Mathematical Programming 43 (1989) 235-256
North-Holland

NONLINEAR PROGRAMMING AND NONSMOOTH OPTIMIZATION BY SUCCESSIVE LINEAR PROGRAMMING

R. FLETCHER
Department of Mathematical Sciences, University of Dundee, Scotland

E. SAINZ de la MAZA
Department of Applied Mathematics, Universidad del Pais Vasco, Bilbao, Spain

Received 6 May 1987
Revised 1 December 1987

Methods are considered for solving nonlinear programming problems using an exact l_1 penalty function. LP-like subproblems incorporating a trust region constraint are solved successively both to estimate the active set and to provide a foundation for proving global convergence. In one particular method, second order information is represented by approximating the reduced Hessian matrix, and Coleman-Conn steps are taken. A criterion for accepting these steps is given which enables the superlinear convergence properties of the Coleman-Conn method to be retained whilst preserving global convergence and avoiding the Maratos effect. The methods generalize to solve a wide range of composite nonsmooth optimization problems and the theory is presented in this general setting. A range of numerical experiments on small test problems is described.

Key words: Nonlinear programming, nonsmooth optimization, global convergence, superlinear convergence, trust region method, Coleman-Conn method.

1. Introduction

The primary motivation for this work is to consider methods for finding a local solution, x^* say, to a nonlinear programming (NLP) problem with equality and inequality constraints:

$$\begin{aligned} &\text{minimize} && f(x), \quad x \in \mathbb{R}^n \\ &\text{subject to} && c_i(x) = 0, \quad i \in E, \\ & && c_i(x) \leqslant 0, \quad i \in I. \end{aligned} \tag{1.1}$$

The methods were developed with the aim of including the following features which contribute either to reliability or efficiency. The methods are first derivative methods so that we avoid the inconvenience associated with deriving expressions for second derivatives and the overheads of storing and manipulating them. The methods solve LP-like subproblems on every iteration: this avoids the extra complexity of solving QP-like subproblems (as in the SQP method) and the need to manipulate a full $(n \times n)$ Hessian matrix. The LP-like subproblem includes a trust region constraint:

together with the use of an l_1 exact penalty function this enables global convergence to be proved under very mild conditions. The trust region constraint also enables the advantage of QP-like methods to be retained in which an accurate estimate of the active constraints at the solution is obtained. Second order information is represented in the form of a reduced Hessian matrix, the dimension of which is $(n-t)\times(n-t)$ where t is the number of active constraints. Together with the feature of solving LP-like subproblems, this potentially allows very large problems to be solved, as long as the dimension of the reduced space (i.e. $n-t$) is not too large. Rapid local convergence when second order effects are important is obtained by using any method which uses a reduced Hessian matrix. In this paper we concentrate on the method of Coleman and Conn (1982a, b) which we regard as most suitable, although we have studied other possibilities (Sainz de la Maza, 1987). An important part of this paper is the way in which the Coleman–Conn step is interfaced with the technique of solving an LP-like subproblem. The latter provides an estimate of the active constraints and a trust region bound is included so that only locally active constraints are located. However the step defined by the LP-like subproblem is only used if the Coleman–Conn step fails and second order information is inadequate. A criterion for accepting a Coleman–Conn step is devised which is related to the reduction predicted by the LP-like subproblem. This enables Coleman–Conn steps to be accepted asymptotically whilst preserving global convergence properties, and so avoids the Maratos effect.

In Section 2 the LP-like subproblem is specified and the criterion for a sufficient reduction in the penalty function is derived. A prototype algorithm is given which contains the basic features that are sufficient for convergence whilst allowing refinements to improve practical performance. Global convergence of the prototype algorithm is proved under very general conditions, in particular not requiring any linear independence assumptions. Convergence of the multiplier estimates is proved, and some properties of the estimate of the active set are derived. In Section 3 the Coleman–Conn method is described in a general way which allows the null space matrix to be calculated by a variety of schemes (generalized elimination). Subject to certain standard conditions and suitable asymptotic behaviour of the approximate reduced Hessian matrix, an expression is obtained for the actual reduction in the penalty function given by the Coleman–Conn step. This is shown to be asymptotically equal to the predicted reduction of a certain QP-like problem. This result is then used to show asymptotically that the Coleman–Conn step satisfies the sufficient reduction criterion derived in Section 2, and that the correct active set is determined. Hence the method avoids the Maratos effect and inherits the local convergence properties of the Coleman–Conn method (Byrd, 1984). In Section 4 a range of numerical experiments with a pilot code on small NLP problems is described. This suggests that despite solving only LP-like subproblems and using only reduced Hessian information, the method is nonetheless comparable with various other types of method that are currently attracting attention. Some suggestions for refining the algorithm are also presented.

The methods are based on using the l_1 exact penalty function

$$\phi(x) = f(x) + \sum_{i \in E} |c_i(x)| + \sum_{i \in I} \max(c_i(x), 0). \tag{1.2}$$

It is well known that if $f(x)$ is scaled so that the Lagrange mulipliers λ^* for (1.1) satisfy $|\lambda_i^*| < 1$ then under mild conditions x^* is also a local minimizer of this function. It happens that the type of method that we have developed is more widely applicable to a range of composite nonsmooth optimization (CNSO) problems, having the form

$$\text{minimize} \quad \phi(x) \triangleq f(x) + h(c(x)) \tag{1.3}$$

in which $f(x)$ $(\mathbb{R}^n \to \mathbb{R})$ and $c(x)$ $(\mathbb{R}^n \to \mathbb{R}^m)$ are smooth $(\mathbb{C}^2)$ and $h(c)$ $(\mathbb{R}^m \to \mathbb{R})$ is a polyhedral convex function

$$h(c) = \max_i h_i^{\mathrm{T}} c + \beta_i. \tag{1.4}$$

This includes not only l_1 and l_∞ penalty functions but also minimax problems and problems in l_1 and l_∞ approximation. In all these applications $\beta_i = 0$ for all i, and in each case an appropriate set of vectors h_i can readily be determined. (A review of this material is given by Fletcher, 1981). Therefore the paper is presented in this wider setting whilst attempting to relate to the NLP problem at a number of places in the text.

First order necessary conditions for x^* to solve (1.3) are that there exists a vector of multipliers $\lambda^* \in \partial h(c^*)$ such that

$$g^* + A^* \lambda^* = 0 \tag{1.5}$$

where $g(x) \triangleq \nabla f(x)$, $A(x) \triangleq \nabla c^{\mathrm{T}}(x)$ and where c^*, $g^*, \ldots$ denote $c(x^*)$, $g(x^*), \ldots$, etc. For convenience these conditions are referred to as KT conditions and x^* as a KT point in what follows. For a polyhedral convex function (1.4) it can be shown (e.g. Fletcher, 1981) that the subdifferential set is given by

$$\partial h(c) = \operatorname*{conv}_{i \in \mathscr{A}(c)} h_i \tag{1.6}$$

where $\mathscr{A}(c) = \{i: h_i^{\mathrm{T}} c + \beta_i = h(c)\}$ is the set of indices at which the max is attained. When (1.3) refers to the l_1 exact penalty function (1.2) then (1.6) can be expressed as

$$\begin{aligned} \partial h(c) = \{\lambda: -1 \leqslant \lambda_i \leqslant 1 \quad &\text{if} \quad c_i = 0 \quad \text{and} \quad i \in E, \\ 0 \leqslant \lambda_i \leqslant 1 \quad &\text{if} \quad c_i = 0 \quad \text{and} \quad i \in I, \\ \lambda_i = \operatorname{sign} c_i \quad &\text{otherwise}\}. \end{aligned} \tag{1.7}$$

It is also useful to have an alternative representation of the set $\partial h(c)$ in the following form. Let λ_0 be an arbitrary vector in $\partial h(c)$. If the dimension of $\partial h(c)$ is t, then there exists an $n \times t$ matrix D whose columns $d_1, d_2, \ldots, d_t \in \mathbb{R}^n$ are basis vectors for the set $\partial h(c) - \lambda_0$. That is to say, $\partial h(c)$ can be expressed as

$$\partial h(c) = \{\lambda: \lambda = \lambda_0 + Du, u \in U\} \tag{1.8}$$

where $U \in \mathbb{R}^t$ is some set which exists. This is essentially the idea contained in Osborne's (1985) concept of *structure functionals* for polyhedral convex functions, taken up more generally by Womersley (1984). For the l_1 exact penalty function it follows from (1.7) that the most natural choice for λ_0 is given by

$$(\lambda_0)_i = \begin{cases} 0 & \text{if } c_i = 0, \\ \text{sign } c_i & \text{otherwise.} \end{cases} \tag{1.9}$$

Also the columns of D are simply the coordinate vectors e_i for indices i of active l_1 terms ($c_i = 0$) in the penalty function. Similar expressions hold in other common cases, e.g. Womersley (1984), Fletcher (1987). For example in a minimax problem, the columns of D are the vectors $e_i - e_q$, $i \in \mathscr{A} \setminus q$ where $q \in \mathscr{A}$ is arbitrary.

When considering the set $\partial h(c^*)$ which arises in (1.5), it is often convenient to choose λ^* as the arbitrary vector in place of λ_0. This merely translates the set U. Thus we may define

$$\partial h(c^*) = \partial h^* = \{\lambda : \lambda = \lambda^* + D^* u,\ u \in U^*\}. \tag{1.10}$$

In addition, *strict complementarity* is said to hold if λ^* is in the relative interior of ∂h^*, or equivalently if $0 \in \text{int } U^*$. In the case of the l_1 exact penalty function, this requires that λ^* *strictly* satisfies the inequalities in (1.7) (for $c = c^*$), and is analogous to the usual meaning of the term.

