may now apply Lemma 10 again, and we obtain that

$$h_{m_i} \leqq 2\sigma_1^3 \sigma_2^2 \varepsilon_2 \leqq \tfrac{1}{2} \sigma_1 \varepsilon_1 \tag{153}$$

when i is sufficiently large. This contradicts (142) and proves the theorem. □

As above, we note that (AS.3) and (AS.5) imply a scaled equivalent of (141), i.e.,

$$\lim_{k \to \infty} h_k^s = 0. \tag{154}$$

We also note that the technique of the proof of Theorem 11 cannot be used to replace the limit inferior by the true limit in (130) because of the lack of continuity of a_k and a_k^s as functions of the point x_k.

Before dealing with the extension of convergence results involving second-order information to the bounded case, we wish to analyse the behaviour of the algorithm when the sequence converges to a critical point. In particular, we will show that, if the iterates converge to a critical point satisfying the strict complementarity conditions, and if the condition

$$I(x_k^C) \subseteq I(x_k + s_k) \tag{155}$$

holds, then the set of constraints that are active at this critical point is correctly identified in a finite number of iterations. Condition (155) ensures that all bounds that are active at the generalized Cauchy point x_k^C are also active at $x_k + s_k$. This is imposed in [5], and we will assume it from now on.

In the argument that follows, we consider $\{x_k\}$ an arbitrary sequence generated by the algorithm. We also define L to be the set of all limit points of this sequence. This set is nonempty because $\mathscr{L}$ is compact. We then restrict our attention to the case where the following assumption holds.

(AS.7) All limit points in L satisfy the strict complementarity slackness condition

$$i \in I(x_*) \Rightarrow |[\nabla f(x_*)]_i| > 0 \tag{156}$$

for every $x_* \in L$, where i is the index of a bound.

This condition has the following consequence.

LEMMA 12. *Assume* (AS.1)-(AS.7) *hold. Then each limit point* $x_* \in L$ *belongs to an unique connected set of limit points of* L, X_* *say, and there exists a set* $I(X_*)$ *such that*

$$I(x_*) = I(X_*) \tag{157}$$

for all $x_* \in X_*$.

Proof. The fact that x_* belongs to an unique connected set of limit point X_* is obvious. We still have to show the existence of $I(X_*)$ and (157). First, Theorem 11 ensures that all points in L are critical. This implies that, for all $x_* \in X_*$,

$$j \notin I(x_*) \Rightarrow [\nabla f(x_*)]_j = 0. \tag{158}$$

On the other hand, since the set $\{x_* \in L \mid j \in I(x_*)\}$ must be closed and hence compact, (AS.7) then implies that there exists an $\varepsilon > 0$ such that

$$j \in I(x_*) \Rightarrow |[\nabla f(x_*)]_j| \geqq \varepsilon \tag{159}$$

for all $x_* \in X_*$. We can then deduce from (158) and (159), connectivity of X_* and the continuity of the gradient that, if $j \in I(\bar{x}_*)$ for some $\bar{x}_* \in X_*$, then j must belong to $I(x_*)$ for all $x_* \in X_*$. Hence we can choose $I(X_*) = I(x_*)$ for any such x_*, and the lemma is proved. □

We now show that, once a constraint is picked up as the iterates approach a limit point, it will be active for the rest of the calculation.

LEMMA 13. *Assume that* (AS.1)-(AS.7) *and* (155) *hold. Then*

$$I(x_k) \subseteq I(x_{k+m}) \tag{160}$$

for all $m \geqq 0$ *and* k *sufficiently large.*

Proof. We first observe that, because $\mathscr{L}$ is compact, it is possible to choose k_1 sufficiently large and $\delta > 0$ such that, given an iterate x_k with $k \geqq k_1$ there exists a connected set of limit points $X_*^{(k)} \subseteq L$ such that

$$x_k \in \mathcal{N}(X_*^{(k)}, \delta) \overset{\text{def}}{=} \{x \in \mathscr{L} \mid d(x, X_*^{(k)}) \leqq \delta\}, \tag{161}$$

