J. Inst. Maths Applics (1975) **15**, 319–342

An Ideal Penalty Function for Constrained Optimization

R. Fletcher
University of Dundee[†]

[Received 8 January 1974 and in revised form 9 May 1974]

A well known approach to constrained optimization is via a sequence of unconstrained minimization calculations applied to a penalty function. This paper shows how it is possible to generalize Powell's penalty function to solve constrained problems with both equality *and* inequality constraints. The resulting methods are equivalent to the Hestenes' *method of multipliers*, and a generalization of this to inequality constraints suggested by Rockafellar. Local duality results (not all of which have appeared before) for these methods are reviewed, with particular emphasis on those of practical importance. It is shown that various strategies for varying control parameters are possible, all of which can be viewed as Newton or Newton-like iterations applied to the dual problem. Practical strategies for guaranteeing convergence are also discussed. A wide selection of numerical evidence is reported, and the algorithms are compared both amongst themselves and with other penalty function methods. The new penalty function is well conditioned, without singularities, and it is not necessary for the control parameters to tend to infinity in order to force convergence. The rate of convergence is rapid and high accuracy is achieved in few unconstrained minimizations. Furthermore the computational effort for successive minimizations goes down rapidly. The methods are very easy to program efficiently, using an established quasi-Newton subroutine for unconstrained minimization.

1. Introduction

Powell (1969) has suggested that to solve the problem

$$\begin{aligned} &\text{minimize} \quad F(\mathbf{x}), && \mathbf{x} \in R^n, \\ &\text{subject to} \quad c_i(\mathbf{x}) = 0 && i = 1, 2, \ldots, m, \end{aligned} \tag{1.1}$$

in the sense of finding a local minimizer $\mathbf{x}^*$, a suitable penalty function is

$$\begin{aligned} \phi(\mathbf{x}, \boldsymbol{\theta}, S) &= F(\mathbf{x}) + \tfrac{1}{2}(\mathbf{c}(\mathbf{x}) - \boldsymbol{\theta})^T S(\mathbf{c}(\mathbf{x}) - \boldsymbol{\theta}) \\ &= F(\mathbf{x}) + \tfrac{1}{2}\sum_i \sigma_i (c_i(\mathbf{x}) - \theta_i)^2 \end{aligned} \tag{1.2}$$

where $\boldsymbol{\theta} \in R^m$, and S is an $m \times m$ diagonal matrix with diagonal elements $\sigma_i > 0$. (In this presentation the signs of the θ_i have been changed from those used by Powell, and a factor $\frac{1}{2}$ introduced in (1.2) to simplify the later analysis.) The penalty function is used in the usual way, that is for any given value of the parameters $\boldsymbol{\theta}$, S, a vector $\mathbf{x}(\boldsymbol{\theta}, S)$ is obtained which minimizes $\phi(\mathbf{x}, \boldsymbol{\theta}, S)$ without constraints. There is an outer iteration in which $\boldsymbol{\theta}$ and S are changed so as to cause the solutions $\mathbf{x}(\boldsymbol{\theta}, S) \to \mathbf{x}^*$. A well known penalty function is one with $\boldsymbol{\theta} = 0$, in which case this convergence is ensured by letting $\sigma_i \to \infty$, $i = 1, 2, \ldots, m$. However Powell suggests an outer iteration for use with (1.2) such that it is not necessary to force $\sigma_i \to \infty$ in order to

[†] Much of this work was carried out whilst the author was at AERE, Harwell.

achieve convergence. Rather the aim is to keep S constant and to let $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is an optimum vector of parameters satisfying

$$\theta_i^* \sigma_i = \lambda_i^* \qquad i = 1, 2, \ldots, m \tag{1.3}$$

where $\boldsymbol{\lambda}^*$ is the vector of Lagrange multipliers for the solution $\mathbf{x}^*$ to (1.1). It is only necessary to increase the σ_i when the rate of convergence of $\mathbf{x}(\boldsymbol{\theta}, S)$ to $\mathbf{x}^*$ is not sufficiently rapid. The method is explained in more detail in Sections 3 and 4.

At about the same time, and independently of Powell, Hestenes (1969) put forward what he called the *method of multipliers*. In this he suggested using the penalty function

$$\psi(\mathbf{x}, \boldsymbol{\lambda}, S) = F(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) + \tfrac{1}{2}\mathbf{c}(\mathbf{x})^T S \mathbf{c}(\mathbf{x}) \tag{1.4}$$

where $\boldsymbol{\lambda} \in R^m$ and S is as above. (In fact Hestenes uses $S = \sigma I$ and therefore implicitly assumes that the constraints are well scaled.) If (1.4) is minimized for fixed $\boldsymbol{\lambda}$, S, then a vector $\mathbf{x}(\boldsymbol{\lambda}, S)$ is obtained. It is clear on expanding (1.2) that if

$$\theta_i \sigma_i = \lambda_i \qquad i = 1, 2, \ldots, m, \tag{1.5}$$

then

$$\phi(\mathbf{x}, \boldsymbol{\theta}, S) = \psi(\mathbf{x}, \boldsymbol{\lambda}, S) + \tfrac{1}{2}\sum_i \lambda_i^2/\sigma_i. \tag{1.6}$$

Because the difference between ϕ and ψ is independent of $\mathbf{x}$, it follows that $\mathbf{x}(\boldsymbol{\lambda}, S) = \mathbf{x}(\boldsymbol{\theta}, S)$ for any S, if $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are related by (1.5). However the penalty function values $\phi(\mathbf{x}(\boldsymbol{\theta}, S), \boldsymbol{\theta}, S)$ and $\psi(\mathbf{x}(\boldsymbol{\lambda}, S), \boldsymbol{\lambda}, S)$ differ, and this difference turns out to be important. Given these relationships between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ the iterative methods suggested by Powell and by Hestenes for changing the $\boldsymbol{\theta}$ (or $\boldsymbol{\lambda}$) parameters are the same. However Powell goes into the situation in much more detail and also suggests an algorithm for increasing S which enables him to prove strong convergence results.

The work in this paper was originally motivated by attempting to modify Powell's function (1.2) to solve the inequality problem

$$\begin{aligned} &\text{minimize} \quad F(\mathbf{x}) \\ &\text{subject to} \quad c_i(\mathbf{x}) \geqslant 0, \qquad i = 1, 2, \ldots, m, \end{aligned} \tag{1.7}$$

by using the penalty function

$$\Phi(\mathbf{x}, \boldsymbol{\theta}, S) = F(\mathbf{x}) + \tfrac{1}{2}\sum_i \sigma_i (c_i(\mathbf{x}) - \theta_i)_-^2, \tag{1.8}$$

where capital Φ denotes the generalization of ϕ to the inequality problem, and where the function a_- is defined as

$$a_- = \min(a, 0) = \begin{cases} a & \text{if } a < 0 \\ 0 & \text{if } a \geqslant 0. \end{cases}$$

In fact there is no difficulty in generalizing (1.8) further to deal with problems with mixed equality/inequality constraints, but the aim here is to keep the notation simple. As before, the case $\boldsymbol{\theta} = 0$ is well known, and convergence can be forced by letting $\sigma_i \to \infty$ $i = 1, 2, \ldots, m$. However difficulties then arise because the second derivative jump discontinuities in (1.8) tend to infinity. Also as the minimizer $\mathbf{x}(0, S)$ converges to the point $\mathbf{x}^*$ at which a discontinuity occurs, so the latter can be expected to become more troublesome. The effect of using the $\boldsymbol{\theta}$ parameters of (1.8) to solve an inequality problem can be illustrated simply. Consider the one variable problem: minimize $F(x)$ subject to $c(x) \geqslant 0$. If an initial choice $\theta = 0$, $\sigma = 1$ is made (assuming the latter is sufficiently large), then the penalty term is only effective for $c < 0$ and the minimum of $\Phi(x, 0, 1)$ is at $c(x) = c_{\min} < 0$ (see Fig. 1). If the correction $\theta' = \theta - c_{\min}$ is made (as suggested by Powell and Hestenes), then for the function $\Phi(x, \theta', 1)$, the penalty term is effective when $c < \theta'$, and a minimum of Φ is created in the neighbourhood of the solution at $c(x) = 0$.