A further significance of the matrix D^* arises by considering the NLP problem

$$\begin{aligned} &\text{minimize} && f(x) + c(x)^{\mathrm{T}} \lambda^* \\ &\text{subject to} && D^{*\mathrm{T}}(c(x) - c^*) = 0. \end{aligned} \tag{1.11}$$

This problem is locally equivalent to the CNSO problem (1.3) (in the sense that x^* minimizes (1.3) iff x^* solves (1.11)) under the mild assumptions of second order sufficiency and strict complementarity. Moreover $u^* = 0$ is the KT multiplier vector for (1.11). A precise statement of this result is given by Fletcher (1987). In fact if the definition of $h(c)$ in (1.4) is restricted to have $\beta_i = 0$ for all i, then a more simple equivalent problem

$$\begin{aligned} &\text{minimize} && f(x) + c(x)^{\mathrm{T}} \lambda^* \\ &\text{subject to} && D^{*\mathrm{T}} c(x) = 0 \end{aligned} \tag{1.12}$$

is obtained. This result follows directly from Lemma 1.1 given at the end of this section. For reasons of clarity this form of the equivalent problem is used in the rest of the paper. Since $\beta_i = 0$ in all common cases, little is lost by imposing this condition, although the theory can be worked through in the more general case.

In the case of the l_1 exact penalty function, (1.12) can be rearranged to give the problem

$$\begin{aligned} &\text{minimize} && f(x) + c(x)^{\mathrm{T}} \lambda_0 \\ &\text{subject to} && c_i(x) = 0 \quad \forall i:\ c_i(x^*) = 0 \end{aligned} \tag{1.13}$$

(λ_0 is calculated using c^* in (1.9)) and the local equivalence to (1.2) is well known. Thus the constraint $D^{*\mathrm{T}} c(x) = 0$ in (1.12) simply picks out the active constraints

from the vector $c(x)$, and these are in fact Osborne's structure functionals in this case. Likewise columns of the matrix A^*D^* are the constraint gradients for the active constraints at x^*. Also the term $c(x)^{\mathrm{T}}\lambda^*$ in (1.12) has the effect of inserting an l_1 penalty for each violated constraint into the objective function.

Our methods involve the solution of LP-like subproblems which provide a current basis $D^{(k)}$ and multiplier estimate λ^k. We can therefore define the current estimate of the equivalent NLP problem as

$$\begin{aligned} &\text{minimize} && f(x)+\lambda^{(k)\mathrm{T}}c(x) \\ &\text{subject to} && D^{(k)\mathrm{T}}c(x)=0 \end{aligned} \tag{1.14}$$

in which the constraints can be regarded as the current estimate of the set of active constraints. In fact the use of $\lambda^{(k)}$ can be varied in practical computation without changing the minimizer of this problem. We return to this point in Section 3 (after eq. (3.8)).

Finally we state two lemmas giving some properties of the set $\partial h(c)$. If the reader wishes to restrict attention to the case of the l_1 exact penalty function then these results are easily verified from (1.7).

Lemma 1.1. *Let $h(c)$ be a polyhedral convex function* (1.4) *in which $\beta_i = 0$ for all i. If $\lambda' \in \partial h(c')$ for some vector c', and D' is a basis for $\partial h(c') - \lambda'$, then both*

$$\lambda^{\mathrm{T}}c' = h(c') \quad \forall \lambda \in \partial h(c') \tag{1.15}$$

and

$$D'^{\mathrm{T}}c' = 0. \tag{1.16}$$

Proof. If $\beta_i = 0$ for all i, then a consequence of (1.4) is that $h(\alpha c) = \alpha h(c)$ for all $\alpha \geq 0$. It follows from the subgradient inequality

$$h(c) \geq h(c') + \max_{\lambda \in \partial h(c')} \lambda^{\mathrm{T}}(c - c') \tag{1.17}$$

that

$$h(c') + (\alpha - 1)\lambda^{\mathrm{T}}c' \leq h(\alpha c') = \alpha h(c')$$

for all $\lambda \in \partial h(c')$, and taking $\alpha = 1 \pm \varepsilon$ gives $\lambda^{\mathrm{T}}c' = h(c')$. Since $\lambda' \in \partial h(c')$ it follows that $\lambda'^{\mathrm{T}}c' = h(c')$ and hence $(\lambda - \lambda')^{\mathrm{T}}c' = 0$. Defining D' as in (1.8) with $\lambda_0 = \lambda'$ as the arbitrary vector, it can be deduced that

$$u^{\mathrm{T}}D'^{\mathrm{T}}c' = 0$$

for all $u \in U' \subset \mathbb{R}^{t'}$. Since the dimension of U' is t', it follows that $D'^{\mathrm{T}}c' = 0$. □

Lemma 1.2. *For $h(c)$ as in Lemma* 1.1, *if c is sufficiently close to c' then*

$$\partial h(c) \subset \partial h(c'). \tag{1.18}$$

In addition if $D'^{\mathrm{T}}c = 0$ then

$$\partial h(c) = \partial h(c'). \tag{1.19}$$

Proof. See Fletcher (1987), Lemmas 14.4.1 and 14.4.2. □

2. Global properties

In this section a prototype algorithm is described which provides global convergence properties for a variety of possible methods. The main feature of this algorithm is that it is a trust region algorithm based on solving LP-like subproblems. This enables asymptotically accurate estimates of multipliers and active constraints to be made. However the correction determined by the subproblem is usually not used directly because of the possibility of slow local convergence on certain types of problem. Rather a Newton-like step is taken which allows superlinear convergence to occur. Only if this step is unsuccessful does the algorithm use the step determined by the subproblem. The interrelation of these different types of step is such that usually the trust region radius ($\rho^{(k)}$ below) does not shrink to zero and so does not affect the superlinear convergence of the Newton-like step.

First of all the *linearized subproblem* at a current point $x^{(k)}$ is defined as

$$\begin{aligned} &\underset{\delta}{\text{minimize}} \quad && l^{(k)}(\delta) \triangleq f^{(k)} + g^{(k)\mathrm{T}}\delta + h(c^{(k)} + A^{(k)\mathrm{T}}\delta) \\ &\text{subject to} \quad && \|\delta\| \leqslant \rho^{(k)} \end{aligned} \tag{2.1}$$

where $c^{(k)}$ denotes $c(x^{(k)})$ etc. The condition $\|\delta\| \leqslant \rho^{(k)}$ is the *trust region constraint* and here we use the l_∞ norm for convenience. In this case, if $h(c)$ is a polyhedral function, (2.1) can be transformed to give an LP calculation, although in practice a more efficient method of solution might exist, as in the case of an l_1 exact penalty function. Denote the solution of (2.1) by $\bar{\delta}^{(k)}$ and let $\bar{c}^{(k)} = c^{(k)} + A^{(k)\mathrm{T}}\bar{\delta}^{(k)}$. First order conditions (e.g. Fletcher, 1987) are that there exist multipliers $\lambda^{(k)} \in \partial h(\bar{c}^{(k)})$, $w^{(k)} \in \partial\|\bar{\delta}^{(k)}\|$ and $\pi^{(k)} \geqslant 0$ such that

$$g^{(k)} + A^{(k)}\lambda^{(k)} + \pi^{(k)} w^{(k)} = 0 \tag{2.2}$$

and

$$\pi^{(k)}(\|\bar{\delta}^{(k)}\| - \rho^{(k)}) = 0. \tag{2.3}$$

Again these are referred to as KT conditions in what follows. The *linearized reduction* given by $\bar{\delta}^{(k)}$ can be defined by

$$\Delta l^{(k)} = l^{(k)}(0) - l^{(k)}(\bar{\delta}^{(k)}) = \phi^{(k)} - l^{(k)}(\bar{\delta}^{(k)}).$$

where $\phi^{(k)}$ denotes $\phi(x^{(k)})$. Also some suitable basis $D^{(k)}$ for $\partial h(\bar{c}^{(k)}) - \lambda^{(k)}$ is chosen, and a set $U^{(k)}$ exists in a similar way to (1.10). The matrix $D^{(k)}$ provides the current estimate of the set of active constraints, and it follows from Lemma 1.1 that

$$D^{(k)\mathrm{T}}\bar{c}^{(k)} = 0. \tag{2.4}$$

In the prototype algorithm we wish to allow the actual correction $\delta^{(k)}$ to be determined by a Newton-like step for (1.14), in which case the correction $\bar{\delta}^{(k)}$ is not used. However the linearized reduction $\Delta l^{(k)}$ is used to determine a criterion which measures whether any step gives a sufficient reduction in $\phi(x)$. This criterion also makes use of any second-order information that is available. Let the matrix $B^{(k)}$ be a positive semi-definite approximation to the Hessian of the Lagrangian function, from which

$$b^{(k)} = \bar{\delta}^{(k)\mathrm{T}} B^{(k)} \bar{\delta}^{(k)} \geq 0 \tag{2.5}$$

can be calculated. Define

$$q^{(k)}(\delta) \triangleq l^{(k)}(\delta) + \tfrac{1}{2}\delta^{\mathrm{T}} B^{(k)} \delta \tag{2.6}$$

and denote the predicted reduction for this quadratic model by

$$\Delta q^{(k)} = q^{(k)}(0) - \min_{\delta} q^{(k)}(\delta) \tag{2.7}$$

(allowing $\Delta q^{(k)} = \infty$ if $q^{(k)}(\delta)$ is unbounded below.) We do not calculate $\Delta q^{(k)}$ directly but a bound can be obtained by using the following lemma.