where the distance $d(x, X)$ is defined by

$$d(x, X) \overset{\text{def}}{=} \min_{y \in X} \|x - y\| \tag{162}$$

for any vector x and any compact set X. Using continuity, Lemma 12 and (AS.7), we can also assume, without loss of generality, that δ is small enough to ensure that, for all $x \in \mathcal{N}(X_*^{(k)}, \delta)$

$$I(x) \subseteq I(X_*^{(k)}), \tag{163}$$

$$\operatorname{sgn}([\nabla f(x)]_j) = \operatorname{sgn}([\nabla f(x_*)]_j) \quad \text{and} \quad |[\nabla f(x)]_j| \geqq \tfrac{1}{2}|[\nabla f(x_*)]_j| \tag{164}$$

for all $j \in I(x)$ and all $x_* \in X_*^{(k)}$. Consider now a $k \geqq k_1$. If the kth iteration is unsuccessful, then $x_{k+1} = x_k$ and, clearly, $I(x_k) \subseteq I(x_{k+1})$. If it is successful, condition (155) ensures that $I(x_k^C) \subseteq I(x_{k+1})$. But (163) and (164) then imply that $I(x_k) \subseteq I(x_k^C)$. Therefore, we obtain that $I(x_k) \subseteq I(x_{k+1})$ for all k sufficiently large, and (160) is proved. □

We now show that the correct active set is identified by the algorithm after a finite number of iterations.

THEOREM 14. *Assume* (AS.1)-(AS.7) *and* (155) *hold. Then*

$$I(x_k) = I(x_*) \tag{165}$$

for k sufficiently large, where x_* is arbitrary in L.

Proof. First choose an arbitrary subsequence $\{x_{k_i}\}$ converging to a limit point $x_* \in L$, say, and denote by X_* the connected set of limit point to which x_* belongs, according to Lemma 12. Since X_* is connected, it is possible to find a $\psi > 0$ such that the distance from X_* to any other limit point not in X_* is bounded below by ψ. As in the previous lemma, we can choose $\delta \in (0, \frac{1}{4}\psi]$ sufficiently small and a k_1 sufficiently large so as to guarantee that (163) and (164) hold and that

$$d(x_k, X_*) > \delta \Rightarrow d(x_k, X_*) \geqq \tfrac{1}{2}\psi \tag{166}$$

for all $k \geqq k_1$. We now define the subsequence $\{k_j\}$ as

$$\{k_j\} = \{k \geqq k_1 \mid x_k \in \mathcal{N}(X_*, \delta) \text{ and } k \in S\}, \tag{167}$$

where, as in the proof of Theorem 8, S denotes the set of indices of successful iterations. By definition of X_*, the subsequence (167) has infinitely many terms. Assume, finally, for the purpose of obtaining a contradiction, that, for all j,

$$I(x_{k_j}^C) \subset I(X_*), \tag{168}$$

where the inclusion is strict. Then there must be a nonempty set $T \subseteq I(X_*)$ such that $T \cap I(x_{k_j}^C)$ is empty for all j. Therefore, from (AS.5) and (164), we may deduce that, for all j,

$$[a_{k_j}^s]^2 \geqq \frac{1}{\sigma_2^2} \sum_{t \in T} [g_{k_j}]_t^2 \geqq \frac{1}{4\sigma_2^2} |J| \min_{t \in T} |[\nabla f(x_*)]_t^2| \geqq \varepsilon^2 \tag{169}$$

for some $\varepsilon > 0$. Lemma 6 then implies that

$$b_{k_j}[f(x_{k_j}) - f(x_{k_j+1})] \geqq \tfrac{1}{2}\mu c_3 \varepsilon^2 \min[\varepsilon^2, b_{k_j}\Delta_{k_j}], \tag{170}$$

and (AS.6) thus gives that

$$\lim_{j \to \infty} b_{k_j}\Delta_{k_j} = 0. \tag{171}$$

The inequality $b_{k_j} \geqq 1$, (AS.3) and (171) show that

$$\|s_{k_j}\| \leqq \sigma_1\beta_2\Delta_{k_j} \leqq \tfrac{1}{2}\delta \tag{172}$$

