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

Research Report CS-86-31

August 1986

This report is simultaneously issued at the University of Waterloo, the A.E.R.E. at Harwell, and the FUNDP at Namur.

Global convergence of a class of trust region algorithms for optimization with simple bounds.

by A.R. Conn†, N.I.M. Gould‡ and Ph.L. Toint§

August 17, 1986

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 size of the Hessian approximations are considered. It is also shown that the proposed algorithms reduce to an unconstrained calculation after finitely many iterations, allowing a fast asymptotic rate of convergence.

†Department of Computer Sciences
University of Waterloo
Waterloo, Canada

‡A.E.R.E.
Harwell, Great-Britain

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

Keywords : Trust regions, convergence theory, optimization with bounds.
Abbreviated title : Trust regions for bounded optimization.
The work of the first author was supported in part by NSERC grant A8639.

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, it is, in a way, 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 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 reason 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 contraints have been investigated, and a variety of solutions 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 details, while section 3 is devoted to the convergence theory. Some conclusions are drawn in section 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

$$l_i \leq x_i \leq u_i \quad (i = 1, \ldots, n). \tag{2}$$

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 $\mathcal{L}$ as the intersection of the set

$$\{x \in \mathbf{R}^n \mid f(x) \leq f(x_0)\}$$

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

AS.1 The set $\mathcal{L}$ is compact and has a non empty interior.

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

The first of these assumptions says that the optimization problem is non trivial and cannot be reduced to a lower dimension. In particular, infinite (positive and/or negative) bounds are allowed provided the set $\mathcal{L}$ 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 active at the point x and

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

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

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

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

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

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

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

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 a direction w_k satisfying 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 non-negative.) This vector 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

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

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 $t \geq 0$, 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 constraint defined by

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

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 picture.

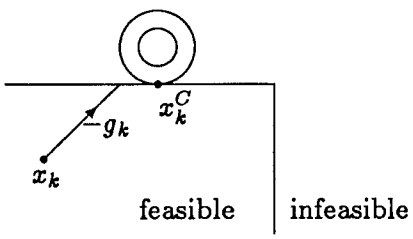

Fig. 1: The generalized Cauchy point.

We require that the step s_k satisfies the following three conditions,

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

$$\|D_k s_k\| \leq \beta_2 \Delta_k \tag{13}$$

and

$$I(x_k^C) \subseteq I(x_k + s_k). \tag{14}$$

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

$$\beta_1 \in (0, 1] \text{ and } \beta_2 \geq \nu. \tag{15}$$

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 a extended trust region, and the third condition ensures that all bounds that are active at the generalized Cauchy point are still active at $x_k + s_k$.

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 constant ν and β_2 in (10) and (13) respectively, other norms 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.

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

$$s_k = -B_k^{-1} g_k, \tag{16}$$

where B_k is positive definite, whenever $\|D_k s_k\| < \Delta_k$, in contrast with [14]. This assumption may indeed be undesirable when n is large. In this context, the calculation of such a direction 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

$$C_k = D_k^T D_k. \tag{17}$$

Note that, in practical implementations of this flexible method, a more sophisticate 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

$$0 < \gamma_0 \leq \gamma_1 < 1 \leq \gamma_2. \tag{18}$$

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(x_k + s_k)}. \tag{19}$$

step 3 In the case where

$$\rho_k > \mu, \tag{20}$$

set

$$x_{k+1} = x_k + s_k, \quad g_{k+1} = \nabla f(x_{k+1}) \tag{21}$$

and

$$\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k] \text{ if } \rho_k \geq \eta \tag{22}$$

or

$$\Delta_{k+1} \in [\gamma_1 \Delta_k, \Delta_k] \text{ if } \rho_k < \eta. \tag{23}$$

Otherwise, set

$$x_{k+1} = x_k, \quad g_{k+1} = g_k \tag{24}$$

and

$$\Delta_{k+1} \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k]. \tag{25}$$

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 (20) 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 (20) 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 uniformly bounded inverses, that is

$$\|D_k^{-1}\| \leq \sigma_1 \tag{26}$$

for some $\sigma_1 \geq 1$.

Condition (26) is identical to that imposed in [12]. As in this last reference, we observe that this condition does not imply that the scaling matrices have uniformly bounded condition numbers.

(AS.3) also allows us to characterize critical points of our problem as expressed in the following statement.

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

$$P[x - tw] = x \tag{27}$$

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

$$D^T D w = \nabla f(x). \tag{28}$$

This lemma results immediately from Proposition 1.35 in [1]. Observe now that, if we define

$$d_k(t) = P[x_k - tw_k] - x_k \tag{29}$$

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 $t \geq 0$. The quantity

$$h_k \stackrel{\text{def}}{=} \|P[x_k - w_k] - x_k\| = \|d_k(1)\|, \tag{30}$$

can therefore be regarded as a measure of the "criticality" of the kth iterate. The geometric interpretation of this quantity is illustrated in Fig. 2.

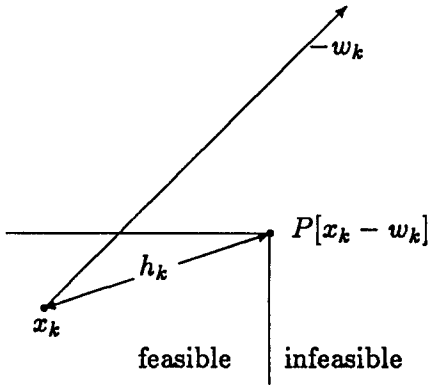

Fig. 2: The critical length h_k.