Lemma 2.1. *Let $l(\alpha) \in \mathbb{C}^0$ $(\mathbb{R} \to \mathbb{R})$ be convex and be minimized in $[0, 1]$ by $\alpha = 1$, with $l(0) > l(1)$. Define $q(\alpha) = l(\alpha) + \frac{1}{2}b\alpha^2$ where $b \geq 0$. Let $q_{\min}$ be the minimum value of $q(\alpha)$ in $[0, 1]$, and denote $\Delta l = l(0) - l(1) > 0$. Then*

$$q(0) - q_{\min} \geq \tfrac{1}{2}\Delta l \min(\Delta l / b, 1).$$

Proof. Consider the chord $c(\alpha) = (1-\alpha)l(0) + \alpha l(1)$. Let α' minimize $c(\alpha) + \frac{1}{2}b\alpha^2$ in $[0, 1]$. Clearly $\alpha' > 0$. If $\alpha' < 1$ then $\alpha' = \Delta l / b$ and hence

$$c(0) - (c(\alpha') + \tfrac{1}{2}b\alpha'^2) = \tfrac{1}{2}\Delta l^2 / b.$$

If $\alpha' = 1$ then

$$c(0) - (c(1) + \tfrac{1}{2}b) = \Delta l - \tfrac{1}{2}b = \Delta l(1 - \tfrac{1}{2}b/\Delta l) \geq \tfrac{1}{2}\Delta l$$

since $\Delta l / b \geq 1$ by definition of α'. But $q(\alpha') \leq c(\alpha') + \frac{1}{2}b\alpha'^2$ by convexity of l and use of the chord, so the lemma follows by definition of $q_{\min} \leq q(\alpha')$ and the above results. □

Now if we set $\delta = \alpha\bar{\delta}^{(k)}$ in (2.6), where $\alpha \in [0, 1]$, then it follows from Lemma 2.1 that

$$\Delta q^{(k)} \geq \tfrac{1}{2}\Delta l^{(k)} \min(\Delta l^{(k)} / b^{(k)}, 1). \tag{2.8}$$

In our algorithm we assess the value of any correction $\delta^{(k)}$ by first computing the *actual reduction* defined by

$$\Delta\phi^{(k)} = \phi^{(k)} - \phi(x^{(k)} + \delta^{(k)}). \tag{2.9}$$

Under certain conditions on $B^{(k)}$ we can expect that $\Delta\phi^{(k)}/\Delta q^{(k)} \to 1$. Together with (2.8), this suggests using a sufficient reduction criterion of the form

$$\Delta\phi^{(k)} \geq \theta \Delta l^{(k)} \min(\Delta l^{(k)}/b^{(k)}, 1), \tag{2.10}$$

where $\theta \in (0, \frac{1}{2})$ is a fixed preset parameter.

It is now possible to describe the trust region algorithm which is used as the prototype for the methods in this paper. On iteration k, the aim is first to try a Newton-like correction, $\delta^{(k)} = d^{(k)}$ say, derived from the NLP problem (1.14), possibly with the addition of a line search. However, if these steps fail to give a sufficient reduction in the sense of (2.10), then the step $\delta^{(k)} = \bar{\delta}^{(k)}$ given by the linearized subproblem is tried. The algorithm changes the trust region radius in a fairly standard way, reducing it when (2.10) is not satisfied and allowing an increase when (2.10) is satisfied and $\|\delta^{(k)}\| \geq \rho^{(k)}$. However it is not necessary for the Newton-like step to lie within the trust region. The algorithm is initialized with values of $x^{(i)}$ and $\rho^{(i)}$, and terminates if $x^{(k)}$ is a KT point for $\phi(x)$. Iteration k of the algorithm is

(i) solve the linearized subproblem determined by $x^{(k)}$ and $\rho^{(k)}$ giving $\bar{\delta}^{(k)}$, $\lambda^{(k)}$, $D^{(k)}$ and $\Delta l^{(k)}$;

(iia) evaluate $\phi(x^{(k)} + d^{(k)})$:
if $\phi^{(k)} - \phi(x^{(k)} + d^{(k)}) \geq \theta \Delta l^{(k)} \min(\Delta l^{(k)}/b^{(k)}, 1)$
then set $\delta^{(k)} = d^{(k)}$ and omit step (iib);

(iib) set $\delta^{(k)} = \bar{\delta}^{(k)}$ and evaluate $\phi(x^{(k)} + \delta^{(k)})$;

(iii) **if** $\Delta\phi^{(k)} < \theta \Delta l^{(k)} \min(\Delta l^{(k)}/b^{(k)}, 1)$
then set $\rho^{(k+1)} \in [\sigma_1 \|\bar{\delta}^{(k)}\|, \sigma_2 \|\bar{\delta}^{(k)}\|]$ (2.11)
else if $\|\delta^{(k)}\| \geq \rho^{(k)}$
then set $\rho^{(k+i)} \in [\rho^{(k)}, \min(\sigma_3 \rho^{(k)}, \rho_{\max})]$
else set $\rho^{(k+i)} = \rho^{(k)}$;

(iv) **if** $\Delta\phi^{(k)} \leq 0$ **then** set $x^{(k+i)} = x^{(k)}$
else set $x^{(k+i)} = x^{(k)} + \delta^{(k)}$.

In this algorithm $\theta \in (0, \frac{1}{2})$, σ_1, σ_2 $(0 < \sigma_1 < \sigma_2 < 1)$ and σ_3 $(\sigma_3 > 1)$ are fixed parameters, and $\rho_{\max}$ is a user supplied upper limit on $\rho^{(k)}$.

The main theoretical properties that are satisfied by this prototype algorithm include global convergence to a KT point of $\phi(x)$ and convergence of the multiplier estimates $\lambda^{(k)}$. It is also possible to say something about the active set basis matrices $D^{(k)}$. These results are given in the three theorems that follow. For these theorems, $d^{(k)}$ in step (iia) is an arbitrary vector, so we can also allow step (iia) to be repeated a finite number of times with different choices of $d^{(k)}$, without affecting the conclusions. It is important to observe that these theorems do not require any linear independence assumptions.

Theorem 2.1 (Global convergence of the algorithm (2.11)). *Either the sequence* $\{x^{(k)}\}$ *terminates at a KT point, or* $\phi^{(k)} \to -\infty$, *or if the sequence* $\{x^{(k)}\}$ *is bounded, and if* $B^{(k)}$ *is bounded above independently of* k, *then there exists a subsequence* S *with an*

accumulation point x^∞ which satisfies KT conditions, that is

$$\max_{\lambda \in \partial h^\infty} s^T(g^\infty + A^\infty \lambda) \geq 0 \quad \textit{for all } s. \tag{2.12}$$

This condition is equivalent to the statement of the KT conditions given in (1.5).

Proof. We need only consider the case that the sequence fails to terminate and $\{\phi^{(k)}\}$ is bounded below. By considering whether $\inf \rho^{(k)} = 0$ or not, and because $\{x^{(k)}\}$ is bounded, there exists a subsequence S of iterations with $x^{(k)} \to x^\infty$, $k \in S$, for which either

(a) $\delta^{(k)}$ does not satisfy (2.10), $\rho^{(k+1)} \to 0$ and hence $\|\bar{\delta}^{(k)}\| \to 0$ for all $k \in S$, or

(b) $\delta^{(k)}$ satisfies (2.10) and $\inf \rho^{(k)} > 0$, for all $k \in S$.

In case (a) let there exist a descent direction s ($\|s\| = 1$) at x^∞, that is

$$\max_{\lambda \in \partial h^\infty} s^T(g^\infty + A^\infty \lambda) = -\beta, \quad \beta > 0. \tag{2.13}$$

By optimality of $\bar{\delta}^{(k)}$, a consequence of (2.13) is

$$\begin{aligned} \Delta l^{(k)} = \phi^{(k)} - l^{(k)}(\bar{\delta}^{(k)}) &\geq (\phi^{(k)} - l^{(k)}(\|\bar{\delta}^{(k)}\| s)) \\ &\geq \beta \|\bar{\delta}^{(k)}\| + o(\|\bar{\delta}^{(k)}\|) \end{aligned} \tag{2.14}$$

by the corollary to Lemma 14.5.1 of Fletcher (1981). We can use this inequality to deduce a contradition. A consequence of $\mathbb{C}^1$ continuity of f and c, convexity of $h(c)$ and boundedness of $\partial h(c)$ is that

$$\Delta \phi^{(k)} = \Delta l^{(k)} + o(\|\bar{\delta}^{(k)}\|)$$

and hence that $\Delta \phi^{(k)} / \Delta l^{(k)} \to 1$. However (2.14) and (2.5) imply that

$$\Delta l^{(k)} / b^{(k)} \geq \beta(1 + o(1))/(\|\bar{\delta}^{(k)}\| \|B^{(k)}\|)$$

and it follows from $\bar{\delta}^{(k)} \to 0$ and the bound on $B^{(k)}$ that

$$\min(\Delta l^{(k)} / b^{(k)}, 1) = 1$$

for all k sufficiently large. The fact that (2.10) fails for all $k \in S$ thus implies that $\Delta \phi^{(k)} \leq \theta \Delta l^{(k)}$ which establishes the contradiction. Therefore there are no descent directions at x^∞ and the theorem follows for this subsequence.

In case (b) it can be assumed that $\inf \rho^{(k)} > \bar{\rho} > 0$. Because $\phi^{(1)} - \phi^\infty \geq \sum_{k \in S} \Delta \phi^{(k)}$, it follows from (2.10) and the bounds on $B^{(k)}$ and $\rho^{(k)}$ that $\Delta l^{(k)} \to 0$. Define $l^\infty(\delta) = f^\infty + g^{\infty T}\delta + h(c^\infty + A^{\infty T}\delta)$. Let $\bar{\delta}$ minimize $l^\infty(\delta)$ subject to $\|\delta\| \leq \bar{\rho}$ and denote $\bar{x} = x^\infty + \bar{\delta}$. Then

$$\|\bar{x} - x^{(k)}\| \leq \|\bar{x} - x^\infty\| + \|x^\infty - x^{(k)}\| = \|\bar{\delta}\| + o(1) \leq \bar{\rho} + o(1) \leq \rho^{(k)}$$

for all k sufficiently large. Thus $\bar{x}$ is feasible in the subproblem so

$$l^{(k)}(\bar{x} - x^{(k)}) \geq l^{(k)}(\bar{\delta}^{(k)}) = \phi^{(k)} - \Delta l^{(k)}.$$

In the limit, for $k \in S$, $g^{(k)} \to g^\infty$, $c^{(k)} \to c^\infty$, $A^{(k)} \to A^\infty$, $\bar{x} - x^{(k)} \to \bar{\delta}$ and $\Delta l^{(k)} \to 0$, so it follows that $l^\infty(\bar{\delta}) \geq \phi^\infty = l^\infty(0)$. Thus $\delta = 0$ also minimizes $l^\infty(\delta)$ subject to $\|\delta\| \leq \bar{\rho}$, and since the latter constraint is not active it follows that x^∞ is a KT point. □

Subsequently the accumulation point x^∞ is referred to as x^*, and $c^* = c(x^*)$, $\partial h^* = \partial h(c^*)$, etc.