for j larger than $j_1 \geqq 1$, say. Then, for all $j \geqq j_1$,

$$d(x_{k_j+1}, X_*) \leqq d(x_{k_j}, X_*) + \|s_{k_j}\| \leqq \frac{3\delta}{2} \leqq \frac{3}{8}\psi, \tag{173}$$

because of the definition (176). Hence, because of (166), the iterates cannot jump outside $\mathcal{N}(X_*, \delta)$, and we must have that $x_{k_j+1} \in \mathcal{N}(X_*, \delta)$ again. This implies that $x_{k_{j+1}} \in \mathcal{N}(X_*, \delta)$. Therefore, the subsequence $\{k_j\}$ is identical to the complete sequence of successful iterates with $k \geqq k_{j_1}$. Hence we may deduce from (171) that

$$\lim_{\substack{k \to \infty \\ k \in S}} b_k \Delta_k = 0. \tag{174}$$

But the mechanism of the algorithm then implies that

$$\lim_{k\to\infty} b_k \Delta_k = 0. \tag{175}$$

Using (169), (175) and Lemma 6 again, we obtain that, for all k sufficiently large,

$$f(x_k) - m_k(x_k + s_k) \geqq \tfrac{1}{2} c_3 \varepsilon^2 \Delta_k. \tag{176}$$

Since $b_k \geqq 1$, the inequalities (107) and (176) now imply that

$$|\rho_k - 1| \leqq \frac{\beta_2^2(c_5\sigma_1^2+1)}{c_3\varepsilon^2} b_k \Delta_k \tag{177}$$

for k sufficiently large. The limit (175) then yields that $\rho_k \geqq \eta$ for k large enough. But this prevents the sequence $\{\Delta_k\}$ from converging to zero, which contradicts (175). Hence our assumption (168) is false, and we obtain that there exists a subsequence $\{k_t\} \subseteq \{k_j\}$ such that $I(X_*) \subseteq I(x_{k_t}^C)$ for all t. Lemma 13, (155) and (163) then give (165). □

This last result is important because it shows that, under very mild conditions, the asymptotic behaviour of the algorithm is that of a purely unconstrained method, restricted to the subspace of variables that are not at their bounds at the solution. Hence rate of convergence analysis for the unconstrained case can be applied in our context without any modification.

It is also interesting to observe that Lemma 13 and Theorem 14 together imply that all limit points of the sequence of iterates generated by the algorithm have the same active set.

3.3. Convergence to local minimizers. In this section, we consider exploiting second-order information in the model and the objective functions to ensure stronger convergence results. Our analysis follows the broad lines of the developments in [12], recasting the results presented therein for unconstrained optimization into the context of bounded minimization. We first examine some conditions that guarantee that the complete sequence of iterates generated by the algorithm converges.

First define $\lambda^1[X]$ as the minimum eigenvalue of the symmetric matrix X restricted to the subspace $C(x_*)$. Then we can state the following result.

THEOREM 15. *Assume that* (AS.1)-(AS.7) *and* (155) *hold, and assume that* $\{x_{k_i}\}$ *is a subsequence of iterates, generated by the algorithm, converging to the critical point* x_*. *Also assume that there is an* $\varepsilon > 0$ *such that*

$$\liminf_{i\to\infty} \lambda^1[B_{k_i}] \geqq \varepsilon. \tag{178}$$

Assume finally that $\nabla^2 f(x_*)$ *is nonsingular on the subspace* $C(x_*)$. *Then the complete sequence of iterates* $\{x_k\}$ *converges to* x_* *and all iterates lie in* $A(x_*)$ *after finitely many iterations.*