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

$$b_k = 1 + \max_{i=0,\ldots,k} \|D_i^{-T} B_i D_i^{-1}\|, \tag{31}$$

we require that

$$\sum_{k=0}^{\infty} \frac{1}{b_k} = +\infty. \tag{32}$$

Note that we 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 (32) 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 (32) is violated, the algorithm can converge to a noncritical point. It is also worth remembering that the well known BFGS secant update does satisfies (AS.4) on convex problems (see [13]). This is also the case of a suitably safeguarded Symmetric Rank One update applied on 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 $\mathcal{L}$, 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 generalize the important condition that, for a given successful iteration, iteration k say,

$$f(x_k) - f(x_k + s_k) \geq c_1 \|D_k^{-T} g_k\| \min[\Delta_k, \frac{\|D_k^{-T} g_k\|}{\|D_k^{-T} B_k D_k^{-1}\| + 1}]. \tag{33}$$

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 the algorithm determines the set of bounds that are active at the solution in a finite number of iterations. This means that, asymptotically, its rate of convergence is that of a purely unconstrained method. The third part of the analysis is concerned with guaranteeing the convergence to a local minimum, and not merely a critical point, and with the conditions to impose on the step s_k to 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

$$x_k^s = D_k x_k, \quad s_k^s = D_k s_k, \quad d_k^s(t) = D_k d_k(t) \tag{34}$$

and

$$g_k^s = D_k^{-T} g_k, \quad B_k^s = D_k^{-T} B_k D_k^{-1}. \tag{35}$$

Observe also that

$$w_k^s = D_k w_k = g_k^s, \tag{36}$$

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

$$\begin{aligned} q_k(t) &= m_k(x_k + d_k(t)) \\ &= f(x_k) + g_k^T d_k(t) + \tfrac{1}{2}[d_k(t)]^T B_k d_k(t) \\ &= f^s(x_k^s) + [g_k^s]^T d_k^s(t) + \tfrac{1}{2}[d_k^s(t)]^T B_k^s d_k^s(t), \end{aligned} \tag{37}$$

where we redefine the objective function on the scaled space by

$$f^s(x^s) = f(D_k^{-1} x^s) = f(x). \tag{38}$$

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

$$l_i^s \leq x_i^s \leq u_i^s \quad (i = 1, \ldots, n), \tag{39}$$

with

$$l_i^s = (D_k)_{ii} l_i \text{ and } u_i^s = (D_k)_{ii} u_i. \tag{40}$$

These bounds, in turn, allow the definition of an index of the scaled active constraints

$$I^s(x^s) = I(x), \tag{41}$$

and of a scaled $P^s[\cdot]$ operator as in (5), corresponding to the projection onto the feasible domain defined by (39). Similarly also, a scaled $\mathcal{L}^s$ can be defined using (38) and (39). Then we have that

$$d_k^s(t) = D_k(P[x_k - tw_k] - x_k) = P^s[x_k^s - tg_k^s] - x_k^s \tag{42}$$

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

$$\|P^s[x_k^s - tg_k^s] - x_k^s\| \leq \nu\Delta_k \tag{43}$$

and

$$\|s_k^s\| \leq \beta_2\Delta_k \tag{44}$$

respectively. Observe finally that, for all x,

$$\|x\| \leq \|D_k^{-1}\|.\|x^s\| \leq \sigma_1\|x^s\| \tag{45}$$

To simplify the notation, we will use this one to one correspondence between the original and scaled spaces at iteration k, and therefore assume that the scaling matrix $D_k = I$. 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 $0 < t_1 \leq t_2$ and all k, we have that*

$$I(x_k + d_k(t_1)) \subseteq I(x_k + d_k(t_2)). \tag{46}$$

This lemma is trivial once we observe that the set of active bounds can only be increased 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,

$$[z_k(y)]_i \stackrel{\text{def}}{=} \begin{cases} [g_k]_i & \text{if } i \notin I(y), \\ 0 & \text{otherwise,} \end{cases} \tag{47}$$

where the subscript i denotes the ith component of the vector.

Lemma 3 *Assume that (AS.1)-(AS.2) hold. Assume also that $h_k > 0$. Then, if*

$$c_2 \stackrel{\text{def}}{=} \max\left[1, \max_{x \in \mathcal{L}} \|\nabla f(x)\|\right], \tag{48}$$

we obtain that

$$\|z_k(x_k + d_k(t_k^{(1)}))\| \geq \frac{1}{2}h_k \tag{49}$$

where

$$t_k^{(1)} = \frac{1}{2}\frac{h_k}{c_2}. \tag{50}$$

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

$$\|d_k(t_k^{(1)})\| \leq \|t_k^{(1)} g_k\| \leq \frac{1}{2}\frac{h_k}{c_2}\|g_k\| \leq \frac{1}{2}h_k. \tag{51}$$

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

$$\|d_k(t)\| = h_k. \tag{52}$$

Then, using (51), one obtains that

$$0 < t_k^{(1)} < t_k^{(2)} \leq 1. \tag{53}$$

Hence, because of (51) and (52),

$$\begin{aligned} \|z_k(x_k + d_k(t_k^{(1)}))\| &\geq (t_k^{(2)} - t_k^{(1)})\|z_k(x_k + d_k(t_k^{(1)}))\| \\ &\geq \|d_k(t_k^{(2)}) - d_k(t_k^{(1)})\| \\ &\geq \|d_k(t_k^{(2)})\| - \|d_k(t_k^{(1)})\| \\ &\geq \tfrac{1}{2}h_k, \end{aligned} \tag{54}$$

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 non critical point.