Theorem 2.2 (Convergence of multipliers). *If the subsequence S in the statement of Theorem 2.1 exists, then $\pi^{(k)} \to 0$ for $k \in S$. Moreover any accumulation point, λ^∞ say, of the multiplier vectors $\lambda^{(k)}$, $k \in S$, satisfies $\lambda^\infty \in \Lambda^*$ where $\Lambda^* = \{\lambda : \lambda$ satisfies KT conditions at $x^*\}$, and such an accumulation point exists.*

Proof. The definition of $l^{(k)}(\delta)$ in (2.1) and the subgradient inequality give

$$\begin{aligned} \Delta l^{(k)} &= -g^{(k)\mathrm{T}}\bar{\delta}^{(k)} + h(c^{(k)}) - h(\bar{c}^{(k)}) \\ &\geqslant -g^{(k)\mathrm{T}}\bar{\delta}^{(k)} - \lambda^{(k)\mathrm{T}}A^{(k)\mathrm{T}}\bar{\delta}^{(k)}. \end{aligned} \tag{2.15}$$

It follows from (2.2) above and eq. (14.3.7) of Fletcher (1981) that

$$\Delta l^{(k)} \geqslant \pi^{(k)} w^{(k)\mathrm{T}} \delta^{(k)} = \pi^{(k)} \|\bar{\delta}^{(k)}\|. \tag{2.16}$$

In both case (a) and case (b) of Theorem 2.1 it follows that $\Delta l^{(k)} \to 0$ and hence $\pi^{(k)} \|\bar{\delta}^{(k)}\| \to 0$ from (2.16). From this it can be deduced that $\pi^{(k)} \to 0$ as follows.

Conversely let $\pi^{(k)} \geqslant \beta > 0$ on some subsequence $S' \subset S$. It follows that $\bar{\delta}^{(k)} \to 0$ and also from (2.3) that $\|\bar{\delta}^{(k)}\| = \rho^{(k)}$ for $k \in S'$. These conditions contradict $\rho^{(k)} > \bar{\rho} > 0$ in case (b), so S must be the subsequence that arises from case (a). But in this case we have seen in the argument following (2.14) that the inequality $\Delta l^{(k)} \geqslant \beta \|\bar{\delta}^{(k)}\|$ leads to a contradiction. Thus $\pi^{(k)} \to 0$ for $k \in S$.

Now consider the sequence $\lambda^{(k)}$ for $k \in S$. Because $x^{(k)} \to x^*$ and $\rho^{(k)} \leqslant \rho_{\max}$, it follows that the vectors $\bar{\delta}^{(k)}$ and $\bar{c}^{(k)}$ are bounded. Existence of an accumulation point is then a consequence of Lemma 14.2.1 of Fletcher (1981). Let $\lambda^{(k)} \to \lambda^\infty$ for $k \in S' \subset S$. In the limit it now follows from (2.2) and $\pi^{(k)} \to 0$ that

$$g^* + A^*\lambda^\infty = 0. \tag{2.17}$$

Moreover the subgradient inequality and $\lambda^{(k)} \in \partial h(\bar{c}^{(k)})$ give

$$\begin{aligned} h(c) &\geqslant h(\bar{c}^{(k)}) + (c - \bar{c}^{(k)})^{\mathrm{T}}\lambda^{(k)} \quad \forall c \\ &= h(c^{(k)}) - \Delta l^{(k)} - g^{(k)\mathrm{T}}\bar{\delta}^{(k)} + (c - c^{(k)} - A^{(k)\mathrm{T}}\bar{\delta}^{(k)})^{\mathrm{T}}\lambda^{(k)} \\ &= h(c^{(k)}) - \Delta l^{(k)} + (c - c^{(k)})^{\mathrm{T}}\lambda^{(k)} + \pi^{(k)}\|\bar{\delta}^{(k)}\| \end{aligned}$$

using the definitions of $\Delta l^{(k)}$ and $\bar{c}^{(k)}$, and then (2.2) as above. In the limit $\Delta l^{(k)} \to 0$, $c^{(k)} \to c^*$, $h(c^{(k)}) \to h(c^*)$, $\lambda^{(k)} \to \lambda^\infty$ and $\pi^{(k)}\|\bar{\delta}^{(k)}\| \to 0$, so it follows that

$$h(c) \geqslant h(c^*) + (c - c^*)^{\mathrm{T}}\lambda^\infty \quad \forall c,$$

that is $\lambda^\infty \in \partial h^*$. Together with (2.17) we see that λ^∞ satisfies KT conditions at x^*. □

The next result concerns the convergence of the subdifferential sets in the solution of the linearized subproblem. Because $h(c)$ is polyhedral, it follows from (1.6) that

there are only a finite number of possible sets $\partial h(c)$. If strict complementarity holds then it is shown that the set $\partial h(\bar{c}^{(k)})$ at the solution of the linearized subproblem contains the set $\partial h(c^*)$ at the limit point x^* for sufficiently large k. The opposite inclusion can be proved if the condition $D^{(k)\mathrm{T}}c^* = 0$ holds. Now this condition cannot be deduced from $D^{(k)\mathrm{T}}\bar{c}^{(k)} = 0$ because $\bar{c}^{(k)}$ does not usually converge to c^*. However it is later shown in Theorem 3.2 that under certain assumptions it can be deduced from properties of the Newton-like step. We are then able to conclude that $\partial h(\bar{c}^{(k)}) = \partial h^*$ which implies that it is possible to select $D^{(k)}$ so that asymptotically the correct active set $D^{(k)} = D^*$ is determined by the algorithm.

Theorem 2.3 (Convergence of subdifferential sets). *Assume that the set Λ^* defined in Theorem 2.2 contains only the single vector λ^* (i.e. $\lambda^\infty = \lambda^*$). Consider $k \in S$. Let $\partial h(\bar{c}^{(k)})$ be any fixed subdifferential set which occurs infinitely and let $D^{(k)}$ be a fixed basis for $\partial h(\bar{c}^{(k)}) - \lambda^{(k)}$.*
If $D^{(k)\mathrm{T}}c^* = 0$ then

$$\partial h(\bar{c}^{(k)}) \subset \partial h^*. \tag{2.18}$$

Alternatively, if λ^* is in the relative interior of ∂h^* (i.e. $u^* = 0 \in \operatorname{int} U^*$ in (1.10)), then

$$\partial h(\bar{c}^{(k)}) \supset \partial h^*. \tag{2.19}$$

Proof. Together with Theorem 2.2, the assumption that Λ^* only contains the single vector λ^* implies that $\lambda^\infty = \lambda^*$ and hence that $\lambda^{(k)} \to \lambda^*$, $k \in S$. Because $\partial h(\bar{c}^{(k)})$ is closed it follows that $\lambda^* \in \partial h(\bar{c}^{(k)})$.

The subgradient inequality about c^* and $\bar{c}^{(k)}$ implies that

$$(c^* - \bar{c}^{(k)})^{\mathrm{T}}\lambda^* \geq h(c^*) - h(\bar{c}^{(k)}) \geq (c^* - \bar{c}^{(k)})^{\mathrm{T}}\lambda^{(k)}$$

and because the vectors $\bar{c}^{(k)}$ are bounded it follows that

$$h(c^*) - h(\bar{c}^{(k)}) = (c^* - \bar{c}^{(k)})^{\mathrm{T}}\lambda^* + \mathrm{o}(1). \tag{2.20}$$

Let $S' \subset S$ be the subsequence on which the set $\partial h(\bar{c}^{(k)})$ occurs and consider $k \in S'$. It follows from (1.16) and the assumption of $D^{(k)\mathrm{T}}c^* = 0$ that

$$D^{(k)\mathrm{T}}(\bar{c}^{(k)} - c^*) = 0.$$

Hence by definition of $D^{(k)}$ and $\lambda^* \in \partial h(\bar{c}^{(k)})$ it follows that

$$(\lambda - \lambda^*)^{\mathrm{T}}(\bar{c}^{(k)} - c^*) = 0, \quad \lambda \in \partial h(\bar{c}^{(k)}). \tag{2.21}$$

Consider $\lambda \in \partial h(\bar{c}^{(k)})$. By the subgradient inequality

$$\begin{aligned} h(c) &\geq h(\bar{c}^{(k)}) + (c - \bar{c}^{(k)})^{\mathrm{T}}\lambda \quad \forall c \\ &= h(c^*) + (c - c^*)^{\mathrm{T}}\lambda + \mathrm{o}(1) \end{aligned}$$

from (2.20) and (2.21). By taking the limit it follows that $\lambda \in \partial h^*$ and hence (2.18) is established.