Proof. The criticality of x_* is ensured by Theorem 11. We first choose a $\delta > 0$ smaller than the distance from x_* to the nearest bound not active at x_*. Then we choose an i_1 sufficiently large to ensure that $x_{k_i} \in \mathcal{N}(x_*, \delta)$ for all $i \geqq i_1$. If we denote by Z_* a matrix whose columns form an orthonormal basis of the subspace $C(x_*)$ and by ψ the minimum singular value of $Z_*^T \nabla^2 f(x_*) Z_*$ (that is $\nabla^2 f(x_*)$ restricted to the subspace $C(x_*)$), we also require δ to be small enough to ensure that

$$\|\nabla^2 f(x) - \nabla^2 f(x_*)\| \leqq c_7 \overset{\text{def}}{=} \frac{1}{4} \min\left[1, \psi, \frac{\varepsilon}{2}\right] \tag{179}$$

for all $x \in A(x_*)$ with $\|x - x_*\| \leqq \delta$. This is possible because of (AS.2). We also apply Theorem 14 and deduce that

$$I(x_{k_i}) = I(x_*) \tag{180}$$

for all i sufficiently large. We can then choose $i_2 \geqq i_1$ large enough so that

$$\lambda^1[B_{k_i}] \geqq \tfrac{1}{2}\varepsilon, \tag{181}$$

$$\|x_{k_i} - x_*\| \leqq \frac{\varepsilon\delta}{4c_7 + \varepsilon} \stackrel{\text{def}}{=} \delta_1 \tag{182}$$

and (180) hold for all $i \geqq i_2$, and also so that

$$h_k \leqq \frac{c_7\delta_1}{\sigma_1^2\sigma_2^2} < \delta \tag{183}$$

for all $k \geqq k_{i_2}$. The first of these inequalities has to hold for large enough k because of Theorem 11. The second results from the definitions of c_7, δ_1, σ_1 and σ_2 in (179), (182), (AS.3) and (AS.5), respectively. For all $i \geqq i_2$ we now observe that, using (180),

$$s_{k_i} \in C(x_*) \tag{184}$$

and we decompose w_{k_i} and g_{k_i} as

$$w_{k_i} = w_{k_i}^R + w_{k_i}^N \quad \text{and} \quad g_{k_i} = g_{k_i}^R + g_{k_i}^N \tag{185}$$

where the vectors superscripted with N belong to $C(x_*)$, while those superscripted with R belong to $C(x_*)^\perp$. Now, the definition of δ_1 and the second part of (183) imply that, for $i \geqq i_2$, $h_{k_1} = \|w_{k_i}^N\|$. Observe then that the vector $g_{k_i}^N$ is the gradient of the model at x_{k_i} projected onto $C(x_*)$, and that it has the property that, for $i \geqq i_2$,

$$\|g_{k_i}^N\| \leqq \sigma_2^2 \|w_{k_i}^N\| = \sigma_2^2 h_{k_i} \leqq c_7\delta_1, \tag{186}$$

where we used (7), (AS.5), the first part of (183) and the fact that $\sigma_1 \geqq 1$. Consider now the one-dimensional strictly convex quadratic function of the parameter τ defined by

$$\phi(\tau) \stackrel{\text{def}}{=} m_{k_i}(x_{k_i} + \tau s_{k_i}) - f(x_{k_i}). \tag{187}$$

Then, since $\phi(0) = 0$ and $\phi(1) \leqq 0$ by construction of the step s_{k_i}, we obtain that

$$\tau_* \stackrel{\text{def}}{=} \arg\min \phi(\tau) \geqq \tfrac{1}{2}. \tag{188}$$

But an easy computation shows that

$$\tau_* = \frac{|g_{k_i}^T s_{k_i}|}{s_{k_i}^T B_{k_i} s_{k_i}} \leqq \frac{2\|g_{k_i}^N\|}{\varepsilon \|s_{k_i}\|}, \tag{189}$$

where we have used the Cauchy-Schwarz inequality, (181) and (184) to derive the last part of this bound. Therefore, we can deduce that

$$\|s_{k_i}\| \leqq \frac{4}{\varepsilon} \|g_{k_i}^N\|. \tag{190}$$

Hence, gathering (182), (186) and (190), we obtain that

$$\|x_{k_i+1} - x_*\| \leqq \|s_{k_i}\| + \|x_{k_i} - x_*\| \leqq \left[\frac{4c_7}{\varepsilon} + 1\right] \delta_1 = \delta. \tag{191}$$