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

$$\alpha_k \stackrel{\text{def}}{=} \|z_k(x_k + d_k(t_k^{(3)}))\| > 0. \tag{55}$$

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

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

and

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

where (56) only holds where the derivative of $q_k(t)$ *is defined, and where*

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

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. Consider now the set T of points in $[0, t_k^{(3)}]$ where the piecewise quadratic $q_k(t)$ is differentiable, that is the complete interval 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 $i = 1, \ldots, n$. We now assume, without loss of generality, that if bounds 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,

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

where the function $\delta(condition)$ 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):

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

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

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

Therefore we obtain the following equivalences

$$\begin{array}{lcl} t \in [0, \omega_i] & \Leftrightarrow & i \geq r, \\ t \in (\omega_i, t_k^{(3)}] & \Leftrightarrow & i \leq r-1, \end{array} \tag{67}$$

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

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

while

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

Gathering (68) and (69), we obtain

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

where B_k^* is defined by

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

Observe now that

$$\|B_k^*\| \leq \|B_k^*\|_F \leq \|B_k\|_F \leq \sqrt{n}\|B_k\|. \tag{72}$$

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

$$\frac{d}{dt}q_k(t) \leq -\|v_k\|^2 + t[v_k]^T B_k v_k + t_k^{(3)}\sqrt{n}\|g_k\|^2\|B_k\|. \tag{73}$$

Hence, for all $t \in T$,

$$\frac{d}{dt}q_k(t) \leq -\alpha_k^2 + t_k^{(3)}c_2^2(1+\sqrt{n})\|B_k\|, \tag{74}$$

where we have used (48), the Cauchy-Schwarz inequality and the fact that Lemma 2 implies the inequality $\|v_k\| \geq \alpha_k$. In particular, if we consider $t_k^{(4)}$ as defined by (58), we obtain that

$$\frac{d}{dt}q_k(t) \leq -\alpha_k^2 + t_k^{(4)}c_2^2(1+\sqrt{n})\|B_k\| \leq -\frac{1}{2}\alpha_k^2 \tag{75}$$

(when the derivative exists), and, consequently, that

$$q_k(t) \leq f(x_k) - \frac{1}{2}\alpha_k^2 t \tag{76}$$

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.

We can now state the equivalent of condition (33) 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. Assume also that*

$$h_k^s = \|D_k(P[x_k - w_k] - x_k)\| > 0. \tag{77}$$

Then

$$f(x_k) - m_k(x_k + s_k) \geq \frac{1}{2}c_3[h_k^s]^2 \min\left[\frac{[h_k^s]^2}{b_k}, \Delta_k\right], \tag{78}$$

where

$$c_3 \stackrel{\text{def}}{=} \min\left[\frac{\beta_2^2}{1-\eta}, \frac{\min(1,\nu)\beta_1}{32c_2^2\sigma_1^2(1+\sqrt{n})}\right] \leq 1. \tag{79}$$

If the kth iteration is successful, we have also that

$$f(x_k) - f(x_k + s_k) \geq \frac{1}{2}\mu c_3[h_k^s]^2 \min\left[\frac{[h_k^s]^2}{b_k}, \Delta_k\right]. \tag{80}$$

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 $\alpha_k \geq \frac{1}{2} h_k > 0$, and relation (57) implies that

$$q_k(t) \leq f(x_k) - \frac{1}{8} h_k^2 t, \tag{81}$$

provided that t is in an interval whose upper bound is

$$t_k^{(5)} \stackrel{\text{def}}{=} \min \left[\frac{h_k}{2c_2}, \frac{1}{8c_2^2(1+\sqrt{n})} \frac{h_k^2}{(\|B_k\|+1)} \right] = \frac{1}{8c_2^2(1+\sqrt{n})} \frac{h_k^2}{(\|B_k\|+1)}, \tag{82}$$

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

$$f(x_k) - m_k(x_k + s_k) \geq \frac{1}{8} \beta_1 h_k^2 t_k^{(5)}. \tag{83}$$

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

$$\|d_k \left(\frac{\nu \Delta_k}{c_2} \right) \| \leq \frac{\nu \Delta_k}{c_2} \|g_k\| \leq \nu \Delta_k, \tag{84}$$

we can deduce that

$$t_k^C \geq \frac{\nu \Delta_k}{c_2}. \tag{85}$$

Therefore (81) implies that

$$f(x_k) - m_k(x_k + s_k) \geq \frac{\nu \beta_1}{8c_2} h_k^2 \Delta_k. \tag{86}$$

Gathering (82), (83) and (86), we obtain that

$$f(x_k) - m_k(x_k + s_k) \geq \frac{\min(1, \nu)\beta_1}{64c_2^2(1+\sqrt{n})} h_k^2 \min \left[\frac{h_k^2}{\|B_k\|+1}, \Delta_k \right]. \tag{87}$$

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 (26). 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 (87) is thus nothing but the scaled form of (78), except that the constant c_3 is possibly further reduced, in order to simplify another development below. The inequality (80) then results from (78), (19) and (20). □

This closes the first part of our convergence theory, devoted to finding a suitable generalization of condition (33) 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 will be indeed 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 6 *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 $\epsilon > 0$ such that, for all k,*

$$h_k^s \geq \epsilon, \tag{88}$$

where h_k^s is defined by (77). Then there exist a constant $c_4 > 0$ such that, for all $k \geq 1$,

$$\Delta_k \geq \frac{c_4}{b_k}, \tag{89}$$

where b_k is defined by (31).