In this paper it will be assumed that $F(\mathbf{x})$ and $c_i(\mathbf{x})$ $i = 1, 2, \ldots, m$, are twice continuously differentiable. Under these circumstances, $\Phi(\mathbf{x})$ is also twice continuously differentiable except at points $\mathbf{x}$ for which any $c_i(\mathbf{x}) = \theta_i$, where the second derivative has a jump discontinuity. However the size of this discontinuity is bounded above when S is bounded, and usually is remote from the minimum, as in Fig. 1, where it does not much affect convergence of the minimization routine.

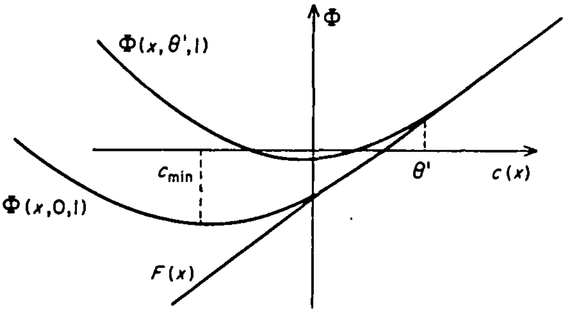

FIG. 1. The penalty function (1.8) for an inequality problem.

In fact a function closely related to (1.8) has already been suggested by Rockafellar, originally in spoken form at the 7th International Mathematical Programming Symposium at The Hague (1970), and more recently in manuscript form (Rockafellar, 1973*a*,*b*, 1974). The idea is to modify the Hestenes function (1.4) giving

$$\Psi(\mathbf{x}, \boldsymbol{\lambda}, S) = F(\mathbf{x}) + \sum_i \begin{cases} -\lambda_i c_i + \frac{1}{2}\sigma_i c_i^2 & \text{if } c_i < \lambda_i/\sigma_i \\ -\frac{1}{2}\lambda_i^2/\sigma_i & \text{if } c_i \geqslant \lambda_i/\sigma_i. \end{cases} \tag{1.9}$$

Actually Rockafellar also considers the more simple case $\sigma_i = \sigma$ $i = 1, 2, \ldots, m$. It is easy to see that the same relationship to (1.6) holds between Ψ and Φ, namely

$$\Phi(\mathbf{x}, \boldsymbol{\theta}, S) = \Psi(\mathbf{x}, \boldsymbol{\lambda}, S) + \tfrac{1}{2} \sum_i \lambda_i^2/\sigma_i, \tag{1.10}$$

and that if $\mathbf{x}(\boldsymbol{\lambda}, S)$ minimizes (1.9), then $\mathbf{x}(\lambda, S) = \mathbf{x}(\boldsymbol{\theta}, S)$ where $\mathbf{x}(\boldsymbol{\theta}, S)$ minimizes (1.8), assuming (1.5). Rockafellar has been concerned primarily with the structure of the function $\Psi(\mathbf{x}(\boldsymbol{\lambda}, S), \boldsymbol{\lambda}, S)$ because strong duality results can be derived when the the original problem (1.7) satisfies certain convexity assumptions. Arrow *et al.* (1973) have also considered the idea and they give local duality results, but their results are unnecessarily restrictive. In his Ph.D. thesis, Buys (1972) is also interested in local results proved without convexity assumptions. Mangasarian (1973) has also investigated developments of the same idea in various directions and there is now much interest in the field, evident by the number of abstracts submitted to the recent 8th Mathematical Programming Symposium at Stanford.

Another antecedent to (1.8) is the barrier function

$$\Phi(\mathbf{x}, \boldsymbol{\theta}, S) = F(\mathbf{x}) - \sum_i \sigma_i \log_e(c_i(\mathbf{x}) - \theta_i)$$

suggested by M. Osborne (private communication), in which the idea is to introduce $\boldsymbol{\theta}$ parameters into other well known barrier functions. Clearly (1.8) is a variation of this same idea.

In Section 2 of this paper a review of the local duality results is given, with particular emphasis on those of practical importance. Buys' (1972) development for the equality problem is largely followed, but in stating the corresponding results for the inequality problem (without proof), Buys seems to be mistaken about the result given here in (2.18) and which is of much importance. The result given here in Theorem 4 is also thought to be new. In Section 3 of this paper it is shown what implications the results of Section 2 have for Newton-like methods to adjust the $\boldsymbol{\lambda}$ parameters. Again Powell (1969), Buys (1972) and Mangasarian (1973) are all aware of this possibility, but it is shown here that such methods can usefully be used even when second derivatives are not available, by minimizing $\phi(\mathbf{x}, \boldsymbol{\theta}, S)$ by a quasi-Newton method, and extracting the approximate hessian for use in the $\boldsymbol{\lambda}$ iteration. Buys is the only person who has considered computational problems for inequality constraints, but because of his apparent mistake about (2.18), he recommends avoiding the Newton-like iteration in this case. In fact this paper indicates theoretically that the iteration is eminently suitable and this is strongly backed up by computational experience. Another way of iterating in the $\boldsymbol{\lambda}$ parameters is that due to Powell and to Hestenes, and a new simple derivation is given of its relation to the Newton iteration.

The convergence of the iterations described in Section 3 is only local, and the iteration must be supplemented by a strategy for increasing S so as to force convergence. The strategy due to Powell (1969) is described, and it is pointed out that it is readily adapted to the inequality problem. Powell's strategy varies either $\boldsymbol{\lambda}$ or S on an iteration, but not both. In fact it is shown that it is consistent to consider algorithms which vary both $\boldsymbol{\lambda}$ and S at every iteration. Three possibilities are suggested. In Section 5 the results of extensive numerical tests are described. All the variations are shown to perform well and their individual merits are discussed. The best variation is compared against other penalty and barrier functions, and it is argued that it is superior from both theoretical and practical considerations.

As to notation, the operators ∇ and ∇^2 will refer to $[\partial/\partial x_i]$ and $[\partial^2/(\partial x_i\,\partial x_j)]$ respectively unless specifically qualified (as ∇_λ). The definitions $\mathbf{g} = \nabla\phi = \nabla\psi$ and $G = \nabla^2\phi = \nabla^2\psi$ will be used. For the equality problem (1.1), $N(\mathbf{x})$ will refer to the matrix $[\nabla c_1, \nabla c_2, \ldots, \nabla c_m]_\mathbf{x}$ and it will be assumed that $N(\mathbf{x}^*) = N^*$ is of full rank. In this case there exist unique multipliers $\boldsymbol{\lambda}^*$ such that

$$\nabla F^* = N^*\boldsymbol{\lambda}^*. \tag{1.11}$$

For the inequality problem, if $\mathbf{x}^*$ is a strong local minimizer of (1.7) then the set of *weakly active* constraints $\mathscr{A}^* = \{i\colon c_i(\mathbf{x}^*) = 0\}$ can be defined. It will be assumed that the vectors ∇c_i $i \in \mathscr{A}^*$ are linearly independent, in which case there exist unique multipliers $\boldsymbol{\lambda}^*$ such that

$$\lambda_i^* \geqslant 0 \qquad i \in \mathscr{A}^*, \qquad \lambda_i^* = 0 \qquad i \notin \mathscr{A}^* \tag{1.12a}$$

$$\nabla F^* = \sum_{i \in A^*} \nabla c_i(\mathbf{x}^*)\lambda_i^* \tag{1.12b}$$

where $A^* = \{i\colon \lambda_i^* > 0\}$ is the set of *strongly active* constraints. Note that $\mathscr{A}^* \supseteq A^*$.