Now assume that

$$\|x_{k_i+1} - x_*\| > \delta_1 \tag{192}$$

and observe that

$$g^N_{k_i+1} = Q_* g_{k_i+1} = Q_*[\nabla f(x_*) + G_*(x_{k_i+1} - x_*)] \tag{193}$$

where

$$G_* = \int_0^1 \nabla^2 f(x_{k_i+1} + t(x_* - x_{k_i+1}))\, dt \tag{194}$$

and $Q_* \overset{\text{def}}{=} Z_* Z_*^T$ is the orthogonal projector onto the subspace $C(x_*)$. But $Q_* \nabla f(x_*) = 0$ because x_* is critical and, by (180) and (184), $x_{k_i+1} - x_* \in C(x_*)$. Hence

$$\|g^N_{k_i+1}\| = \|Q_* \nabla^2 f(x_*) Q_*(x_{k_i+1} - x_*) + Q_*(G_* - \nabla^2 f(x_*)) Q_*(x_{k_i+1} - x_*)\|. \tag{195}$$

This implies, by using the definition of Z_*, the fact that $\|Q_*\| = 1$, (192) and the Cauchy–Schwarz inequality, that

$$\begin{aligned} \|g^N_{k_i+1}\| &\geqq \|Z_*[Z_*^T \nabla^2 f(x_*) Z_*] Z_*^T (x_{k_i+1} - x_*)\| - \|G_* - \nabla^2 f(x_*)\| \, \|x_{k_i+1} - x_*\| \\ &> \psi\delta_1 - \|G_* - \nabla^2 f(x_*)\| \delta. \end{aligned} \tag{196}$$

Now,

$$\begin{aligned} \|G_* - \nabla^2 f(x_*)\| &= \left\| \int_0^1 [\nabla^2 f(x_{k_i+1} + t(x_* - x_{k_i+1})) - \nabla^2 f(x_*)]\, dt \right\| \\ &\leqq \max_{t \in [0,1]} \|\nabla^2 f(x_{k_i+1} + t(x_* - x_{k_i+1})) - \nabla^2 f(x_*)\| \\ &\leqq c_7. \end{aligned} \tag{197}$$

Hence, using (7), (186), (196), (197), the definition of δ_1 in (182), the definition of c_7 in (179) and the fact that $\sigma_1 \geqq 1$,

$$h_{k_i+1} > \frac{\psi\delta_1 - c_7\delta}{\sigma_2^2} \geqq \frac{\delta_1 \psi}{4\sigma_2^2}\left(4 - \frac{4c_7 + \varepsilon}{\varepsilon}\right) \geqq \frac{c_7 \delta_1 (3\varepsilon - 4c_7)}{\sigma_2^2 \varepsilon} \geqq \frac{c_7 \delta_1}{\sigma_1^2 \sigma_2^2}. \tag{198}$$

This is impossible because of (183), and therefore

$$\|x_{k_i+1} - x_*\| \leqq \delta_1. \tag{199}$$

All the conditions that are satisfied at x_{k_i} are thus satisfied again at x_{k_i+1}, and the argument can be applied recursively to show that, for all $j \geqq 1$,

$$\|x_{k_i+j} - x_*\| \leqq \delta_1 \leqq \delta. \tag{200}$$

Since δ is arbitrarily small, this proves the convergence of the complete sequence $\{x_k\}$ to x_*. □

This result is interesting because it confirms the intuition that the algorithm can be forced to converge to a critical point which is not a minimizer, when the Hessian approximations B_k do not reflect adequately the behaviour of $\nabla^2 f(x)$.

Since the final calculations are purely unconstrained, we can also apply Moré's results in [12] and deduce that all iterations are eventually successful, and that the trust region radii Δ_k are bounded away from zero.

We now wish to show that convergence to a point where the necessary second-order conditions for a local minimizer hold can also be established, if one is willing to strengthen the requirements on the step s_k.

We shall impose the following conditions instead of (12).