In order to prove the lemma, we first define

$$c_5 \stackrel{\text{def}}{=} \max_{x \in \mathcal{L}} \|\nabla^2 f(x)\|, \tag{90}$$

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

$$\epsilon < \min\left(1, \beta_2 \sqrt{\frac{(c_5\sigma_1^2 + b_0)\Delta_0}{\gamma_0 c_3 (1-\eta)}}\right). \tag{91}$$

We have also that

$$|f(x_k + s_k) - m_k(x_k + s_k)| = \frac{1}{2}|s_k^T (G_k - B_k) s_k| \tag{92}$$

because of (6), where

$$G_k = \int_0^1 \nabla^2 f(x_k + t s_k)\, dt. \tag{93}$$

This yields that

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq \frac{1}{2}\beta_2^2 \Delta_k^2 \|D_k^{-T}(G_k - B_k) D_k^{-1}\| \leq \frac{1}{2}\beta_2^2 \Delta_k^2 [c_5\sigma_1^2 + b_k], \tag{94}$$

where we have used (AS.3), (13), (31),(34), and the inequality $\|G_k\| \leq c_5$. Assume now that there is a k such that

$$(c_5\sigma_1^2 + b_k)\Delta_k \leq \gamma_0(1-\eta)c_3 \frac{\epsilon^2}{\beta_2^2}, \tag{95}$$

and define r as the first iteration number such that (95) holds. (Note that (91) implies that (95) does not hold for $k = 0$, and hence $r \geq 1$.) Then the mechanism of the algorithm ensures that

$$b_{r-1}\Delta_{r-1} \leq (c_5\sigma_1^2 + b_{r-1})\Delta_{r-1} \leq (c_5\sigma_1^2 + b_r)\frac{\Delta_r}{\gamma_0} \leq (1-\eta)c_3 \frac{\epsilon^2}{\beta_2^2} \leq \epsilon^2, \tag{96}$$

where we used (79) to derive the last inequality. This last inequality, (31), (88), (91) and Lemma 5 then imply that

$$f(x_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1}) \geq \frac{1}{2} c_3 \epsilon^2 \min[\frac{\epsilon^2}{b_{r-1}}, \Delta_{r-1}] \geq \frac{1}{2} c_3 \epsilon^2 \Delta_{r-1}. \tag{97}$$

Combining (94) and (97), we obtain that

$$|\rho_{r-1} - 1| \leq \frac{|f(x_{r-1} + s_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1})|}{|f(x_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1})|} \leq \frac{\beta_2^2 (c_5 \sigma_1^2 + b_{r-1}) \Delta_{r-1}}{c_3 \epsilon^2} \leq 1 - \eta, \tag{98}$$

where we also used (19) and the middle part of (96). This imposes $\rho_{r-1} \geq \eta$, and therefore (22) implies that $\Delta_r \geq \Delta_{r-1}$. This, in turn, gives that

$$(c_5 \sigma_1^2 + b_{r-1}) \Delta_{r-1} \leq (c_5 \sigma_1^2 + b_r) \Delta_r \leq \gamma_0 (1 - \eta) c_3 \frac{\epsilon^2}{\beta_2^2}, \tag{99}$$

which contradicts the assumption that r was the first index with (95) satisfied. Hence, (95) never holds, and we obtain that

$$(c_5 \sigma_1^2 + b_k) \Delta_k > \gamma_0 (1 - \eta) c_3 \frac{\epsilon^2}{\beta_2^2} \tag{100}$$

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

$$c_5 \sigma_1^2 + b_k \leq c_5 \sigma_1^2 (b_k + 1) \leq 2 c_5 \sigma_1^2 b_k \tag{101}$$

for all such k, we have proved (89) with

$$c_4 = \frac{\gamma_0 (1 - \eta) c_3 \epsilon^2}{2 c_5 \sigma_1^2 \beta_2^2}. \tag{102}$$

□

We are now ready for our main convergence theorem.

Theorem 7 *Assume that (AS.1)-(AS.4) hold. Assume also that $\{x_k\}$ is a sequence of iterates generated by the algorithm. Then*

$$\liminf_{k \to \infty} h_k = 0, \tag{103}$$

where h_k is defined in (30).

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

$$h_k \geq \epsilon \sigma_1 \tag{104}$$

for all k, implying that

$$h_k^s \geq \epsilon \tag{105}$$

because of (45). Now Lemma 5 and the fact that the objective function is bounded below on the set $\mathcal{L}$ imply that

$$\frac{1}{2} \mu c_3 \epsilon^2 \sum_{k \in S} \min \left[\frac{\epsilon^2}{b_k}, \Delta_k \right] \leq \sum_{k \in S} [f(x_k) - f(x_{k+1})] < +\infty, \tag{106}$$

where S is the set of successful iterations. Using this inequality and Lemma 6, we may deduce that the sum

$$\min(\epsilon^2, c_4) \sum_{k \in S} \frac{1}{b_k} \leq \sum_{k \in S} \min \left[\frac{\epsilon^2}{b_k}, \frac{c_4}{b_k} \right] \leq \sum_{k \in S} \min \left[\frac{\epsilon^2}{b_k}, \Delta_k \right] \tag{107}$$

converges to a finite limit. Let now p be defined as an integer such that

$$\gamma_2 \gamma_1^{p-1} < 1 \tag{108}$$

and define also

$$S(k) = |S \cap \{0, \ldots, k\}| \tag{109}$$

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

$$J_1 = \{k \mid k \leq pS(k)\} \text{ and } J_2 = \{k \mid k > pS(k)\} \tag{110}$$

We now want to show that both sums

$$\sum_{k \in J_1} \frac{1}{b_k} \text{ and } \sum_{k \in J_2} \frac{1}{b_k} \tag{111}$$