Now consider (2.19). Assume conversely that there exists $\lambda' \in \partial h^*$ for which λ' is not in $\partial h(\bar{c}^{(k)})$. Define $\lambda = \lambda^* + \varepsilon(\lambda^* - \lambda')$ and observe for sufficiently small $\varepsilon > 0$ that $\lambda \in \partial h^*$ by the relative interior property. The subgradient inequality about c^* and $\lambda' \in \partial h^*$ give

$$h(\bar{c}^{(k)}) \geq h(c^*) + (\bar{c}^{(k)} - c^*)^{\mathrm{T}}\lambda' = \bar{c}^{(k)\mathrm{T}}\lambda'$$

by (1.15). Since λ' is not in $\partial h(\bar{c}^{(k)})$ it follows from (1.15) that

$$h(\bar{c}^{(k)}) > \bar{c}^{(k)\mathrm{T}}\lambda'. \tag{2.22}$$

By definition of λ and $\lambda^* \in \partial h(\bar{c}^{(k)})$,

$$\begin{aligned}\bar{c}^{(k)\mathrm{T}}\lambda &= (1+\varepsilon)\bar{c}^{(k)\mathrm{T}}\lambda^* - \varepsilon\bar{c}^{(k)\mathrm{T}}\lambda' \\ &> (1+\varepsilon)h(\bar{c}^{(k)}) - \varepsilon h(\bar{c}^{(k)}) = h(\bar{c}^{(k)})\end{aligned}$$

using (2.22). Hence from (1.15), and because $\lambda \in \partial h^*$,

$$h(\bar{c}^{(k)}) < h(c^*) + (\bar{c}^{(k)} - c^*)^{\mathrm{T}}\lambda$$

which contradicts the subgradient inequality at c^*. Thus λ' not in $\partial h(\bar{c}^{(k)})$ is contradicted and (2.19) follows. □

3. Local properties

In this section it is assumed for the main sequence that $x^{(k)} \to x^*$, $\lambda^{(k)} \to \lambda^*$ and $\pi^{(k)} \to 0$, and local properties of the algorithm are considered. It has been observed that the CNSO problem (1.3) has an equivalent NLP problem (1.12). Using the active set matrix $D^{(k)}$ and multiplier estimate $\lambda^{(k)}$ determined by the linearized subproblem (2.1), the current guess at the equivalent NLP problem is

$$\begin{aligned}&\text{minimize} \quad f(x) + \lambda^{(k)\mathrm{T}}c(x), \\ &\text{subject to} \quad D^{(k)\mathrm{T}}c(x) = 0.\end{aligned} \tag{3.1}$$

In the prototype algorithm (2.11) the first trial correction in step (iia) is to take a Newton-like step for (3.1). In this paper we concentrate on using the Coleman-Conn horizontal + vertical step method. This may be followed by a finite search of certain other points in an attempt to obtain a sufficient reduction, before resorting to the descent step $\bar{\delta}^{(k)}$ in step (iib). The aim of this section is to describe the Coleman-Conn step in this general setting and to show that asymptotically the step satisfies the sufficient reduction criterion (2.10) under mild conditions. It then follows that the local properties of the Coleman-Conn method (two-step superlinear convergence, one-step superlinear convergence in $x^{(k)} + h^{(k)}$, Byrd (1984)) are valid for the prototype algorithm.

Usually the Coleman-Conn method is described in terms of a null space matrix Z having orthonormal columns. However it is possible to use the more general

formulation (generalized elimination) given for example by Fletcher (1981). It is easy to extend the results of Byrd to this situation and it allows us for example to use direct elimination methods to calculate Z. In practice it is most efficient if we determine Z and the other matrices that we require from the factors calculated when solving the linearized subproblem. This is of particular value for example when solving large nonlinear network problems or large sparse nonlinear programming problems. The Jacobian of the constraints in (3.1) is the matrix $A^{(k)}D^{(k)}$ and the required matrices are defined by the equation

$$[A^{(k)}D^{(k)}: V]^{-\mathrm{T}} = [Y: Z]. \tag{3.2}$$

It is assumed that $V = V(x)$ can be regarded as being $\mathbb{C}^1$ in a neighbourhood of x^* (a mild assumption), and that $[A^*D^*: V^*]$ is nonsingular. A consequence of (3.2) is the set of equations

$$\begin{aligned} Y^{\mathrm{T}}A^{(k)}D^{(k)} &= I, \qquad Y^{\mathrm{T}}V = 0, \\ Z^{\mathrm{T}}A^{(k)}D^{(k)} &= 0, \qquad Z^{\mathrm{T}}V = I. \end{aligned} \tag{3.3}$$

The matrices obtained from (3.2) on iteration k are referred to as $V^{(k)}$, $Y^{(k)}$ and $Z^{(k)}$. Of course these matrices may be obtained indirectly as a consequence of some factorized form.

The Coleman-Conn method is one of a number of methods which are based on the use of a matrix, $M^{(k)}$ say, which approximates the reduced Hessian matrix $Z^{*\mathrm{T}}W^*Z^*$, where W^* denotes $\nabla^2(f+\lambda^{*\mathrm{T}}c)$ evaluated at x^*. For theoretical purposes we assume that a sequence of matrices can be calculated for which

$$(M^{(k)} - Z^{(k)\mathrm{T}}W^{(k)}Z^{(k)})V^{(k)\mathrm{T}}\delta^{(k)} = o(\|\delta^{(k)}\|) \tag{3.4}$$

where $W^{(k)} = \nabla^2(f+\lambda^{(k)\mathrm{T}}c)$, evaluated at $x^{(k)}$. This is the usual condition assumed by Byrd in his analysis of local convergence. In our numerical work we have updated a positive definite matrix $M^{(k)}$ using the BFGS method with some modifications proposed by Nocedal and Overton (1985), and rely on this condition to occur. However this is at present an open question since their analysis depends on the assumption that $M^{(1)}$ is sufficiently close to $Z^{*\mathrm{T}}W^*Z^*$, which is unrealistic in practice. We also assume that $M^{(k)}$ is bounded for global convergence (Theorem 2.1), and that $(M^{(k)})^{-1}$ is bounded to prove the local convergence results. The use of a reduced Hessian approximation $M^{(k)}$ implies that the corresponding estimate of W^* can be regarded as being the matrix

$$B^{(k)} = V^{(k)}M^{(k)}V^{(k)\mathrm{T}}. \tag{3.5}$$

It is this matrix that is used to compute $b^{(k)}$ in (2.5) for use in the sufficient reduction test (2.10). Of course this matrix is usually deficient as an estimate of W^*, lacking information about $Y^{\mathrm{T}}W^*Y$ and $Y^{\mathrm{T}}W^*Z$. However the former term does not contribute to the Newton step for solving the KT conditions, and the lack of the latter term is compensated for by the additional evaluation of $c(x)$ used in the Coleman-Conn step.

The Coleman-Conn step is derived directly from the current equivalent NLP problem (3.1). Thus the total step $d^{(k)}$ is the sum of a *horizontal step* $h^{(k)}$ and a *vertical step* $v^{(k)}$ defined by

$$d^{(k)} = h^{(k)} + v^{(k)}, \tag{3.6}$$

$$h^{(k)} = -Z^{(k)}(M^{(k)})^{-1}Z^{(k)\mathrm{T}}(g^{(k)} + A^{(k)}\lambda^{(k)}), \tag{3.7}$$

$$v^{(k)} = -Y^{(k)}D^{(k)\mathrm{T}}c(x^{(k)} + h^{(k)}). \tag{3.8}$$

(Note that $h^{(k)}$ refers to the step given by (3.7) and not to $h(c^{(k)})$). If there are n active constraints (the dimension of $\partial h(\bar{c}^{(k)})$ is n) then $h^{(k)} = 0$, and if there are no active constraints then $v^{(k)} = 0$. Otherwise the method requires the additional evaluation of $c(x^{(k)} + h^{(k)})$ as specified in (3.8). One particular aspect deserves some clarification. The vector $Z^{(k)\mathrm{T}}(g^{(k)} + A^{(k)}\lambda^{(k)})$ in (3.7) occurs frequently: in fact it can be equivalently written as $Z^{(k)\mathrm{T}}(g^{(k)} + A^{(k)}\lambda_0^{(k)})$ where $\lambda_0^{(k)}$ is any other vector in $\partial h(\bar{c}^{(k)})$, and this form may be more convenient for calculation. In particular, for the l_1 exact penalty function $\lambda_0^{(k)}$ could be conveniently chosen as indicated in (1.9), using the vector $\bar{c}^{(k)}$. However the text continues to use the notation in (3.7) for simplicity. Likewise the equivalent NLP problem ((3.1) or (1.14)) could also be given with $\lambda^{(k)}$ replaced by $\lambda_0^{(k)}$, and this indicates that the equivalent problem is solely determined by the estimate of the active constraints derived from $\partial h(\bar{c}^{(k)})$ and is not affected by how close $\lambda^{(k)}$ is to λ^*.

To establish the local properties of the Coleman-Conn step we assume that certain well known *standard conditions* hold, which are:

$f(x)$ and $c(x)$ are $\mathbb{C}^2$ functions,

strict complementarity (λ^* is such that $u^* = 0 \in \text{int } U^*$),

the reduced Hessian $Z^{*\mathrm{T}}W^*Z^*$ is positive definite, and A^*D^* has full rank (linearly independent active constraint gradients).

These conditions are often used and are in the nature of second order sufficient conditions for a minimizer of the CNSO problem (1.3). They are used by Byrd (1984) to derive the local convergence results for the Coleman-Conn method. The final condition ensures that V^* can be chosen to make $[A^*D^* : V^*]$ nonsingular and also ensures that λ^* is unique. Using strict complementarity we can deduce asymptotically from Theorem 2.3 that $\partial h(\bar{c}^{(k)}) \supset \partial h(c^*)$. We also assume for any fixed set $\partial h(\bar{c}^{(k)})$ that $D^{(k)}$ is a fixed matrix: this causes no difficulty, especially for the l_1 exact penalty function in which case the columns of $D^{(k)}$ are simply columns of a unit matrix.

Firstly we consider those iterations, $k \in S^*$ say, for which $\partial h(\bar{c}^{(k)}) = \partial h^*$. Asymptotically, by the full rank assumption, $Z^{(k)}$ and $Y^{(k)}$ are bounded and the estimates

$$\begin{aligned} &\|h^{(k)}\| = \mathrm{O}(\|d^{(k)}\|), \qquad \|v^{(k)}\| = \mathrm{O}(\|d^{(k)}\|), \\ &\|h^{(k)}\| \sim \|Z^{(k)\mathrm{T}}(g^{(k)} + A^{(k)}\lambda^{(k)})\|, \\ &\|v^{(k)}\| \sim \|D^{*\mathrm{T}}c(x^{(k)} + h^{(k)})\| \end{aligned} \tag{3.9}$$

follow from (3.3), (3.6) and the assumed bounds on $M^{(k)}$ and $(M^{(k)})^{-1}$ ($a \sim b \Leftrightarrow a = O(b)$ and $b = O(a)$). It also follows from $g^{(k)} + A^{(k)}\lambda^{(k)} \to g^* + A^*\lambda^* = 0$ that $h^{(k)} \to 0$. Hence $D^{*T}c(x^{(k)} + h^{(k)}) \to D^{*T}c^* = 0$ implies that $v^{(k)} \to 0$ and hence $d^{(k)} \to 0$.