(AS.8) The choice of the step s_k ensures that

$$f(x_k) - m_k(x_k + s_k) \geqq \beta_1[f(x_k) - \min m_k(x_k^C + p_k)], \tag{201}$$

where the minimum is taken over eigenvectors $p_k \in C(x_k^C)$ associated with $\lambda^1[B_k]$ that are scaled so that the point $x_k^C + p_k$ still lies in the trust region.

We also require that some step along a direction of negative curvature can be made, when such a direction is found, as ensured by the condition

$$\beta_2 > \nu. \tag{202}$$

(AS.9)

We will finally require that the model reflects the behaviour of the objective function more accurately (for example, by using exact Hessians or finite difference approximations). Thus, we require that the following two conditions hold.

(AS.10) The matrices B_k satisfy the conditions

$$\lambda^1[B_k] \leqq c_8 \lambda^1[\nabla^2 f(x_k)] \quad \text{when } \lambda^1[\nabla^2 f(x_k)] < 0, \tag{203}$$

where c_8 is some positive constant, and

$$\lim_{k\to\infty} |r(B_k, s_k) - r(\nabla^2 f(x_k), s_k)| = 0 \quad \text{whenever } \lim_{k\to\infty} \|s_k\| = 0, \tag{204}$$

where $r(X, s) = s^T X s / \|s\|^2$ is the Rayleigh quotient of the symmetric matrix X with respect to the direction s (it can be viewed as a measure of the curvature of the quadratic form defined by X along s).

We first consider a consequence of these conditions on the model decrease.

LEMMA 16. *Assume that* (AS.1)-(AS.3), (AS.5) *and* (AS.8) *hold. Assume, furthermore, that*

$$\lambda^1[B_k] < 0 \tag{205}$$

for some k. Then

$$f(x_k) - m_k(x_k + s_k) \geqq -\tfrac{1}{2} c_9 \lambda^1[B_k] \Delta_k^2, \tag{206}$$

where the constant c_9 is defined by

$$c_9 \overset{\text{def}}{=} \beta_1 \left(\frac{\beta_2 - \nu}{\sigma_2} \right)^2. \tag{207}$$

Proof. Consider first $p_k \in C(x_*)$ an eigenvector of B_k associated with $\lambda^1[B_k]$, and assume it is scaled so that

$$p_k^T (g_k + B_k(x_k^C - x_k)) \leqq 0 \tag{208}$$

and

$$\|D_k(x_k^C + p_k - x_k)\| = \beta_2 \Delta_k. \tag{209}$$

If we set $x_k^P \overset{\text{def}}{=} x_k^C + p_k$, we obtain that

$$\begin{aligned} m_k(x_k^P) &= m_k(x_k^C) + p_k^T (g_k + B_k(x_k^C - x_k)) + \tfrac{1}{2} p_k^T B_k p_k \\ &\leqq f(x_k) + \tfrac{1}{2} \lambda^1[B_k] \|p_k\|^2, \end{aligned} \tag{210}$$

where we used the inequality $m_k(x_k^C) < f(x_k)$, (208) and the definition of p_k. Now observe that

$$\frac{\|D_k p_k\|}{\|D_k(x_k^C - x_k)\|} \geqq \frac{\|D_k(x_k^P - x_k)\|}{\|D_k(x_k^C - x_k)\|} - 1 \geqq \frac{\beta_2 - \nu}{\nu}, \tag{211}$$

because of (10) and (209), and hence

$$\beta_2 \Delta_k = \|D_k(x_k^P - x_k)\| \leqq \|D_k p_k\| + \|D_k(x_k^C - x_k)\| \leqq \left(1 + \frac{\nu}{\beta_2 - \nu}\right) \sigma_2 \|p_k\|. \tag{212}$$

This last inequality and relation (210) together then imply that

$$f(x_k) - m_k(x_k^P) \geqq -\frac{1}{2} \lambda^1[B_k] \left(\frac{\beta_2 - \nu}{\sigma_2}\right)^2 \Delta_k^2. \tag{213}$$

To complete the proof, one only needs to notice that

$$f(x_k) - m_k(x_k + s_k) \geqq \beta_1 [f(x_k) - m_k(x_k^P)] \tag{214}$$

because of (201). □

We now show that there is a limit point where the second-order necessary conditions for a minimizer are satisfied.