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 that

$$\sum_{k \in J_1} \frac{1}{b_k} \leq \sum_{k \in J_3} \frac{1}{b_k} = p \sum_{k \in S} \frac{1}{b_k} \leq +\infty \tag{112}$$

because of the non-decreasing nature of the sequence $\{b_k\}$ and the convergence of the last sum. We now turn our attention to the second sum in (111). Observe that, for $k \in J_2$,

$$\frac{c_4}{b_k} \leq \Delta_k \leq \gamma_2^{S(k)} \gamma_1^{k-S(k)} \Delta_0 \leq \gamma_2^{k/p} \gamma_1^{k-k/p} \Delta_0 \leq \left(\gamma_2 \gamma_1^{p-1} \right)^{k/p} \Delta_0, \tag{113}$$

where we have used Lemma 6 and the definition of J_2 in (110). This yields that

$$\sum_{k \in J_2} \frac{1}{b_k} \leq \frac{\Delta_0}{c_4} \sum_{k \in J_2} \left(\gamma_2 \gamma_1^{p-1} \right)^{k/p} < +\infty \tag{114}$$

and the second sum is convergent. Therefore the sum

$$\sum_{k=0}^{\infty} \frac{1}{b_k} = \sum_{k \in J_1} \frac{1}{b_k} + \sum_{k \in J_2} \frac{1}{b_k} \tag{115}$$

is finite, which contradicts (AS.3). Hence condition (104) is impossible and (103) is true. □

Directly from this proof, we also obtain the following important corollary.

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

$$\liminf_{k\to\infty} h_k^s = 0. \tag{116}$$

We note that (116) gives an scaled equivalent of (103). We often prefer (116) 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).

We are also able to prove that the "liminf" in (103) can be replaced by a true limit, if we somewhat strengthen our assumptions, as is shown in the next theorem.

We first complete our assumptions on the scaling matrices.

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

$$\|D_k\| \leq \sigma_2. \tag{117}$$

This and (AS.3) clearly implies that the scaling matrices have uniformly bounded condition numbers. Although slightly stronger than the condition in [12], we note that it is rather natural, because it prevents h_k going to zero when the sequence x_k does not approach a critical point. More precisely, we obtain the following result.

Lemma 9 *Assume that (AS.3) and (AS.5) hold. Then*

$$\frac{1}{\sigma_1^2} h_k \leq \|P[x_k - g_k] - x_k\| \leq \sigma_2^2 h_k \tag{118}$$

for all k.

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

$$\frac{1}{\sigma_2^2} |[g_k]_j| \leq |[w_k]_j| \leq \sigma_1^2 |[g_k]_j| \tag{119}$$

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

$$v_j = \begin{cases} u_j & \text{if } [g_k]_j < 0, \\ l_j & \text{if } [g_k]_j > 0, \end{cases} \tag{120}$$

we have also that, for all j,

$$\begin{aligned} |[P[x_k - w_k] - x_k]_j| &= \min(|[w_k]_j|, |[x_k]_j - v_j|) \\ &\leq \min(\sigma_1^2 |[g_k]_j|, |[x_k]_j - v_j|) \\ &\leq \sigma_1^2 |[P[x_k - g_k] - x_k]_j|, \end{aligned} \tag{121}$$

Similarily, for all j,

$$\begin{aligned} |[P[x_k - w_k] - x_k]_j| &\geq \min(\tfrac{1}{\sigma_2^2} |[g_k]_j|, |[x_k]_j - v_j|) \\ &\geq \tfrac{1}{\sigma_2^2} |[P[x_k - g_k] - x_k]_j|. \end{aligned} \tag{122}$$

(118) follows immediately. □

We also require the following condition.

(AS.6)

$$\lim_{k\to\infty} b_k[f(x_k) - f(x_{k+1})] = 0. \tag{123}$$

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 uniformly bounded, as assumed in [12]. Note that the fact that $f(\cdot)$ is bounded below and (AS.4) already imply that

$$\liminf_{k\to\infty} b_k[f(x_k) - f(x_{k+1})] = 0. \tag{124}$$

Indeed, assume that

$$b_k[f(x_k) - f(x_{k+1})] \geq \epsilon \tag{125}$$

for some $\epsilon > 0$. Then

$$\sum_{k=0}^{\infty} \frac{1}{b_k} \leq \frac{1}{\epsilon} \sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})] < +\infty \tag{126}$$

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

Theorem 10 *Assume (AS.1)-(AS.6) hold. Assume also that $\{x_k\}$ is a sequence of iterates generated by the algorithm. Then,*

$$\lim_{k\to\infty} h_k = 0. \tag{127}$$

We prove this theorem by contradiction. Assume therefore that there is an $\epsilon_1 \in (0,1)$ and a subsequence $\{m_i\}$ of successful iterates such that, for all m_i in this subsequence

$$h_{m_i} \geq \frac{\epsilon_1}{\sigma_1}, \tag{128}$$

and thus, by (45),

$$h^s_{m_i} \geq \epsilon_1. \tag{129}$$

Corollary 8 guarantees the existence of another subsequence $\{l_i\}$ such that

$$h^s_k \geq \epsilon_2 \text{ for } m_i \leq k < l_i \text{ and } h^s_{l_i} < \epsilon_2, \tag{130}$$

where we have set

$$\epsilon_2 = \frac{\epsilon_1}{4\sigma_1^4\sigma_2^2} < \epsilon_1. \tag{131}$$

Note that the last inequality in (130), (45) and Lemma 9 imply that

$$\|P[x_{l_i} - g_{l_i}] - x_{l_i}\| \leq \sigma_2^2 h_{l_i} \leq \sigma_1\sigma_2^2\epsilon_2. \tag{132}$$

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

$$K = \{k \mid k \in S \text{ and } m_i \leq k < l_i\}, \tag{133}$$

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

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\mu c_3 \epsilon_2^2 \min[\frac{\epsilon_2^2}{b_k}, \Delta_k]. \tag{134}$$