The main result of this section is that asymptotically for $k \in S^*$ the Coleman-Conn method satisfies the sufficient reduction criterion (2.10). Hence most of the effort is devoted to deriving asymptotic estimates of the actual and predicted reductions in ϕ. There are two components of the error in $x^{(k)}$ which must be accounted for. One is the error in the reduced gradient (see (3.9)) which is $\sim\|h^{(k)}\|$. An important feature of what follows is to preserve the significance of the term $\frac{1}{2}h^{(k)T}B^{(k)}h^{(k)}$. Hence we can only allow negligible terms involving $h^{(k)}$ if they are $o(\|h^{(k)}\|^2)$. Terms of order $\|h^{(k)}\|\,\|v^{(k)}\|$ are handled by regarding them as $o(\|v^{(k)}\|)$. The other component of the error is that in the active constraint residuals, that is $D^{*T}c^{(k)}$. A term of the form $(g^{(k)} + A^{(k)}\lambda^{(k)})^T Y^{(k)} D^{*T}c^{(k)}$ arises which is $o(\|D^{*T}c^{(k)}\|)$ since $g^{(k)} + A^{(k)}\lambda^{(k)} \to 0$. By a Taylor series

$$c^{(k)} = c(x^{(k)} + h^{(k)}) - A^{(k)T}h^{(k)} + O(\|h^{(k)}\|^2),$$

so from (3.3)

$$D^{*T}c^{(k)} = D^{*T}c(x^{(k)} + h^{(k)}) + O(\|h^{(k)}\|^2). \tag{3.10}$$

This relates $v^{(k)}$ and $D^{*T}c^{(k)}$, in particular in that from (3.8)

$$\|v^{(k)}\| = O(\|D^{*T}c^{(k)}\|) + O(\|h^{(k)}\|^2). \tag{3.11}$$

Also in both Lemmas 3.1 and 3.2 below a term $h(c^{(k)}) - \lambda^{(k)T}c^{(k)}$ arises. It follows from strict complementarity that

$$h(c^{(k)}) - \lambda^{(k)T}c^{(k)} = \sim\|D^{*T}c^{(k)}\|. \tag{3.12}$$

This result is easily seen in the case of an l_1 exact penalty function. Typically for indices i such that $i \in E$, $c_i^* = 0$, the left hand side of (3.12) includes a term $|c_i^{(k)}| - \lambda_i^{(k)}c_i^{(k)}$ which is $\sim|c_i^{(k)}|$ because $\lambda_i^{(k)} \to \lambda_i^*$ and $-1 < \lambda_i^* < 1$ by strict complementarity. In the general case we can argue as follows. The polyhedral function $h(c)$ is locally linear (see Fletcher, 1987) so asymptotically

$$h(c^{(k)}) = h(c^*) + \max_{\lambda \in \partial h^*} (c^{(k)} - c^*)^T\lambda.$$

Because of $\lambda^{(k)} \in \partial h^*$ and (1.15) it follows that

$$\begin{aligned} h(c^{(k)}) - \lambda^{(k)T}c^{(k)} &= \max_{\lambda \in \partial h^*} (c^{(k)} - c^*)^T\lambda - \lambda^{(k)T}(c^{(k)} - c^*) \\ &= \max_{u \in U^*} u^T D^{*T}c^{(k)} + o(\|D^{*T}c^{(k)}\|) \end{aligned}$$

using $\lambda^{(k)} \to \lambda^*$, (1.10) and (1.16). Strict complementarity is $0 \in \text{int } U^*$ and so (3.12) follows.

We can now prove the main lemmas for actual and predicted reductions in $\phi(x)$ when a Coleman-Conn step is taken.

Lemma 3.1 (Predicted reduction). *If $q^{(k)}(\delta)$ is defined by (2.6) then for sufficiently large $k \in S^*$, under standard conditions, a unique minimizer $\hat{\delta}^{(k)}$ exists and the predicted reduction $\Delta q^{(k)} = q^{(k)}(0) - q^{(k)}(\hat{\delta}^{(k)})$ is given by*

$$\Delta q^{(k)} = \tfrac{1}{2} h^{(k)\mathrm{T}} B^{(k)} h^{(k)} + (g^{(k)} + A^{(k)} \lambda^{(k)})^{\mathrm{T}} Y^{(k)} D^{*\mathrm{T}} c^{(k)} + h(c^{(k)}) - \lambda^{(k)\mathrm{T}} c^{(k)}. \tag{3.13}$$

Proof. Asymptotically, under standard conditions, it is shown in Lemma 4.1 of Womersley (1985) that $\hat{\delta}^{(k)}$ exists and $\partial h(\hat{c}^{(k)}) = \partial h(c^*)$, where $\hat{c}^{(k)}$ denotes $c^{(k)} + A^{(k)\mathrm{T}} \hat{\delta}^{(k)}$. Using the result that relates (1.3) to (1.12), the problem minimize $q^{(k)}(\delta)$ has an equivalent QP problem

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2} \delta^{\mathrm{T}} B^{(k)} \delta + g^{(k)\mathrm{T}} \delta + \lambda^{*\mathrm{T}} (c^{(k)} + A^{(k)\mathrm{T}} \delta) \\ &\text{subject to} && D^{*\mathrm{T}} (c^{(k)} + A^{(k)\mathrm{T}} \delta) = 0. \end{aligned}$$

From the KT conditions, (3.5) and (3.7), and standard conditions, the unique solution of this problem is readily verified to be

$$\hat{\delta}^{(k)} = h^{(k)} - Y^{(k)} D^{*\mathrm{T}} c^{(k)}.$$

It follows from (3.5), (3.3) and (3.7) that both

$$\hat{\delta}^{(k)\mathrm{T}} B^{(k)} \hat{\delta}^{(k)} = h^{(k)\mathrm{T}} B^{(k)} h^{(k)} \tag{3.14}$$

and

$$\hat{\delta}^{(k)\mathrm{T}} g^{(k)} = -h^{(k)\mathrm{T}} B^{(k)} h^{(k)} - (g^{(k)} + A^{(k)} \lambda^{(k)})^{\mathrm{T}} Y^{(k)} D^{*\mathrm{T}} c^{(k)} - \hat{\delta}^{(k)\mathrm{T}} A^{(k)} \lambda^{(k)}. \tag{3.15}$$

Because $\lambda^{(k)} \in \partial h(\bar{c}^{(k)}) = \partial h^* = \partial h(\hat{c}^{(k)})$ it follows from (1.15) and the definition of $\hat{c}^{(k)}$ that

$$h(\hat{c}^{(k)}) = \hat{c}^{(k)\mathrm{T}} \lambda^{(k)} = c^{(k)\mathrm{T}} \lambda^{(k)} + \hat{\delta}^{(k)\mathrm{T}} A^{(k)} \lambda^{(k)}. \tag{3.16}$$

By definition

$$\Delta q^{(k)} = -\hat{\delta}^{(k)\mathrm{T}} g^{(k)} - \tfrac{1}{2} \hat{\delta}^{(k)\mathrm{T}} B^{(k)} \hat{\delta}^{(k)} + h(c^{(k)}) - h(\hat{c}^{(k)})$$

and (3.11) follows directly from (3.14), (3.15) and (3.16). □

Lemma 3.2 (Actual reduction for a Coleman–Conn step). *Asymptotically for $k \in S^*$, under standard conditions at x^*,*

$$\Delta \phi^{(k)} = \Delta q^{(k)} + o(\|h^{(k)}\|^2) + o(\|v^{(k)}\|). \tag{3.17}$$

Proof. In this lemma we denote $\tilde{c}^{(k)} = c(x^{(k)} + h^{(k)}) + A^{(k)\mathrm{T}} v^{(k)}$ and it follows from (3.8) and (3.3) that $D^{(k)\mathrm{T}} \tilde{c}^{(k)} = 0$ and hence $D^{*\mathrm{T}} \tilde{c}^{(k)} = 0$. Then using Taylor series

and boundedness of ∂h,

$$h(c(x^{(k)}+d^{(k)})) = h(c(x^{(k)}+h^{(k)})+A(x^{(k)}+h^{(k)})^{\mathrm{T}}v^{(k)}+o(\|v^{(k)}\|))$$
$$= h(c(x^{(k)}+h^{(k)})+A^{(k)\mathrm{T}}v^{(k)}+o(\|v^{(k)}\|))$$
$$= h(\tilde{c}^{(k)})+o(\|v^{(k)}\|). \tag{3.18}$$

Also by Taylor series

$$\lambda^{(k)\mathrm{T}}(c(x^{(k)}+h^{(k)})-c^{(k)}-A^{(k)\mathrm{T}}h^{(k)})$$
$$=\tfrac{1}{2}h^{(k)\mathrm{T}}(W^{(k)}-G^{(k)})h^{(k)}+o(\|h^{(k)}\|^2) \tag{3.19}$$

using the definition of $W^{(k)}$ after (3.4) and $G^{(k)}=\nabla^2 f(x^{(k)})$. Because $\tilde{c}^{(k)}\to c^*$ and $D^{*\mathrm{T}}\tilde{c}^{(k)}=0$ it follows from Lemma 1.2 that $\partial h(\tilde{c}^{(k)})=\partial h(c^*)=\partial h(\bar{c}^{(k)})$. Hence in a similar way to (3.16) of Lemma 3.1 we can deduce that

$$h(\tilde{c}^{(k)}) = \lambda^{(k)\mathrm{T}}c(x^{(k)}+h^{(k)})+\lambda^{(k)\mathrm{T}}A^{(k)\mathrm{T}}v^{(k)}. \tag{3.20}$$

Finally by definition of $\Delta\phi^{(k)}$ (at $\delta^{(k)}=d^{(k)}$) and Taylor series

$$\Delta\phi^{(k)} = -g^{(k)\mathrm{T}}d^{(k)}-\tfrac{1}{2}d^{(k)\mathrm{T}}G^{(k)}d^{(k)}+o(\|d^{(k)}\|^2)+h(c^{(k)})-h(c(x^{(k)}+d^{(k)}))$$
$$= -g^{(k)\mathrm{T}}d^{(k)}-\tfrac{1}{2}h^{(k)\mathrm{T}}G^{(k)}h^{(k)}+h(c^{(k)})-h(c(x^{(k)}+d^{(k)}))$$
$$+o(\|h^{(k)}\|^2)+o(\|v^{(k)}\|)$$

from (3.6) and (3.9). Merging this with (3.18), (3.19) and (3.20), and rearranging using (3.5), (3.7) and (3.8) gives

$$\Delta\phi^{(k)} = h^{(k)\mathrm{T}}B^{(k)}h^{(k)}-\tfrac{1}{2}h^{(k)\mathrm{T}}W^{(k)}h^{(k)}$$
$$+(g^{(k)}+A^{(k)}\lambda^{(k)})^{\mathrm{T}}Y^{(k)}D^{(k)\mathrm{T}}c(x^{(k)}+h^{(k)})$$
$$+h(c^{(k)})-\lambda^{(k)\mathrm{T}}c^{(k)}+o(\|h^{(k)}\|^2)+o(\|v^{(k)}\|).$$

The quadratic terms can be combined using (3.4) and the next term can be rearranged using (3.10), $D^{(k)}=D^*$ and $g^{(k)}+A^{(k)}\lambda^{(k)}\to 0$. Equation (3.17) then follows from (3.13). □

We can now state the first main result of this section which is that asymptotically for $k\in S^*$, a Coleman-Conn step satisfies the sufficient reduction criterion (2.10).