THEOREM 17. *Assume that* (AS.1)-(AS.5) *and* (AS.8)-(AS.10) *hold. Then there is a limit point* x_* *of the sequence of iterates generated by the algorithm, with* $\nabla^2 f(x_*)$ *positive semidefinite on* $C(x_*)$.

Proof. We proceed again by contradiction, and assume that, for all limit points x_* of the sequence $\{x_k\}$, the Hessian matrix $\nabla^2 f(x_*)$ has an eigenvector in $C(x_*)$ corresponding to a negative eigenvalue bounded above by $-2\varepsilon_1$, where ε_1 is some positive constant. We first want to show that the trust region radii Δ_k tend to zero. Assume it is not true, that is, there exists a subsequence $\{m_i\}$ of successful iterations such that

$$\Delta_{m_i} \geqq \varepsilon_2 \tag{215}$$

for some $\varepsilon_2 > 0$. Then, because of (AS.1), we can exhibit a subsequence of this subsequence which is converging to a limit point x_*, say. Without loss of generality, we assume that the entire sequence $\{x_{m_i}\}$ converges to x_*. We will also assume that i is large enough to ensure that

$$I(x_{m_i}) \subseteq I(x_*) \quad \text{and} \quad \lambda^1[\nabla^2 f(x_{m_i})] \leqq -\varepsilon_1 \tag{216}$$

by using the continuity of the Hessian. Hence (203) implies that

$$\lambda^1[B_{m_i}] \leqq -c_8 \varepsilon_1 \tag{217}$$

for i sufficiently large. Using this bound, Lemma 16, the fact that iteration m_i is successful and (215), we deduce that

$$f(x_{m_i}) - f(x_{m_i+1}) \geqq \tfrac{1}{2} \mu c_8 c_9 \varepsilon_1 \varepsilon_2^2 \tag{218}$$

for large i. But this inequality is clearly impossible because

$$\sum_{i=0}^{\infty} [f(x_{m_i}) - f(x_{m_i+1})] \leqq \sum_{k \in S} [f(x_k) - f(x_{k+1})] \leqq f(x_0) - f(x_*) < +\infty. \tag{219}$$

Hence no subsequence of successful iterations is such that (215) holds, and

$$\lim_{k \in S} \Delta_k = 0. \tag{220}$$

This, in turn, implies that

$$\lim_{k\to\infty} \Delta_k = 0 \quad \text{and} \quad \lim_{k\to\infty} \|s_k\| = 0, \tag{221}$$

by using (13), (22) and (24). We note that, for k sufficiently large, the point x_k is close enough to a limit point to ensure that

$$\lambda^1[\nabla^2 f(x_k)] \leqq -\varepsilon_1 \quad \text{and hence } \lambda^1[B_k] \leqq -c_8\varepsilon_1, \tag{222}$$

because of (203). Combining now (105), the bound $\|s_k\| \leqq \sigma_1\beta_2\Delta_k$ and (206) together, we obtain that

$$|\rho_k - 1| \leqq \frac{\sigma_1^2\beta_2^2}{c_8 c_9 \varepsilon_1} |r(B_k, s_k) - r(\nabla^2 f(x_k), s_k)| \tag{223}$$

and the right-hand side of this expression tends to zero because of the second part of (221) and (204). The updating rules for the trust region radius then prevent Δ_k from tending to zero. However, this contradicts the first part of (221); therefore, there must be a limit point x_* where $\nabla^2 f(x_*)$ is positive semidefinite. □

Strictly speaking, this result does not ensure that x_* is a local minimizer, because we did not show that it is critical, or that it is not a saddle point. Nevertheless it is often the case that x_* is a local minimizer, and criticality can be guaranteed by imposing (AS.6) as shown by Theorem 11.

Finally observe that we really proved that there is a limit point where the matrices B_k are asymptotically positive semidefinite, and it is only because we imposed an adequate relationship between these matrices and the true Hessian that the theorem's statement involves the latter.