2. Optimality Results for Lagrange Multipliers

In this section some important duality results will be developed, showing that the optimum choice of the $\boldsymbol{\lambda}$ (or $\boldsymbol{\theta}$) parameters for the Powell/Hestenes/Rockafellar penalty function is determined by maximization problem in terms of these parameters. This problem is unconstrained, even for the inequality penalty function. First of all the equality problem (1.1) is considered, subject to the various assumptions of Section 1. Initially a theorem is proved which shows that if the optimum multipliers $\boldsymbol{\lambda}^*$ are chosen in constructing $\psi(\mathbf{x}, \boldsymbol{\lambda}^*, S)$ or $\phi(\mathbf{x}, \boldsymbol{\theta}^*, S)$ where $\boldsymbol{\lambda}^* = S\boldsymbol{\theta}^*$, then $\mathbf{x}^*$ is a strong local minimum of both these functions, and hence $\mathbf{x}^*$ is $\mathbf{x}(\boldsymbol{\lambda}^*)$.

Theorem 1. *If the second order conditions* (2.3 *below*) *on the problem are satisfied, then there exists an* $S' > 0$ *such that for any* $S \geqslant S'$, $\mathbf{x}^*$ *is a strong local minimum with respect to* $\mathbf{x}$ *of both* $\phi(\mathbf{x}, \boldsymbol{\theta}^*, S)$ *and* $\psi(\mathbf{x}, \boldsymbol{\lambda}^*, S)$.

Proof. The argument is first to show that the necessary conditions $\mathbf{c}(\mathbf{x}^*) = 0$ and (1.11) imply that $\nabla\phi(\mathbf{x}^*, \boldsymbol{\theta}^*, S) = 0$ which is a necessary condition for $\mathbf{x}^*$ to be a local minimum of ϕ. This result follows directly from the equation

$$\nabla\phi(\mathbf{x}, \boldsymbol{\theta}, S) = \nabla F(\mathbf{x}) + NS(\mathbf{c}(\mathbf{x}) - \boldsymbol{\theta}). \tag{2.1}$$

Given this result, then a sufficient condition for the theorem to hold is to show that $\nabla^2\phi(\mathbf{x}^*, \boldsymbol{\theta}^*, S)$ is positive definite. Now from (2.1)

$$\nabla^2\phi(\mathbf{x}^*, \boldsymbol{\theta}^*, S) = L^* + N^*SN^{*T}$$

where

$$L^* = \nabla^2 F(\mathbf{x}^*) - \sum_i \lambda_i^* \nabla^2 c_i(\mathbf{x}^*).$$

Consider any vector $\mathbf{u}$ such that $\|\mathbf{u}\|_2 = 1$ and let $\mathbf{u}$ be written as a component $\mathbf{v}$ orthogonal to the columns of N^*, plus a component which is a linear combination of the columns of N^*. This can be conveniently written

$$\mathbf{u} = \mathbf{v} + N^{*+T}\mathbf{w}$$

where $A^+ = (A^TA)^{-1}A^T$. Then

$$\mathbf{u}^T\nabla^2\phi\mathbf{u} = \mathbf{v}^TL^*\mathbf{v} + 2\mathbf{v}^TL^*N^{*+T}\mathbf{w} + \mathbf{w}^TN^{*+}L^*N^{*+T}\mathbf{w} + \mathbf{w}^TS\mathbf{w}. \tag{2.2}$$

Now a sufficient condition for a solution of (1.1) is that there exists an $a > 0$ such that

$$\mathbf{v}^TL^*\mathbf{v} \geqslant a\|\mathbf{v}\|_2^2 \ \forall \ \mathbf{v}\colon N^{*T}\mathbf{v} = 0. \tag{2.3}$$

Writing $\|L^*N^{*+T}\|_2 = b$ and $\|N^{*+}L^*N^{*+T}\|_2 = d$, then from (2.2)

$$\mathbf{u}^T\nabla^2\phi\mathbf{u} \geqslant a\|\mathbf{v}\|_2^2 - 2b\|\mathbf{v}\|_2\|\mathbf{w}\|_2 + (\min_i \sigma_i - d)\|\mathbf{w}\|_2^2.$$

Because $\mathbf{v} = \mathbf{w} = 0$ cannot hold, if $S' = \sigma' I$ is chosen so that $\sigma' > d + b^2/a$, then for all $S \geqslant S'$ it follows that $\mathbf{u}^T\nabla^2\phi\mathbf{u} > 0$. This establishes the positive definiteness of $\nabla^2\phi$ and the theorem follows in respect of ϕ. The same result holds for $\psi(\mathbf{x}, \boldsymbol{\lambda}^*, S)$ because (1.6) shows that the difference between ϕ and ψ is independent of $\mathbf{x}$. □

It is instructive that it is not worthwhile attempting to strengthen this theorem, because when the conditions (2.3) do not hold, then it may not be possible to solve (1.1) by minimizing the function $\phi(\mathbf{x}, \boldsymbol{\theta}^*, S)$ for any S. A simple example due to Hestenes (private communication) illustrates the point well. Let $F(\mathbf{x}) = x_1^4 + x_1x_2$ and $c(\mathbf{x}) = x_2$. Then although $\mathbf{x}^* = 0$ solves (1.1) with optimum multiplier $\lambda^* = 0$, there is no value of σ for which $\phi(\mathbf{x}, 0, S) = x_1^4 + x_1x_2 + \frac{1}{2}\sigma x_2^2$ is minimized at $\mathbf{x}^*$. Henceforth it will be assumed that the conditions (2.3) hold.

For much of this paper it will be the case that a fixed value of S has been chosen which satisfies Theorem 1, and that the interest is in determining optimum $\boldsymbol{\lambda}$ (or $\boldsymbol{\theta}$) parameters. In these cases the explicit dependence of ϕ, ψ, $\mathbf{x}$, etc. on S can and will be dropped. The subsequent analysis can be carried out equivalently either in terms of the parameters $\boldsymbol{\theta}$ or the parameters $\boldsymbol{\lambda}$, assuming that $\boldsymbol{\lambda} = S\boldsymbol{\theta}$. However it turns out to be convenient to work in terms of the $\boldsymbol{\lambda}$ parameters for the most part.

It is useful to regard $\mathbf{x}(\boldsymbol{\lambda})$ as a function implicitly determined by solving the nonlinear equations

$$\nabla\psi(\mathbf{x}, \boldsymbol{\lambda}) = 0. \tag{2.4}$$

Because $\nabla^2\psi(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite, it follows from the implicit function theorem (Hestenes, 1966, for example) that there exist open neighbourhoods $\Omega_\lambda \subset R^m$ about $\boldsymbol{\lambda}^*$ and $\Omega_x \subset R^n$ about $\mathbf{x}^*$ such that for any $\boldsymbol{\lambda} \in \Omega_\lambda$ there exists a unique $\mathbf{x}$ in Ω_x satisfying (2.4), this being the vector $\mathbf{x}(\boldsymbol{\lambda})$. Furthermore $\mathbf{x}(\boldsymbol{\lambda})$ is continuous and continuously differentiable and $\nabla^2\psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ is positive definite for all $\boldsymbol{\lambda} \in \Omega_\lambda$. It may be of course that $\psi(\mathbf{x}, \boldsymbol{\lambda})$ has local minima, so that various solutions to (2.4) exist. It is assumed however that a consistent choice of $\mathbf{x}(\boldsymbol{\lambda})$ is made by the minimization routine, that is the unique solution which exists in Ω_x by virtue of the implicit function theorem.