Because of (123), we have then that

$$\lim_{k \in K} b_k \Delta_k = 0. \tag{135}$$

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

$$\begin{aligned} \|x_{m_i} - x_{l_i}\| &\leq \sum_{k=m_i}^{l_i-1} \|x_{k+1} - x_k\| \\ &\leq \beta_2 \sigma_1 \sum_{k=m_i}^{l_i-1}{}^{(K)} \Delta_k \\ &\leq c_6 \sum_{k=m_i}^{l_i-1}{}^{(K)} [f(x_k) - f(x_{k+1})] \\ &\leq c_6 [f(x_{m_i}) - f(x_{l_i})], \end{aligned} \tag{136}$$

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

$$c_6 \stackrel{\text{def}}{=} \frac{2\beta_2 \sigma_1}{\mu c_3 \epsilon_2^2}. \tag{137}$$

But the last right hand side of relation (136) tends to zero as i tends to infinity, and thus continuity and (132) imply that

$$\|P[x_{m_i} - g_{m_i}] - x_{m_i}\| \leq 2\sigma_1 \sigma_2^2 \epsilon_2 \tag{138}$$

for i sufficiently large. We may now apply Lemma 9 again, and we obtain that

$$h_{m_i} \leq 2\sigma_1^3 \sigma_2^2 \epsilon_2 \leq \frac{\epsilon_1}{2\sigma_1} \tag{139}$$

when i is sufficiently large. This contradicts (128) and proves the theorem. □

As above, we note that (AS.3) and (AS.5) imply a scaled equivalent of (127), i.e.

$$\lim_{k \to \infty} h_k^s = 0. \tag{140}$$

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, then the set of constraints that are active at this critical point is correctly identified in a finite number of iterations.

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 non-empty because $\mathcal{L}$ 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

$$i \in I(x_*) \Rightarrow |(\nabla f(x_*))_i| > 0 \tag{141}$$

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

We first 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 11 *Assume that (AS.1)-(AS.4) and (AS.7) hold. Then L is finite and*

$$I(x_k) \subseteq I(x_{k+m}) \tag{142}$$

for all $m \geq 0$ and k sufficiently large.

After a finite number of iterations, every iterate lies in the neighbourhood of a limit point. More precisely, we can choose a $\delta > 0$ such that, given an iterate x_k with k sufficiently large, there exists a limit point $x_* \in L$ such that

$$x_k \in \mathcal{N}(x_*, \delta) \stackrel{\text{def}}{=} \{x \in \mathcal{L} \mid \|x - x_*\| \leq \delta\}. \tag{143}$$

Using continuity, compactness of L and (AS.7), we can also assume, without loss of generality, that δ is small enough to ensure that, for all $x_* \in L$, all $x \in \mathcal{N}(x_*, \delta)$ and all $j \in I(x)$,

$$j \in I(x_*) \tag{144}$$

$$\text{sgn}([\nabla f(x)]_j) = \text{sgn}([\nabla f(x_*)]_j) \text{ and } |[\nabla f(x)]_j| \geq \frac{1}{2}|[\nabla f(x_*)]_j|. \tag{145}$$

Consider now such a large k. 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, the mechanism of the algorithm ensures that $I(x_k^C) \subseteq I(x_{k+1})$. But (144) and (145) 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 (142) is proved. □

We now show that, for every convergent subsequence, the correct active set is identified by the algorithm after a finite number of iterations.

Theorem 12 *Assume (AS.1)-(AS.4) and (AS.7) hold. Furthermore, assume that*

$$\lim_{k \to \infty} h_k = 0. \tag{146}$$

Then, if $\{x_{k_i}\}$ is a subsequence converging to a limit point $x_* \in L$, x_* is critical and

$$I(x_{k_i}) = I(x_*) \tag{147}$$

for i sufficiently large.

As a consequence, if the complete sequence $\{x_k\}$ converges to a single limit point, then this point is critical and its active set is correctly identified by the algorithm after finitely many iterations.

The criticality of x_* follows from (146) and Lemma 1.

Let $I_* = \limsup I(x_{k_i})$. Using Lemma 11,

$$I(x_{k_i}) \subseteq I(x_*), \tag{148}$$

for all i sufficiently large. Hence $I(x_{k_i}) = I_*$ for i sufficiently large. Suppose $I_* \subset I(x_*)$. Then there exists a non-empty set $J \subseteq I(x_*)$ such that $J \cap I_* = \emptyset$. However, from the strict complementarity slackness in (AS.7) and (145), we may deduce that

$$[h^s_{k_i}]^2 \geq \frac{1}{\sigma_2^2} \sum_{j \in J} [g_{k_i}]_j^2 \geq \frac{1}{4\sigma_2^2} |J| \min_{j \in J} |[\nabla f(x_*)]_j^2| \geq \epsilon \tag{149}$$

for some $\epsilon > 0$. Hence $h_{k_i} \geq \epsilon$, for all k large enough. This contradicts (146). □

This last result is important because it shows that 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 11 and Theorem 12 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 subsection, 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.

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

Theorem 13 *Assume that (AS.1)-(AS.7) hold, and assume that $\{x_{k_i}\}$ is a subsequence of iterates, generated by the algorithm, converging to the critical point x_*. Assume also that there is an $\epsilon > 0$ such that*

$$\liminf_{i \to \infty} \lambda^1[B_{k_i}] \geq \epsilon. \tag{150}$$

Assume finally that $\nabla^2 f(x_*)$ is non-singular 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.