Theorem 3.1 (Asymptotic behaviour of Coleman-Conn steps). *For $k\in S^*$, if standard conditions hold, if $M^{(k)}$ and $(M^{(k)})^{-1}$ are bounded independently of k, and if* (3.4) *holds, then both $\Delta\phi^{(k)}=\Delta q^{(k)}(1+o(1))$ and* (2.10) *hold asymptotically.*

Proof. Conversely assume that $\exists$ a subsequence $S' \subset S^*$ such that $|\Delta\phi^{(k)}/\Delta q^{(k)} - 1| \geq \gamma > 0$. We can always find a thinner subsequence $S'' \subset S'$ such that either case (i) or case (ii) below holds, and in either case we prove that $\Delta\phi^{(k)}/\Delta q^{(k)} \to 1$ which is a contradiction.

Case (i): $\|D^{*\mathrm{T}}c^{(k)}\| = o(\|h^{(k)}\|^2)$. In this case $\|D^{*\mathrm{T}}c(x^{(k)}+h^{(k)})\| = O(\|h^{(k)}\|^2)$ from (3.10), and hence $\|v^{(k)}\| = O(\|h^{(k)}\|^2)$ from (3.9). Thus (3.13), (3.17) and (3.12) give

$$\frac{\Delta\phi^{(k)}}{\Delta q^{(k)}} = \frac{\frac{1}{2}h^{(k)\mathrm{T}}B^{(k)}h^{(k)} + o(\|h^{(k)}\|^2)}{\frac{1}{2}h^{(k)\mathrm{T}}B^{(k)}h^{(k)} + o(\|h^{(k)}\|^2)} = 1 + o(1).$$

Case (ii): $\|h^{(k)}\|^2 = O(\|D^{*\mathrm{T}}c^{(k)}\|)$. In this case it follows from (3.11) that $\|v^{(k)}\| = O(\|D^{*\mathrm{T}}c^{(k)}\|)$ and hence

$$\frac{\Delta\phi^{(k)}}{\Delta q^{(k)}} = \frac{\frac{1}{2}h^{(k)\mathrm{T}}B^{(k)}h^{(k)} + h(c^{(k)}) - \lambda^{(k)\mathrm{T}}c^{(k)} + o(\|D^{*\mathrm{T}}c^{(k)}\|)}{\frac{1}{2}h^{(k)\mathrm{T}}B^{(k)}h^{(k)} + h(c^{(k)}) - \lambda^{(k)\mathrm{T}}c^{(k)} + o(\|D^{*\mathrm{T}}c^{(k)}\|)}$$

$$= 1 + o(1)$$

since $\frac{1}{2}h^{(k)\mathrm{T}}B^{(k)}h^{(k)} + h(c^{(k)}) - \lambda^{(k)\mathrm{T}}c^{(k)} = \sim\|D^{*\mathrm{T}}c^{(k)}\|$ from (3.12).

Finally (2.10) follows from $\Delta\phi^{(k)} = \Delta q^{(k)}(1+o(1))$ and (2.8). □

Our remaining theoretical result shows that the correct active set is ultimately identified by the algorithm. In deriving this result we allow step (iia) of algorithm (2.11) to try a finite number of Coleman-Conn corrections in which $h^{(k)}$ in (3.7) is replaced by

$$h^{(k)} = -\alpha Z^{(k)}(M^{(k)})^{-1}Z^{(k)\mathrm{T}}(g^{(k)} + A^{(k)}\lambda^{(k)}) \tag{3.21}$$

for $\alpha > 0$, and this value of $h^{(k)}$ is used in (3.8). It is assumed that the unit step $\alpha = 1$ is tried first and that the values of α are bounded above independently of k.

Theorem 3.2. *If the assumptions of Theorem* 3.1 *hold for all k, then* $k \in S^*$ *for all k sufficiently large. Also* $\rho^{(k)}$ *is uniformly bounded away from zero.*

Proof. Denote $S^{*\perp} = \{k: k \notin S^*\}$ as the subsequence on which $\partial h(\bar{c}^{(k)}) \neq \partial h^*$. Let $\bar{S}$ be a thinner subsequence on which $\partial h(\bar{c}^{(k)})$ is constant and assume that $D^{(k)}$ is also constant. For sufficiently large $k \in \bar{S}$, it follows by Theorem 2.3 and strict complementarity that $D^{(k)\mathrm{T}}c^* \neq 0$. Also (2.19) implies that if A^*D^* has full rank then $A^*D^{(k)}$ has full rank, and hence $Z^{(k)}$ can be bounded above asymptotically. Consider $h^{(k)}$ in (3.21). Since $\alpha(M^{(k)})^{-1}$ is bounded by assumption and $g^{(k)} + A^{(k)}\lambda^{(k)} \to 0$ by Theorem 2.2, it follows that $h^{(k)} \to 0$. Thus $c(x^{(k)} + h^{(k)}) \to 0$ and so from the definition of $\tilde{c}^{(k)}$ in Lemma 3.2 and $D^{(k)\mathrm{T}}\tilde{c}^{(k)} = 0$ it follows that

$$D^{(k)\mathrm{T}}A^{(k)\mathrm{T}}v^{(k)} = -D^{(k)\mathrm{T}}c^* + o(1).$$

Since $A^{(k)}$ is bounded and $D^{(k)\mathrm{T}}c^* \neq 0$, it follows that $v^{(k)}$ is uniformly bounded away from zero, and hence there exists $\gamma > 0$ such that $\|d^{(k)}\| \geq \gamma$. Similarly by considering $0 = D^{(k)\mathrm{T}}\bar{c}^{(k)} = D^{(k)\mathrm{T}}c^{(k)} + D^{(k)\mathrm{T}}A^{(k)\mathrm{T}}\bar{\delta}^{(k)}$, a bound $\|\bar{\delta}^{(k)}\| \geq \gamma$ can be

obtained and this also implies that $\rho^{(k)} \geq \gamma$. Since the number of possible sets $\partial h(\bar{c}^{(k)})$ is finite, there exists γ such that these bounds (that is $\|d^{(k)}\| \geq \gamma$, $\|\bar{\delta}^{(k)}\| \geq \gamma$ and $\rho^{(k)} \geq \gamma$) apply for sufficiently large $k \in S^{*\perp}$. Now the actual step (i.e. $x^{(k+1)} - x^{(k)}$) in algorithm (2.11) is either $d^{(k)}$, $\bar{\delta}^{(k)}$ or 0, and since $x^{(k)} \to x^*$ it follows from the bounds on $d^{(k)}$ and $\bar{\delta}^{(k)}$ that $x^{(k+1)} = x^{(k)}$ for sufficiently large $k \in S^{*\perp}$. Thus asymptotically the only steps that reduce $\phi(x)$ are those for $k \in S^*$. These steps will be unit Coleman-Conn steps because the unit step in (2.21) is tried first, and by virtue of Theorem 3.1. Moreover $d^{(k)} \to 0$ for these steps, and the algorithm can increase $\rho^{(k)}$ by a factor of at most σ_3 only if $\|d^{(k)}\| \geq \rho^{(k)}$. Thus it is asymptotically not possible for a step $k \in S^*$ to increase $\rho^{(k)}$ to give $\rho^{(k+1)} \geq \gamma$. Thus step $k+1 \in S^*$ and hence $k \in S^*$ for all k sufficiently large.

Finally because Theorem 3.1 shows for such k that unit Coleman-Conn steps satisfy (2.10), it follows that $\rho^{(k)}$ is not decreased asymptotically and hence is uniformly bounded away from zero. □

4. Numerical experiments and conclusions

The first part of this section describes some pilot calculations with a rudimentary form of the prototype algorithm, in order to indicate the potential of this type of method in practice. We consider the solution of the NLP problem (1.1) using the l_1 exact penalty function (1.2), having first scaled the objective function (by $f \to \nu f$) so that $\|\lambda^*\|_\infty < 1$. The norm used in the trust region bound in (2.1) is the l_∞ norm and (2.1) has been solved by converting it to an LP problem and using a standard package. (Clearly a special purpose l_1LP solver could be expected to be more efficient here.) In common with other implementations of the Coleman-Conn method we have defined the null space matrix $Z^{(k)}$ by the orthogonal factorization method (e.g. see Fletcher, 1981), using QR factors of the current matrix $A^{(k)}D^{(k)}$ of active constraint gradients. Currently we have not yet tried to exploit other factors, in particular the direct elimination factors arising from the solution of the l_1LP sub-problem.