It is important to look at other quantities derived from $\mathbf{x}(\boldsymbol{\lambda})$ for any given $\boldsymbol{\lambda}$, in particular $\mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}))$ and $\psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$. For convenience these will be written $\mathbf{c}(\boldsymbol{\lambda})$ and $\psi(\boldsymbol{\lambda})$, with the convention that if the dependence on $\mathbf{x}$ is not written explicitly, then $\mathbf{x}(\boldsymbol{\lambda})$ is implied. Because $\mathbf{x}(\boldsymbol{\lambda})$ is differentiable, it is possible to develop an expression for the Jacobian matrix $[d\mathbf{c}/d\boldsymbol{\lambda}]$ of $\mathbf{c}(\boldsymbol{\lambda})$, where $[dc/d\lambda]_{ij} = dc_i/d\lambda_j$ and where the total differential is used in the sense that $\mathbf{x}$ is not held constant but is the function $\mathbf{x}(\boldsymbol{\lambda})$. Because $\mathbf{c}(\boldsymbol{\lambda}) = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}))$, the chain rule implies that

$$[d\mathbf{c}/d\boldsymbol{\lambda}] = [\partial\mathbf{c}/\partial\mathbf{x}][\partial\mathbf{x}/\partial\boldsymbol{\lambda}] = N^T[\partial\mathbf{x}/\partial\boldsymbol{\lambda}]. \tag{2.5}$$

Operating on (2.4) by $[d/d\boldsymbol{\lambda}]$ gives

$$[d\nabla\psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})/d\boldsymbol{\lambda}] = [\partial\nabla\psi/\partial\mathbf{x}][\partial\mathbf{x}/\partial\boldsymbol{\lambda}] + [\partial\nabla\psi/\partial\boldsymbol{\lambda}] = 0. \tag{2.6}$$

But $[\partial\nabla\psi/\partial\mathbf{x}] = \nabla^2\psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = G$ say, and $[\partial\nabla\psi/\partial\boldsymbol{\lambda}] = -N$, so from (2.5) and (2.6) there follows

$$[d\mathbf{c}/d\boldsymbol{\lambda}] = [N^TG^{-1}N]_{\mathbf{x}(\boldsymbol{\lambda})}. \tag{2.7}$$

This result is significant in that it shows that the Jacobian matrix of $\mathbf{c}(\boldsymbol{\lambda})$ is symmetric positive definite at $\boldsymbol{\lambda}^*$ and positive semi-definite for all $\boldsymbol{\lambda} \in \Omega_\lambda$. Therefore it can be expected that there exists a (convex) function of $\boldsymbol{\lambda}$ defined on Ω_λ whose gradient is $\mathbf{c}(\boldsymbol{\lambda})$ and whose hessian matrix is $N^TG^{-1}N$. In fact Buys (1972) points out that $-\psi(\boldsymbol{\lambda})$ is this particular function. This follows because

$$\begin{aligned}[d\psi/d\boldsymbol{\lambda}] &= [\partial\psi/\partial\mathbf{x}][\partial\mathbf{x}/\partial\boldsymbol{\lambda}] + [\partial\psi/\partial\boldsymbol{\lambda}] \\ &= -\mathbf{c}(\boldsymbol{\lambda})\end{aligned}$$

by virtue of equation (2.4) and by $\partial\psi/\partial\lambda_i = -c_i$ from (1.4). Thus $\psi(\boldsymbol{\lambda})$ has derivatives

$$\nabla_\lambda\psi(\boldsymbol{\lambda}) = -\mathbf{c}(\boldsymbol{\lambda}) \tag{2.8a}$$

and

$$\nabla_\lambda^2(\boldsymbol{\lambda}) = -[N^TG^{-1}N]_\lambda. \tag{2.8b}$$

The main optimality result for the equality problem can now be derived. Theorem 1 implies that $\mathbf{x}(\boldsymbol{\lambda}^*)$ solves (1.1) and so is $\mathbf{x}^*$. Therefore $\mathbf{c}(\boldsymbol{\lambda}^*) = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}^*)) = 0$ and hence $\boldsymbol{\lambda}^*$ is a stationary point of $\psi(\boldsymbol{\lambda})$ by (2.8a). Because $\nabla_\lambda^2\psi(\boldsymbol{\lambda}^*)$ is negative definite and because $\psi(\boldsymbol{\lambda})$ is concave, $\boldsymbol{\lambda}^*$ is the only maximizer of $\psi(\boldsymbol{\lambda})$ on Ω_λ and hence is a strong local maximizer of $\psi(\boldsymbol{\lambda})$ on R^m. Thus the problem of finding optimum Lagrange multipliers has been formulated in terms of an optimization problem defined in terms of $\psi(\boldsymbol{\lambda})$. The problem of minimizing the penalty function $\psi(\mathbf{x}, \boldsymbol{\lambda})$ with respect to $\mathbf{x}$ is embedded within this problem of finding a maximizer of $\psi(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$. A study of various iterative methods for solving this latter problem will be made in Section 3. If the functions $F(\mathbf{x})$ and $c_i(\mathbf{x})$ have special features which imply $\mathbf{x}(\boldsymbol{\lambda})$ to be the *global* minimizer of $\psi(\mathbf{x}, \boldsymbol{\lambda})$, then $\boldsymbol{\lambda}^*$ is the *global* maximizer of $\psi(\boldsymbol{\lambda})$. This follows from the inequality

$$\psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \leqslant \psi(\mathbf{x}^*, \boldsymbol{\lambda}) = \psi(\mathbf{x}^*, \boldsymbol{\lambda}^*) \tag{2.9}$$

by virtue of $\mathbf{c}(\mathbf{x}^*) = 0$, a result first used by Powell (1969).

These results concerning the optimality of the multipliers $\boldsymbol{\lambda}^*$ can all be extended to cover the inequality problem (1.7) by using the penalty function $\Psi(\mathbf{x}, \boldsymbol{\lambda})$ defined in (1.9). As before S will be fixed and $\boldsymbol{\lambda}$ or $\boldsymbol{\theta}$ used interchangeably subject to $\boldsymbol{\lambda} = S\boldsymbol{\theta}$. In order to prove a result analogous to Theorem 1, it is useful to state sufficient conditions for the function $\Phi(\mathbf{x}, \boldsymbol{\theta}, S)$ defined in (1.8) to have a strong local minimum. The usual conditions break down when $\nabla^2\Phi$ exhibits a jump discontinuity at $\mathbf{x}(\boldsymbol{\theta})$. Let $\mathbf{x}(\boldsymbol{\lambda})$ minimize $\Psi(\mathbf{x}, \boldsymbol{\lambda})$ and hence $\Phi(\mathbf{x}, \boldsymbol{\theta})$ with respect to $\mathbf{x}$. Then two constraint index sets can be defined,

$$\begin{aligned} M(\mathbf{x}) &= \{i \colon c_i(\mathbf{x}) < \theta_i\} \\ Z(\mathbf{x}) &= \{i \colon c_i(\mathbf{x}) = \theta_i\}, \end{aligned} \tag{2.10}$$

for which the $c_i - \theta_i$ terms are *M*inus and *Z*ero respectively. Constraints in M (only) contribute to $\Phi(\mathbf{x}, \boldsymbol{\theta})$, whilst in some neighbourhood of $\mathbf{x}$, some constraints in Z may also contribute. If R is a general constraint index set then the function

$$\phi_R(\mathbf{x}, \boldsymbol{\theta}) = F(\mathbf{x}) + \tfrac{1}{2}\sum_{i \in R} \sigma_i (c_i(\mathbf{x}) - \theta_i)^2 \tag{2.11}$$

can be defined. A useful result is then as follows.