The criticality of x_* is ensured by theorem 10. We first choose a $\delta > 0$ and a k_1 sufficiently large to ensure (143), and the two conditions thereafter, for all $k \geq k_1$, and that the minimum singular value of $\nabla^2 f(x)$ restricted to the subspace $C(x_*)$ is larger than some constant $c_7 > 0$. We also apply theorem 12 to the subsequence $\{x_{k_i}\}$ and deduce that

$$I(x_{k_i}) = I(x_*) \tag{151}$$

for all i sufficiently large. We can then choose i_1 large enough so that $k_{i_1} \geq k_1$,

$$\lambda^1[B_{k_i}] \geq \frac{1}{2}\epsilon, \tag{152}$$

$$\|x_{k_i} - x_*\| \leq \frac{\epsilon\delta}{4c_7 + \epsilon} \stackrel{\text{def}}{=} \delta_1 \tag{153}$$

and (151) hold for all $i \geq i_1$, and also so that

$$h_k \leq c_7\delta_1 \tag{154}$$

for all $k \geq k_{i_1}$. This last relation has to hold for large enough k because of theorem 10. For all $i \geq i_1$ we now observe that, using (151)

$$s_{k_i} \in C(x_*) \tag{155}$$

and we decompose g_{k_i} as

$$g_{k_i} = g^R_{k_i} + g^N_{k_i} \text{ with } g^N_{k_i} \in C(x_*) \text{ and } g^R_{k_i} \in C(x_*)^\perp. \tag{156}$$

The vector $g^N_{k_i}$ is thus the gradient of the model at x_{k_i} projected onto $C(x_*)$, and has the property that

$$\|g^N_{k_i}\| = h_{k_i} \leq c_7\delta_1 \tag{157}$$

for $i \geq i_1$. Consider now the one dimensional strictly convex quadratic function of the parameter τ defined by

$$\phi(\tau) \stackrel{\text{def}}{=} m_{k_i}(x_{k_i} + \tau s_{k_i}) - f(x_{k_i}). \tag{158}$$

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

$$\tau_* \stackrel{\text{def}}{=} \arg\min \phi(\tau) \geq \frac{1}{2}. \tag{159}$$

But an easy computation shows that

$$\tau_* = \frac{|g^T_{k_i} s_{k_i}|}{s^T_{k_i} B_{k_i} s_{k_i}} \leq \frac{2\|g^N_{k_i}\|}{\epsilon\|s_{k_i}\|}, \tag{160}$$

where we have used the Cauchy-Schwartz inequality, (152) and (155) to derive the last part of this bound. Therefore, we can deduce that

$$\|s_{k_i}\| \leq \frac{4}{\epsilon}\|g^N_{k_i}\|. \tag{161}$$

Hence, gathering (153), (157) and (161), we obtain that

$$\|x_{k_i+1} - x_*\| \leq \|s_{k_i}\| + \|x_{k_i} - x_*\| \leq \left[\frac{4c_7}{\epsilon} + 1\right]\delta_1 \leq \delta. \tag{162}$$

Assume now that

$$\|x_{k_i+1} - x_*\| > \delta_1. \tag{163}$$

Then, using (151), (162) and our assumptions on $\nabla^2 f(x_*)$ and δ, we have that

$$h_{k_i+1} = \|g^N_{k_i+1}\| = \|\nabla^2 f(u)(x_{k_i+1} - x_*)\| > c_7\delta_1, \tag{164}$$

where u is a point in the segment (x_{k_i+1}, x_*). This is impossible because of (154), and therefore

$$\|x_{k_i+1} - x_*\| \leq \delta_1. \tag{165}$$

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 $j \geq 1$.

$$\|x_{k_i+j} - x_*\| \leq \delta_1 \leq \delta. \tag{166}$$

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

$$f(x_k) - m_k(x_k + s_k) \geq \beta_1[f(x_k) - \min m_k(x_k^C + p_k)], \tag{167}$$

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

(AS.9)

$$\beta_2 > \nu. \tag{168}$$

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 two following conditions hold.

(AS.10) The matrices B_k satisfy the conditions

$$\lambda^1[B_k] \leq c_8\lambda^1[\nabla^2 f(x_k)] \text{ when } \lambda^1[\nabla^2 f(x_k)] < 0, \tag{169}$$

where c_8 is some positive constant, and

$$\lim_{k\to\infty} |r(B_k, s_k) - r(\nabla^2 f(x_k), s_k)| = 0 \text{ whenever } \lim_{k\to\infty} \|s_k\| = 0, \tag{170}$$

where $r(X,s) = s^T Xs/\|s\|^2$ 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 14 *Assume that (AS.1)-(AS.3), (AS.5) and (AS.8) hold. Assume furthermore that*

$$\lambda^1[B_k] < 0 \tag{171}$$

for some k. Then,

$$f(x_k) - m_k(x_k + s_k) \geq -\frac{1}{2} c_9 \lambda^1[B_k] \Delta_k^2, \tag{172}$$

where the constant c_9 is defined by

$$c_9 \stackrel{\rm def}{=} \beta_1 \left(\frac{\beta_2 - \nu}{\sigma_2} \right)^2. \tag{173}$$

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

$$p_k^T \left(g_k + B_k(x_k^C - x_k) \right) \leq 0 \tag{174}$$

and

$$\|D_k(x_k^C + p_k - x_k)\| = \beta_2 \Delta_k. \tag{175}$$

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

$$m_k(x_k^P) = m_k(x_k^C) + p_k^T \left(g_k + B_k(x_k^C - x_k) \right) + \frac{1}{2} p_k^T B_k p_k \leq f(x_k) + \frac{1}{2} \lambda^1[B_k] \|p_k\|^2, \tag{176}$$

where we used the inequality $m_k(x_k^C) < f(x_k)$, (174) and the definition of p_k. Observe now that

$$\frac{\|D_k p_k\|}{\|D_k(x_k^C - x_k)\|} \geq \frac{\|D_k(x_k^P - x_k)\|}{\|D_k(x_k^C - x_k)\|} - 1 \geq \frac{\beta_2 - \nu}{\nu}, \tag{177}$$

because of (10) and (175), and hence

$$\beta_2 \Delta_k = \|D_k(x_k^P - x_k)\| \leq \|D_k p_k\| + \|D_k(x_k^C - x_k)\| \leq \left(1 + \frac{\nu}{\beta_2 - \nu} \right) \sigma_2 \|p_k\|. \tag{178}$$