In regard to updating the reduced Hessian approximation $M^{(k)}$ we have tried to use existing technology where possible. However it is by no means certain what is currently best in this respect and any future developments will be very relevant to our algorithm. In fact we do not necessarily need to think in terms of updating $M^{(k)}$ at all: it might be better to consider calculating the reduced Hessian matrix $Z^{(k)\mathrm{T}}W^{(k)}Z^{(k)}$ directly, particularly in the case of sparse NLP problems. Nonetheless in our numerical experiments we have updated the matrix $M^{(k)}$ by the BFGS method

$$M^{(k+1)} = M^{(k)} + \frac{\gamma^{(k)}\gamma^{(k)\mathrm{T}}}{\gamma^{(k)\mathrm{T}}s^{(k)}} - \frac{M^{(k)}s^{(k)}s^{(k)\mathrm{T}}M^{(k)}}{s^{(k)\mathrm{T}}M^{(k)}s^{(k)}}.$$

We have followed Nocedal and Overton (1985) in choosing

$$s^{(k)} = Z^{(k+1)\mathrm{T}}(x^{(k+1)} - x^{(k)}),$$

$$\gamma^{(k)} = Z^{(k+1)\mathrm{T}}(g^{(k+1)} - (g^{(k)} + A^{(k)}\lambda)).$$

In the latter equation λ is any convenient multiplier approximation: we have tried various obvious possibilities with very little difference in our results. We have also used the Nocedal and Overton criterion to skip the update if

$$\| Y^{(k+1)\mathrm{T}}(x^{(k+1)} - x^{(k)})\| \geq \frac{\eta}{(k+1)^{1+\mu}} \|s^{(k)}\| \tag{4.1}$$

choosing $\eta = 1$ and $\mu = 0.01$. This ensures that the BFGS update is always well defined when it is used, and does not impede the local convergence results. On the other hand the dependence on k of the factor on the right-hand side of (4.1) is not very attractive and it may be that a better criterion could be developed.

When the active set changes we need to derive a new positive definite reduced Hessian approximation. We may not always have adequate information in the current matrix so we follow the usual practice of including unit matrix information (e.g. Byrd and Schnabel (1984)). Specifically, if $Q = [Q_1 : Q_2]$ is the orthogonal matrix in the QR factors ($Q_2 = Z$) then we proceed on the assumption that $Z^{\mathrm{T}} W^* Q_i = 0$ and $Q_1{}^{\mathrm{T}} W^* Q_1 = cI$. Thus our current estimate of W^* is $ZMZ^{\mathrm{T}} + cQ_1 Q_1^{\mathrm{T}}$. If we change basis from one null space $Z^{(k)}$ to another $Z^{(k+1)}$, then it follows that the current reduced Hessian approximation M is changed to $T^{\mathrm{T}}(M - cI)T + cI$ where $T = Z^{(k)\mathrm{T}} Z^{(k+1)}$. We have taken the scale factor to be $c = \|M^{(k)}\|$. We only use this transformation when $D^{(k)}$ and $D^{(k+1)}$ differ.

We have performed the experiments on a DEC 10 computer with relative precision $2^{-27} \simeq 0.75_{10} - 8$. We have used a version of algorithm (2.11) in which the search process is a single Coleman-Conn step, with the horizontal step $h^{(k)}$ being truncated if necessary to lie within the trust region box. This truncation allows for $M^{(k)}$ being unduly small and does not seem to occur close to the solution. The simplest way to accomodate this change within our current theory is to regard $M^{(k)}$ as being rescaled accordingly. An alternative would be to show that the results in Section 3 remain valid when only a fraction of the full $h^{(k)}$ step is taken. However it may be that this truncation would be more effectively replaced by a line search along $h^{(k)}$. We have used parameter values $\theta = \frac{1}{4}$, $\sigma_1 = \sigma_2 = \frac{1}{2}$ and $\sigma_3 = 2$ and have made no attempt to optimize these choices. The scaling parameter ν is chosen to be 1 or 0.1 as necessary. We have terminated the iteration when the KT conditions for the NLP problem are satisfied to an accuracy of 10^{-7}. Various small standard test problems have been used: the references for these are given in Sainz de la Maza (1987). The results are set out in Table 4.1. In this table n, m and b denote the numbers of variables, constraints and bounds respectively, and $n - t^*$ is the dimension of the reduced space at the solution. NI, NF, NG and NU denote the numbers of iterations, function + constraint evaluations, gradient evaluations and updates respectively.

If we compare these results with other first derivative methods in the literature we see that our results are often an improvement. Nocedal and Overton (1985) describe a different type of reduced Hessian method that requires 5 and 8 function evaluations for the Wright 1 and 2 problems, 12 function evaluations for the Powell

Table 4.1
Numerical experiments

Problem	n	m	b	$n-t^*$	NI	NF	NG	NU
Wright1	2	1	0	0	4	4	4	3
Wright2	5	3	0	2	6	6	6	4
Chamberlain	2	1	0	0	5	5	5	4
Mukai-Polak	6	2	2	3	13	15	13	9
Powell	5	3	0	2	5	5	5	3
Hock-Schittkowski 100	7	4	0	5	11	19	11	9
Colville2	15	5	15	4	29	40	29	20

problem and 12 for the Hock-Schittkowski 100 problem. Gurwitz (1986) reports various calculations with SQP-type methods which use reduced Hessian approximations. With an l_1 penalty function 63 function evaluations are required for the Chamberlain problem (clearly the Maratos effect is occurring here) and 13 for the Mukai-Polak problem. With an augmented Lagrangian merit function, 6 function evaluations are required for the Chamberlain problem, 14 for the Mukai-Polak problem and 8 for Powell's problem. Gurwitz also reports that the NPSOL package of Gill et al. which approximates full Hessian matrices takes 7, 15 and 10 function evaluations respectively for these problems. For the Colville 2 problem, Powell (1978) requires 17 function evaluations with an SQP-type method which updates full Hessian matrices. Whilst we would not like to make too much of these comparisons, they do suggest that our method is comparable to these other approaches and does not lose out either on account of solving LP rather than QP subproblems, or on account of updating reduced Hessians rather than full Hessians. A point to observe is that all the other methods above find it necessary to use double precision calculation whereas our results are obtained satisfactorily in single precision. However our method does appear to be superior to the original implementation of the Coleman-Conn method (Coleman and Conn, 1982a, b). For example on the Hock-Schittkowski 100 problem they require between 50 and 64 function evaluations to reach a less accurate solution. This can possibly be ascribed to the fact that solving LP-type subproblems locates the correct active set more quickly than the method of tolerances used by Coleman and Conn.

A number of improvements might be made to the pilot code which has been used to derive the above results. A special purpose l_1LP solver would improve the overall efficiency of the code and a sparse matrix version would allow large problems to be solved. The use of direct elimination factors from the l_1LP solver in place of orthogonal factors is another possible saving that could be explored. Some experiments with different parameter selections might be interesting, particularly those that relate to changing $\rho^{(k)}$, but we suspect that our method is relatively insensitive to the values of these parameters. The change that we think would be most useful is to include a line search in step (iia) of algorithm (2.11). Thus we would try points

$x^{(k)} + \alpha h^{(k)}$ ($h^{(k)}$ as in (3.7)) for a range of α-values generated by some combination of sectioning and interpolation. We would follow each of these trials with a vertical step only if an overall reduction in the value of $\phi(x)$ were predicted. This should not only reduce $\phi(x)$ more quickly but should also enable the subsequent update of $M^{(k)}$ to provide more accurate second derivative information. Although we are using the BFGS formula for updating $M^{(k)}$ we are aware that the Nocedal and Overton rule for skipping the update is somewhat ad-hoc and that its theoretical justification is limited. Thus we continue to look for any improvements that arise in this aspect of the subject. Finally we have also considered using other reduced Hessian methods to replace the Coleman-Conn method, in particular the second method analysed Byrd (1984). This work is described by Sainz de la Maza (1987). Sainz de la Maza uses a more complicated algorithm in order to obtain the sufficient decrease property and is not able to detect any advantage to set against this. Thus we have preferred the Coleman-Conn method as our choice of Newton-like step in the algorithm.

Acknowledgements

We wish to acknowledge the financial support of the Spanish Ministry of Education and Science and the U.K. Committee of Vice Chancellors and Principals.

References

R.H. Byrd, "On the convergence of constrained optimization methods with accurate Hessian information on a subspace," Dept. of Computer Science, Report CU-CS-270-84. Univ. of Colorado at Boulder (1984).

R.H. Byrd and R.B. Schnabel "Continuity of the null space basis and constrained optimization," Dept. of Computer Science, Report CU-CS-272-84, Univ. of Colorado at Boulder (1984).

T.F. Coleman and A.R. Conn, "Nonlinear programming via an exact penalty function: Asymptotic analysis," *Mathematical Programming* 24 (1982a) 123-136.

T.F. Coleman and A.R. Conn, "Nonlinear programming via an exact penalty function: Global analysis," *Mathematical Programming* 24 (1982b) 137-161.

R. Fletcher, *Practical Methods of Optimization* 2 (1981); *Constrained Optimization* (Wiley, Chichester, 1981). (References to this volume are also contained in Fletcher (1987) which follows.)

R. Fletcher, *Practical Methods of Optimization*, 2nd Edition (Wiley, Chichester, 1987).

C.B. Gurwitz, "Sequential quadratic programming methods based on approximating a projected Hessian matrix," Comp. Sci. Dept. Report #219, Courant Institute of Mathematical Science (1986).

J. Nocedal and M.L. Overton, "Projected Hessian updating algorithms for nonlinearly constrained optimization," *SIAM Journal on Numerical Analysis* 22 (1985) 821-850.

M.R. Osborne, *Finite Algorithms in Optimization and Data Analysis* (Wiley, Chichester, 1985).

M.J.D. Powell, "A fast algorithm for nonlinearly constrained optimization calculations," G.A. Watson, ed., *Numerical Analysis, Dundee 1977,* Lecture Notes in Mathematics 630 (Springer-Verlag, Berlin, 1978).

E. Sainz de la Maza, "Nonlinear programming algorithms based on l_1 linear programming and reduced Hessian approximations," Ph.D. thesis, Dept. of Mathematical Science, University of Dundee (1987).

R.S. Womersley, "Local properties of algorithms for minimizing nonsmooth composite functions," *Mathematical Programming* 32 (1984) 69-89.