LEMMA 1. $\Phi(\mathbf{x}, \boldsymbol{\theta})$ *has a strong local minimum at* $\mathbf{x}'$ *if* $\nabla\Phi(\mathbf{x}', \boldsymbol{\theta}) = 0$ *and* $\nabla^2\phi_{M(x')}(\mathbf{x}', \boldsymbol{\theta})$ *is positive definite.*

Proof. Because $\nabla\Phi(\mathbf{x}', \boldsymbol{\theta}) = \nabla\phi_{M(\mathbf{x}')}(\mathbf{x}', \boldsymbol{\theta})$, these conditions imply that $\mathbf{x}'$ minimizes $\phi_{M(\mathbf{x}')}(\mathbf{x}, \boldsymbol{\theta})$. But $\Phi(\mathbf{x}', \boldsymbol{\theta}) = \phi_{M(\mathbf{x}')}(\mathbf{x}', \boldsymbol{\theta})$ and $\Phi(\mathbf{x}, \boldsymbol{\theta}) \geqslant \phi_{M(\mathbf{x}')}(\mathbf{x}, \boldsymbol{\theta})$ for all $\mathbf{x}$ in some neighbourhood of $\mathbf{x}'$, so the Lemma follows. □

Some insight into this result is provided by looking at the hessian matrix $\nabla^2\phi_{R(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta})$ where $R(\mathbf{x})$ is any index set such that

$$M(\mathbf{x}) \subseteq R(\mathbf{x}) \subseteq M(\mathbf{x}) \cup Z(\mathbf{x}). \tag{2.12}$$

Now by the definition of Φ (see (1.8)) it follows that

$$\nabla^2\Phi(\mathbf{x}, \boldsymbol{\theta}) = F(\mathbf{x}) + \sum_i \nabla^2 c_i \sigma_i (c_i(\mathbf{x}) - \theta_i)_- + \sum_{i \in M(\mathbf{x})} \sigma_i \nabla c_i \nabla c_i^T$$

which is continuous at $\mathbf{x}$ (for fixed $\boldsymbol{\theta}$) when $Z(\mathbf{x})$ is empty. When $Z(\mathbf{x})$ is not empty however, the last term is discontinuous and makes $\nabla^2\Phi$ indeterminate. However the hessian matrices $\nabla^2\phi_{R(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta})$ for any $R(\mathbf{x})$ such that (2.12) holds, satisfy the bounds

$$\nabla^2\phi_{M(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta}) \leqslant \nabla^2\phi_{R(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta}) \leqslant \nabla^2\phi_{M(\mathbf{x}) \cup Z(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta}). \tag{2.13}$$

Thus Lemma 1 can be paraphrased as follows: although $\nabla^2\Phi$ is not well-determined, the sufficient condition for a strong local minimum of Φ (given $\nabla\Phi = 0$) is that, of all the related hessian matrices, the smallest $\nabla^2\phi_{M(\mathbf{x}')}$ shall be positive definite.

Another useful result which follows from this analysis of hessian matrices is as follows.

LEMMA 2. *If* $\nabla^2\phi_{M(\mathbf{x})}(\mathbf{x}, \boldsymbol{\theta})$ *for fixed* $\boldsymbol{\theta}$, *is positive definite for all* $\mathbf{x}$ *in some open convex subset of* R^n *then* $\Phi(\mathbf{x}, \boldsymbol{\theta})$ *is strictly convex on the subset.*

Proof. Given a point $\mathbf{x}'$ in the subset, and a direction $\mathbf{v}$ ($\|\mathbf{v}\|_2 = 1$), then for all $R(\mathbf{x}')$ satisfying (2.12)

$$\lim_{\alpha\to 0}\frac{\mathbf{v}^T(\nabla\phi_{R(\mathbf{x}')}(\mathbf{x}'+\alpha\mathbf{v})-\nabla\phi_{R(\mathbf{x}')}(\mathbf{x}'))}{\alpha}=\mathbf{v}^T\nabla^2\phi_{R(\mathbf{x}')}(\mathbf{x}')\mathbf{v}. \tag{2.14}$$

Because of (2.13) it follows that $\mathbf{v}^T\nabla^2\phi_{R(\mathbf{x}')}(\mathbf{x}')\mathbf{v}$ can be bounded below by the least eigenvalue of $\nabla^2\phi_{M(\mathbf{x}')}(\mathbf{x}')$, and hence uniformly away from zero. By definition of Φ, for all α sufficiently small it is possible to equate $\nabla\Phi(\mathbf{x}'+\alpha\mathbf{v})$ with some $\nabla\phi_{R(\mathbf{x}')}(\mathbf{x}'+\alpha\mathbf{v})$ and hence by (2.14) there exists an $a > 0$ such that

$$\frac{\mathbf{v}^T(\nabla\Phi(\mathbf{x}'+\alpha\mathbf{v})-\nabla\Phi(\mathbf{x}'))}{\alpha}\geqslant a$$

for all α sufficiently small. It follows that the component of $\nabla\Phi$ along any line in the subset increases strictly monotonically, and it is then a straightforward application of the mean value theorem to show that $\Phi(\mathbf{x})$ is strictly convex. □

Corollary. The weaker assumption of positive semi-definiteness in Lemma 2 implies convexity (not strict). The proof is based similarly on (2.14) and is not difficult. □

A result analogous to Theorem 1 can now be demonstrated.

THEOREM 2. *If second order sufficient conditions* (2.15) *below are satisfied by the inequality problem* (1.7), *then there exists an* $S' > 0$ *such that for all* $S \geqslant S'$, $\mathbf{x}^*$ *is a strong local minimum of* $\Phi(\mathbf{x}, \boldsymbol{\theta}^*, S)$ *with respect to* $\mathbf{x}$, *where* $\boldsymbol{\lambda}^* = S\boldsymbol{\theta}^*$.

Proof. The necessary condition $\nabla\Phi = 0$ follows from (1.12b) in the same way as for Theorem 1. Mildly restrictive sufficient conditions for the problem (1.7) to have a strong local minimum at $\mathbf{x}^*$ are to assume that there exists an $a > 0$ such that

$$\mathbf{v}^TL^*\mathbf{v}\geqslant a\|\mathbf{v}\|_2^2 \quad \forall\ \mathbf{v}:\mathbf{v}^T\nabla c_i^*=0\quad\forall\quad i\in A^*, \tag{2.15}$$

(see Fiacco & McCormick, 1968), where L^* is defined as in Theorem 1. But $A^* \equiv M^*$ (i.e. $M(\mathbf{x}^*)$), so it follows by the proof of Theorem 1 that $\nabla^2\phi_{M^*}(\mathbf{x}^*, \boldsymbol{\theta}^*)$ is positive definite. Theorem 2 now follows from Lemma 1. As in Theorem 1, an identical result holds for $\Psi(\mathbf{x}, \boldsymbol{\lambda}^*, S)$. □

Again it is useful to regard $\mathbf{x}(\boldsymbol{\lambda})$ as a function implicitly determined by solving the non-linear equations

$$\nabla\Psi(\mathbf{x}, \boldsymbol{\lambda}) = 0. \tag{2.16}$$

Similar statements to those in the paragraph containing (2.4) can be made, except that $\mathbf{x}(\boldsymbol{\lambda})$ is differentiable only if $Z(\mathbf{x}(\boldsymbol{\lambda}))$ is empty, and the implication about $\nabla^2\Phi$ is that $\nabla^2\phi_{M(\mathbf{x}(\boldsymbol{\lambda}))}(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\lambda}\in\Omega_\lambda$ (again $\boldsymbol{\lambda} = S\boldsymbol{\theta}$). These statements are a consequence of the implicit function theorem, because if R^* is any index set defined by (2.12) at $\mathbf{x}^*$, then there exist neighbourhoods Ω_λ and Ω_x such that for any $\boldsymbol{\lambda}\in\Omega_\lambda$, the minimizer $\mathbf{x}_{R^*}(\boldsymbol{\lambda})$ of the function $\phi_{R^*}(\mathbf{x}, \boldsymbol{\theta})$ is in Ω_x. If the minimizer of $\Phi(\mathbf{x}, \boldsymbol{\theta})$ is not in Ω_x, then its minimizer in $\bar{\Omega}_x$ (closure) is on the boundary of $\bar{\Omega}_x$ at $\mathbf{x}'$ say. If sets M', Z' are defined at $\mathbf{x}'$ analogously to (2.10), then $\mathbf{x}'$ minimizes $\phi_{M' \cup Z'}$. in Ω_x. If $\Omega_\lambda \times \Omega_x$ is sufficiently small, then $M' \cup Z'$ is one of the above R^* and so the assertion that the ϕ_{R^*} are strictly convex with unique minimizers in Ω_x contradicts the minimality of $\mathbf{x}'$. Hence an $\mathbf{x}(\boldsymbol{\lambda})$ which minimizes $\Phi(\mathbf{x}, \boldsymbol{\lambda})$ exists in Ω_x. Uniqueness follows by strict convexity of Φ (Lemma 2) and continuity by the continuity of the $\mathbf{x}_{R^*}(\boldsymbol{\lambda})$.