4. Conclusions and perspectives. We believe that the theory presented in this paper is interesting for several reasons.

First, it extends most of the convergence results known for unconstrained problems to the very frequent case where bounds on the variables are present. This extension is obtained by generalizing the now classical notion of a Cauchy point in what seems to us a natural way. Quite general conditions on the size of the Hessian approximations are also considered, allowing for a number of specialized implementations.

An important feature of the algorithm presented is that it does not require successive multidimensional unconstrained minimization in a single iteration, at variance with Gay's proposal in [8]. Gay's method may indeed require the complete numerical solution of more than one linear system for computing a single step, and is therefore quite costly when applied on large-dimensional problems. Instead, our proposal replaces this calculation by a simple linesearch along a piecewise linear arc, possibly followed by a single refinement of the step. Efficient iterative methods such as truncated preconditioned conjugate gradients can then be used for this last phase. The new strategy is thus much cheaper to implement for problems involving a large number of variables. Hence, the framework presented here is quite well suited to such problems. In particular, one may consider using it in conjunction with partitioned secant updating techniques on the very general class of partially separable problems [10]. These last techniques have already been used for bounded problems in the Harwell library subroutine VE08, which was shown in [11] to be remarkably efficient, although it still lacks the strong theoretical foundation that we provide for the present proposal.

Finally, preliminary numerical experience with this type of algorithms is rather encouraging (see [5]). The theory developed here therefore may well prove to be useful in practical applications.

Further extensions of this framework to the linearly and nonlinearly constrained case are also of interest. They are the subject of continuing research.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] R. K. BRAYTON AND J. CULLUM, *An algorithm for minimizing a differentiable function subject to box constraints and errors*, J. Optim. Theory Appl., 29 (1979), pp. 521–558.

[3] R. H. BYRD, R. B. SCHNABEL AND G. A. SCHULTZ, *A trust region algorithm for nonlinearly constrained optimization*, Technical Report CU-CS-313-85, Department of Computer Sciences, University of Colorado, Boulder, CO, 1985.

[4] M. R. CELIS, J. E. DENNIS AND R. A. TAPIA, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization 1984, P. T. Boggs, R. H. Byrd and R. B. Schnabel, eds., 1985, pp. 71–82.

[5] A. R. CONN, N. I. M. GOULD AND PH. L. TOINT, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Technical Report #86/3, Department of Mathematics, Facultés Universitaires ND de la Paix, Namur, Belgium, 1986; Math. Comp., to appear.

[6] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] D. M. GAY, *Computing optimal locally constrained steps*, SIAM J. Statist. Sci. Comput., 2 (1981), pp. 186–197.

[8] ———, *A trust region approach to linearly constrained optimization*, in Numerical Analysis: Proceedings Dundee 1983, D. F. Griffiths, ed., Lecture Notes in Mathematics 1066, Springer-Verlag, Berlin, 1984, pp. 72–105.

[9] P. E. GILL, W. MURRAY AND M. H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.

[10] A. GRIEWANK AND PH. L. TOINT, *Partitioned variable metric updates for large structured optimization problems*, Numer. Math., 39 (1982), pp. 119–137.

[11] ———, *Numerical experiments with partially separable optimization problems*, in Numerical Analysis: Proceedings Dundee 1983, D. F. Griffiths, ed., Lecture Notes in Mathematics 1066, Springer-Verlag, Berlin, 1984, pp. 203–220.

[12] 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.

[13] M. J. D. POWELL, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, Society for Industrial and Applied Mathematics–American Mathematical Society Proc. 9, 1976, pp. 53–72.

[14] ———, *On the global convergence of trust region algorithms for unconstrained minimization*, Math. Programming, 29 (1984), pp. 297–303.

[15] M. J. D. POWELL AND Y. YUAN, *A trust region algorithm for equality constrained optimization*, Report DAMTP 1986-NA2, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, 1986.

[16] A. VARDI, *A trust region algorithm for equality constrained minimization: convergence properties and implementation*, this Journal, 22 (1985), pp. 575–591.