This last inequality and relation (176) together then imply that

$$f(x_k) - m_k(x_k^P) \geq -\frac{1}{2} \lambda^1[B_k] \left(\frac{\beta_2 - \nu}{\sigma_2} \right)^2 \Delta_k^2. \tag{179}$$

To complete the proof, one only needs to notice that

$$f(x_k) - m_k(x_k + s_k) \geq \beta_1 [f(x_k) - m_k(x_k^P)] \tag{180}$$

because of (167). □

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

Theorem 15 *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 $C(x_*)$.*

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 $-2\epsilon_1$, 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 sucessful iterations such that

$$\Delta_{m_i} \geq \epsilon_2 \tag{181}$$

for some $\epsilon_2 > 0$. 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 whole sequence $\{x_{m_i}\}$ converges to x_*. We will also assume that i is large enough to ensure that

$$I(x_{m_i}) \subseteq I(x_*) \text{ and } \lambda^1[\nabla^2 f(x_{m_i})] \leq -\epsilon_1 \tag{182}$$

by using the continuity of the Hessian. Hence (169) implies that

$$\lambda^1[B_{m_i}] \leq -c_8\epsilon_1 \tag{183}$$

for i sufficiently large. Using this bound, lemma 14, the fact that iteration m_i is successful and (181), we deduce that

$$f(x_{m_i}) - f(x_{m_i+1}) \geq \frac{1}{2}\mu c_8 c_9 \epsilon_1 \epsilon_2^2 \tag{184}$$

for large i. But this inequality is clearly impossible because

$$\sum_{i=0}^{\infty}[f(x_{m_i}) - f(x_{m_i+1})] \leq \sum_{k\in S}[f(x_k) - f(x_{k+1})] \leq f(x_0) - f(x_*) < +\infty. \tag{185}$$

Hence no subsequence of successful iterations is such that (181) holds, and

$$\lim_{k\in S} \Delta_k = 0. \tag{186}$$

This, in turn, implies that

$$\lim_{k\to\infty} \Delta_k = 0 \text{ and } \lim_{k\to\infty} \|s_k\| = 0, \tag{187}$$

by using (13), (23) and (25). We note that, for k sufficiently large, the point x_k is close enough to a limit point to ensure that

$$\lambda^1[\nabla^2 f(x_k)] \leq -\epsilon_1 \text{ and hence } \lambda^1[B_k] \leq -c_8\epsilon_1, \tag{188}$$

because of (169). Combining now (92), the bound $\|s_k\| \leq \sigma_1\beta_2\Delta_k$ and (172) together, we obtain that

$$|\rho_k - 1| \leq \frac{\sigma_1^2\beta_2^2}{c_8 c_9 \epsilon_1}|r(B_k, s_k) - r(\nabla^2 f(x_k), s_k)| \tag{189}$$

and the right hand side of this expression tends to zero because of the second part of (187) and (170). The updating rules for the trust region radius then prevent Δ_k from tending to zero. This contradicts the first part of (187) however, and 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, nor that it was not a saddle point. It is nevertheless often the case that x_* is a local minimizer, and criticality can be guaranteed by imposing (AS.6), as shown by theorem 10.

Observe finally 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.

Firstly, 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.

At variance with [8], for example, the implementation of our algorithm does not require extensive linear algebra, and what is needed can be accomplished by using efficient iterative methods such as preconditioned conjugate gradients. Therefore, the framework presented here is quite well suited to large dimensional 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 been already 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 may therefore 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.

Bibliography

[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", Journal of Optimization Theory and Applications, vol. 29, # 4, pp. 521-558, 1979.

[3] R.H. Byrd, R.B. Schnabel and G.A. Schultz, "A trust region algorithm for nonlinearly constrained optimization", Department of Computer Sciences Technical Report CU-CS-313-85, University of Colorado at Boulder, 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.), pp.71-82, 1985.

[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", (in preparation).

[6] J.E. Dennis and R.B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations", Prentice-Hall, Englewood Cliffs, 1983.

[7] D.M. Gay, "Computing optimal locally constrained steps", SIAM Journal on Statistical and Scientific Computing, vol. 2, pp. 186-197, 1981.

[8] D.M. Gay, "A trust region approach to linearly constrained optimization", In "Numerical Analysis : Proceedings Dundee 1983" (D.F. Griffiths, ed.), pp. 203-220, Lecture Notes in Mathematics 1066, Springer Verlag, Berlin, 1984.

[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", Numerische Mathematik, vol. 39, pp. 119-137, 1982.

[11] A. Griewank and Ph.L. Toint, "Numerical experiments with partially separable optimization problems", In "Numerical Analysis : Proceedings Dundee 1983" (D.F. Griffiths, ed.), pp. 203-220, Lecture Notes in Mathematics 1066, Springer Verlag, Berlin, 1984.

[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.), pp. 258-287, Springer Verlag, Berlin, 1983.

[13] M.J.D. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches", SIAM-AMS Proc. 9, pp. 53-72, 1976.

[14] M.J.D. Powell, "On the global convergence of trust region algorithms for unconstrained minimization", Mathematical Programming, vol. 29, # 3, pp.297-303, 1984.

[15] M.J.D. Powell and Y. Yuan, "A trust region algorithm for equality constrained optimization", Department of Applied Mathematics and Theoretical Physics, Report DAMTP1986-NA2, University of Cambridge, Cambridge, 1986.

[16] A. Vardi, "A trust region algorithm for equality constrained minimization: convergence properties and implementation" SIAM Journal on Numerical Analysis, vol. 22, # 3, pp. 575-591, 1985.