It is again convenient to write $\Psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ as $\Psi(\boldsymbol{\lambda})$ and $\mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}))$ as $\mathbf{c}(\boldsymbol{\lambda})$. Following the analysis for the equality problem, it is interesting to examine derivatives of $\Psi(\boldsymbol{\lambda})$. Assume first that $Z(\mathbf{x}(\boldsymbol{\lambda}))$ is empty so that

$$[d\Psi/d\boldsymbol{\lambda}] = [\partial\Psi/\partial\mathbf{x}][\partial\mathbf{x}/\partial\boldsymbol{\lambda}] + [\partial\Psi/\partial\boldsymbol{\lambda}].$$

Then because $\mathbf{x}(\boldsymbol{\lambda})$ satisfies (2.16), it follows that

$$d\Psi(\boldsymbol{\lambda})/d\lambda_i = \partial\Psi/\partial\lambda_i = \begin{cases} -c_i & \text{if } c_i < \theta_i \\ -\theta_i & \text{if } c_i > \theta_i. \end{cases}$$

If however $Z(\mathbf{x}(\boldsymbol{\lambda}))$ is not empty, then $[\partial\mathbf{x}/\partial\boldsymbol{\lambda}]$ is not well defined, and differs depending on which set $R(\mathbf{x}(\boldsymbol{\lambda}))$ is chosen. However the resulting value of $\nabla_\lambda \Psi$ is independent of this choice, so it is well determined and can be written

$$d\Psi(\boldsymbol{\lambda})/d\lambda_i = -\min(c_i, \theta_i). \tag{2.17}$$

A further differentiation shows that

$$\nabla_\lambda^2 \Psi = \begin{array}{r} i \in M(\boldsymbol{\lambda}) \, \{ \\ i \notin M(\boldsymbol{\lambda}) \, \{ \end{array} \begin{array}{|c|c|} \hline -N^T G^{-1} N & 0 \\ \hline 0 & -S^{-1} \\ \hline \end{array} \tag{2.18}$$

where $M(\boldsymbol{\lambda})$ means $M(\mathbf{x}(\boldsymbol{\lambda}))$, where $G = \nabla^2 \phi_{M(\lambda)}(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ is derived from (2.11), and where the columns of N are taken from the constraints $i \in M(\boldsymbol{\lambda})$ and those of S^{-1} from the constraints $i \notin M(\boldsymbol{\lambda})$. The matrix $\nabla_\lambda^2 \Psi(\boldsymbol{\lambda})$ is well defined only when $Z(\boldsymbol{\lambda})$ is empty. When this is not so, $\nabla_\lambda^2 \Psi(\boldsymbol{\lambda})$ has a jump discontinuity, and different hessian matrices could be defined in terms of any index set $R(\boldsymbol{\lambda})$ which satisfies $M(\boldsymbol{\lambda}) \subseteq R(\boldsymbol{\lambda}) \subseteq M(\boldsymbol{\lambda}) \cup Z(\boldsymbol{\lambda})$. However these matrices would all be negative definite at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and at least negative semi-definite for all $\boldsymbol{\lambda} \in \Omega_\lambda$. Hence by arguments similar to those of Lemma 2 and its corollary, $\Psi(\boldsymbol{\lambda})$ is concave for all $\boldsymbol{\lambda} \in \Omega_\lambda^?$ and strictly concave for $\boldsymbol{\lambda}$ in some smaller neighbourhood of $\boldsymbol{\lambda}^*$.

These results can be brought together to demonstrate the optimality properties of Lagrange multipliers in the inequality problem. By Theorem 2, $\mathbf{x}^*$ which is a strong local solution to (1.7), is also a minimizer of (1.8) and is therefore $\mathbf{x}(\boldsymbol{\lambda}^*)$. Hence from the necessary conditions (1.12a) and the definition of $\mathscr{A}^*$ it follows that

$$\min(c_i^*, \lambda_i^*/\sigma_i) = 0$$

for all i. Therefore by (2.17), $\boldsymbol{\lambda}^*$ is a stationary point of $\Psi(\boldsymbol{\lambda})$ and hence maximizes $\Psi(\boldsymbol{\lambda})$ by the concavity results of the last paragraph. As for the equality problem, this result can be used to determine iterative methods for locating the optimum multipliers $\boldsymbol{\lambda}^*$, and this is also taken up in Section 3. A stronger result, analogous to (2.9), can be stated when the problem (1.7) is such that $\mathbf{x}(\boldsymbol{\lambda})$ can be guaranteed to be the *global* minimizer of $\Psi(\mathbf{x}, \boldsymbol{\lambda})$. Then the inequality

$$\Psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \leqslant \Psi(\mathbf{x}^*, \boldsymbol{\lambda}) \leqslant \Psi(\mathbf{x}^*, \boldsymbol{\lambda}) + \tfrac{1}{2} \sum_{i:\,\lambda_i < 0} \lambda_i^2/\sigma_i = \Psi(\mathbf{x}^*, \boldsymbol{\lambda}^*)$$

holds, showing that $\boldsymbol{\lambda}^*$ is a *global* maximizer of $\Psi(\boldsymbol{\lambda})$.

Finally two other results will be mentioned which give intuition to the choice of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ parameters in these penalty functions. Both show that minimizing a penalty function is equivalent to solving a constrained minimization problem which is a perturbation of the original problem. For the equality problem Powell (1969) showed that Theorem 3 holds, and Theorem 4 is a new result.

THEOREM 3. $\mathbf{x}(\boldsymbol{\lambda})$ *is a strong minimizer of the problem: minimize* $F(\mathbf{x})$ *subject to* $\mathbf{c}(\mathbf{x}) = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}))$.

The proof (see Powell) is immediate, and the result emphasizes the need to choose $\boldsymbol{\lambda}$ so as to make $\mathbf{c}(\mathbf{x}(\boldsymbol{\lambda})) = 0$. A similar result is demonstrated for the inequality problem.

THEOREM 4. $\mathbf{x}(\boldsymbol{\lambda})$ *is a strong local minimizer of the problem: minimize* $F(\mathbf{x})$ *subject to* $c_i(\mathbf{x}) \geqslant \min(c_i(\boldsymbol{\lambda}), \theta_i)$ *for all i.*

Proof. By the optimality of $\mathbf{x}(\boldsymbol{\lambda})$, for any neighbouring $\mathbf{x}$ it follows that $\Phi(\mathbf{x}, \boldsymbol{\lambda}) > \Phi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ and hence

$$F(\mathbf{x})+\tfrac{1}{2}\sum_i \sigma_i(c_i(\mathbf{x})-\theta_i)_-^2 > F(\mathbf{x}(\boldsymbol{\lambda}))+\tfrac{1}{2}\sum_i \sigma_i(c_i(\mathbf{x}(\boldsymbol{\lambda}))-\theta_i)_-^2.$$

If now $c_i(\mathbf{x}) \geqslant \min(c_i(\mathbf{x}(\boldsymbol{\lambda})), \theta_i)$ for all i, then (i) if $c_i(\mathbf{x}(\boldsymbol{\lambda})) \geqslant \theta_i$ then $(c_i(\mathbf{x})-\theta_i)_- = 0 = (c_i(\mathbf{x}(\boldsymbol{\lambda}))-\theta_i)_-$, and (ii) if $c_i(\mathbf{x}(\boldsymbol{\lambda})) < \theta_i$ then $(c_i(\mathbf{x}(\boldsymbol{\lambda}))-\theta_i)_- \leqslant (c_i(\mathbf{x})-\theta_i)_- \leqslant 0$. Hence for both (i) and (ii), $F(\mathbf{x}) > F(\mathbf{x}(\boldsymbol{\lambda}))$ and the Theorem follows. □

This time the need to choose $\boldsymbol{\lambda}$ so as to make $\min(c_i, \theta_i) = 0$ is emphasized.

3. Strategies for Changing $\boldsymbol{\lambda}$ (or $\boldsymbol{\theta}$)

In this section iterations for correcting the $\boldsymbol{\lambda}$ parameters will be investigated of the form

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}+\Delta\boldsymbol{\lambda}^{(k)} \tag{3.1}$$

where superscripts denote iteration number. Because $\psi(\boldsymbol{\lambda})$ is twice differentiable it is possible to state a Newton algorithm for this iteration. By virtue of (2.8) this is the correction

$$\Delta\boldsymbol{\lambda}^{(k)} = -(N^TG^{-1}N)^{-1}\mathbf{c} \tag{3.2}$$

where $N = N(\mathbf{x}(\boldsymbol{\lambda}^{(k)}))$ etc. For an inequality constraint problem, by virtue of (2.17) and (2.18) the correction (3.2) is appropriate for the multipliers λ_i $i \in M(\boldsymbol{\lambda}^{(k)})$, together with the correction

$$\Delta\lambda_i^{(k)} = -\lambda_i^{(k)} \qquad i \notin M(\boldsymbol{\lambda}^{(k)}) \tag{3.3}$$

for the other multipliers. However the correction (3.2) does not necessarily keep $\boldsymbol{\lambda} \geqslant 0$, and in view of the necessary condition (1.12a) a probably more effective iteration is to choose $\boldsymbol{\lambda}^{(k+1)}$ as the minimizer of the subproblem

$$\text{minimize } Q^{(k)}(\boldsymbol{\lambda}) \qquad \text{subject to } \boldsymbol{\lambda} \geqslant 0 \tag{3.4}$$

where

$$Q^{(k)}(\boldsymbol{\lambda}) = \tfrac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^{(k)})^T\{\nabla_\lambda^2\Psi(\boldsymbol{\lambda}^{(k)})\}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^{(k)})+(\boldsymbol{\lambda}-\boldsymbol{\lambda}^{(k)})^T\nabla_\lambda\Psi(\boldsymbol{\lambda}^{(k)}) \tag{3.5}$$

comprises the first and second terms of the Taylor expansion of $\Psi(\boldsymbol{\lambda})$ about $\boldsymbol{\lambda}^{(k)}$. In fact the solution of (3.4) for the multipliers λ_i $i \notin M(\boldsymbol{\lambda}^{(k)})$ is also given by (3.3), so it is only necessary to pose (3.4) in terms of the multipliers λ_i $i \in M(\boldsymbol{\lambda}^{(k)})$, and there are usually not more than n elements in this set. The subproblem (3.4) is a quadratic program with only simple lower bounds and can be solved efficiently (Fletcher & Jackson, 1974).

If exact values of $\nabla_\lambda \Psi$ and $\nabla_\lambda^2 \Psi$ are available, and if convergence of (3.2) to the optimum parameters $\boldsymbol{\lambda}^*$ does occur, then the rate of convergence is known to be second order. This is also true for (3.4) assuming that $A^* \equiv \mathscr{A}^*$ (strict complementarity) because ultimately the bounds $\lambda_i \geqslant 0$ $i \in M(\boldsymbol{\lambda}^{(k)})$ are not active. However I would expect a second order rate of convergence to hold even when strict complementarity does not occur.

In view of the fact that $\boldsymbol{\lambda}^*$ is the unconstrained maximum of $\Psi(\boldsymbol{\lambda})$ however, there is no necessity to solve the quadratic program (3.4). The Newton iteration based on (3.2) and (3.3) is more simple and avoids the need for a routine for (3.4). It would also be possible to modify the solution to (3.2) to ensure $\boldsymbol{\lambda} \geqslant 0$ by any convenient *ad hoc* rule, and the resulting algorithm would probably not be much inferior to (3.4).

It is well known that a disadvantage of Newton's method is that both first and second derivatives of the objective function must be available. The same is true here because to evaluate $\nabla_\lambda^2 \Psi$ requires the matrix $G = \nabla^2 \Psi(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ which involves second derivatives of the functions $F(\mathbf{x})$ and $c_i(\mathbf{x})$. Nevertheless most good unconstrained minimization routines build up estimates of G (or G^{-1}) either directly as in quasi-Newton methods, or indirectly as in conjugate direction methods. Thus if this estimate is used in (3.2) or (3.4) a Newton-like algorithm results for which a rapid rate of convergence is likely even though only first derivatives of the functions F and c_i are required.

An even more simple correction formula is suggested for the equality problem by Powell (1969) and by Hestenes (1969). It will be shown simply how this formula can be derived from (3.2) and also what the corresponding formula is for the inequality problem. Consider the matrix $G^*(S) = \nabla^2 \psi(\mathbf{x}^*, \boldsymbol{\lambda}^*, S)$ and let S be increased by adding a large positive diagonal matrix D. Then

$$G^*(S+D) = G^*(S) + N^* D N^{*T} \tag{3.5}$$

follows from the definition of ψ, (1.4). Use of the well known formula for the correction to the inverse matrix (see Householder, 1964, for example) enables an expression for $[G^*(S+D)]^{-1}$ to be obtained. By applying the same formula again it follows that

$$(N^{*T}[G^*(S+D)]^{-1}N^*)^{-1} = (N^{*T}[G^*(S)]^{-1}N^*)^{-1} + D. \tag{3.6}$$

Consider now the iteration (3.1) and let the correction be determined by

$$\Delta\boldsymbol{\lambda}^{(k)} = -A^{(k)}\mathbf{c}^{(k)}.$$

The resulting iteration is superlinearly convergent if $A^{(k)} \to (N^{*T}G^{*-1}N^*)^{-1}$. Furthermore linear convergence at any desired rate can be achieved if $\lim_{k\to\infty} A^{(k)}$ is sufficiently close to $(N^{*T}G^{*-1}N^*)^{-1}$, relatively speaking. Now if all the elements of D are large relative to S, then it follows that

$$(N^{*T}[G^*(S+D)]^{-1}N^*)^{-1} = (S+D) + O(S). \tag{3.7}$$

That is to say, if $S := S+D$ is made sufficiently large, then arbitrarily good relative agreement between S and $[\nabla_\lambda^2 \Psi]^{-1}$ is obtained. If $A^{(k)}$ is replaced by S then the correction becomes

$$\Delta\boldsymbol{\lambda}^{(k)} = -S\mathbf{c}^{(k)}, \tag{3.8a}$$

or in terms of the $\boldsymbol{\theta}$ parameters

$$\Delta\boldsymbol{\theta}^{(k)} = -\mathbf{c}^{(k)}, \tag{3.8b}$$

and these are the simple formulae suggested by Hestenes and Powell respectively.

One especial merit is that the formulae do not require the derivatives N and G and therefore are suitable when the minimization routine applied in $\mathbf{x}$-space does not require derivatives. An equally simple correction formula holds for the inequality problem. Equation (3.7) shows that the hessian of the quadratic function $Q^{(k)}(\boldsymbol{\lambda})$ in (3.4) tends to the diagonal matrix S^{-1} as $k \to \infty$ and $S \to \infty$. Using this fact to determine the correction yields the formula

$$\Delta\theta_i^{(k)} = -\min(c_i^{(k)}, \theta_i^{(k)}) \qquad i = 1, 2, \ldots, m, \tag{3.9a}$$

or in terms of the $\boldsymbol{\lambda}$ parameters,

$$\Delta\lambda_i^{(k)} = -\min(\sigma_i^{(k)}c_i^{(k)}, \lambda_i^{(k)}), \quad i = 1, 2, \ldots, m. \tag{3.9b}$$

In Section 5 some numerical results are given comparing the Newton-like iteration based on (3.4) against the Powell/Hestenes iteration (3.9).

The Newton correction (3.2) has also been suggested by Buys when second derivatives are explicitly available, but he does not seem to consider using approximations to G. Furthermore he concludes (wrongly in my view—see (2.18)) that $\nabla_\lambda^2\Psi(\boldsymbol{\lambda})$ is always singular and that this precludes the use of Newton-like algorithms for varying $\boldsymbol{\lambda}$ in the inequality problem. In my opinion (backed up by the numerical experience of Section 5) the local rate of convergence of these algorithms is very satisfactory.

4. Convergence: Variation of S

So far attention has only been given to local convergence results for iterative methods which find the optimum multipliers $\boldsymbol{\lambda}^*$. In this section consideration will be given to ways of obtaining guaranteed convergence by varying the elements of S. In fact under certain conditions it is possible to prove convergence of (3.8) without increasing S at all, providing that Theorem 1 is satisfied. To do this requires an inequality given by Rockafellar for the case $S = \sigma I$, that for all $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}'$

$$\Psi(\boldsymbol{\lambda}') \geqslant \Psi(\boldsymbol{\lambda}) + (\boldsymbol{\lambda}'-\boldsymbol{\lambda})^T\nabla_\lambda\Psi(\boldsymbol{\lambda}) - \tfrac{1}{2}\sigma^{-1}(\boldsymbol{\lambda}'-\boldsymbol{\lambda})^T(\boldsymbol{\lambda}'-\boldsymbol{\lambda}).$$

This inequality shows that the matrix $-\nabla_\lambda^2\Psi$ is bounded away from zero by $\sigma^{-1}I$. However the derivation of this inequality depends strongly on convexity assumptions about the original problem, and I am not clear to what extent such conclusions are valid in general.

Most global convergence results make the assumption that $\mathbf{x}(\boldsymbol{\lambda})$ is the *global* minimizer of the penalty function. It is not known to what extent this is inevitable, and it would be of interest to conduct a study into this problem. The role of S in forcing convergence is illustrated by the result that if $\sigma_i^{(k)} \to \infty$ and $\boldsymbol{\lambda}$ is fixed, then $c_i^{(k)} \to 0$, and behaves asymptotically like $c_i^{(k)} \sim \text{const}/\sigma_i^{(k)}$. Thus it would be expected that a sensible way of using S is to increase S so as to force the iterates $\boldsymbol{\lambda}^{(k)}$ into a region about $\boldsymbol{\lambda}^*$ in which the local convergence results are valid. Once in this region then S stays fixed and $\boldsymbol{\lambda}$ alone is varied, and $\boldsymbol{\lambda} \to \boldsymbol{\lambda}^*$ at the appropriate rate.

A suitable way of achieving this aim is suggested by Powell (1969) in regard to the equality problem (1.1). He measures convergence in terms of

$$K = \|\mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}), S)\|_\infty. \tag{4.1}$$

If an iteration is deemed to be completed when a better value of K is found, and if the corresponding $\boldsymbol{\lambda}$, S are denoted by $\boldsymbol{\lambda}^{(k)}$, $S^{(k)}$, then Powell's method is the following.

Initially set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(1)}$, $S = S^{(1)}$, $k = 0$, $K^{(0)} = \infty$.

```
(i) Evaluate $\mathbf{x}(\boldsymbol{\lambda}, S)$ and $\mathbf{c} = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}, S))$.          (4.2)
(ii) Find $\{i\colon |c_i| \geqslant K^{(k)}/4\}$. If $\|\mathbf{c}\|_\infty \geqslant K^{(k)}$ goto (v).
(iii) Set $k = k+1$, $\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}$, $S^{(k)} = S$, $K^{(k)} = \|\mathbf{c}\|_\infty$. Finish if $K^{(k)} \leqslant \varepsilon$.
(iv) If $K^{(k)} \leqslant K^{(k-1)}/4$ or $\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}^{(k-1)}$, (set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)} - S^{(k)}\mathbf{c}^{(k)}$, goto (i)).
(v) Set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)}$, $\sigma_i = 10\sigma_i \; \forall\, i \in \{i\}$, goto (i).
```

The test $\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}^{(k-1)}$ in (iv) is equivalent to testing whether control last flowed into (i) from (v). This iteration is chosen so that ultimately $K^{(k)}$ is reduced by a factor of at least 1/4 at each iteration. Although an early iteration may involve more than one unconstrained minimization of the penalty function, each iteration cannot fail to terminate. Powell proves convergence of his iteration in that the test $K^{(k)} \leqslant \varepsilon$ must ultimately be satisfied. To do this requires the global result (2.9), and also the assumption that $F(\mathbf{x})$ is bounded below, although this latter assumption can be dispensed with. A similar process to (4.2) can be used to force convergence of the Newton iteration (3.2), merely by an appropriate change of the formula $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)} - S^{(k)}\mathbf{c}^{(k)}$ in (4.2 (iv)).

An important observation is that any minimizer $\mathbf{x}(\boldsymbol{\lambda}, S)$ of the penalty function $\psi(\mathbf{x}, \boldsymbol{\lambda}, S)$ could have been obtained from any one of an infinity of values of $\boldsymbol{\lambda}, S$. If $\boldsymbol{\lambda}', S'$ is one of these values, then it is only necessary for $\boldsymbol{\lambda}', S'$ to be related to $\boldsymbol{\lambda}, S$ by

$$\boldsymbol{\lambda}' - S'\mathbf{c} = \boldsymbol{\lambda} - S\mathbf{c} \tag{4.3}$$

where $\mathbf{c} = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}, S)) = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}', S'))$. This is by virtue of the condition

$$\nabla\psi(\mathbf{x}(\boldsymbol{\lambda}, S), \boldsymbol{\lambda}, S) = \nabla F(\mathbf{x}(\boldsymbol{\lambda}, S)) - N(\boldsymbol{\lambda} - S\mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}, S))) = 0$$

which remains satisfied when the change (4.3) is made. It is possible therefore to change $\boldsymbol{\lambda}, S$ subject to (4.3) after a minimization is completed (subject to S being sufficiently large so that $\nabla^2\psi$ is still positive definite) and it is interesting to observe what happens if the subsequent iteration is carried out in terms of the new values $\boldsymbol{\lambda}', S'$. In fact because $\boldsymbol{\lambda} - S\mathbf{c}$ remains constant, and because the new iterate $\boldsymbol{\lambda}^{(k+1)}$ is just $\boldsymbol{\lambda} - S\mathbf{c}$, it is easy to see that $\boldsymbol{\lambda}^{(k+1)}$ remains unaffected when the Powell/Hestenes correction is used on the values $\boldsymbol{\lambda}', S'$. The same is also true when $\boldsymbol{\lambda}^{(k+1)}$ is determined by Newton's method (3.2) although the analysis is less obvious. This must be so however because the Newton prediction is exact for some problems and therefore should be independent of how it is predicted. The only effect of making the change (4.3) then is that the value of $S^{(k+1)}$ which is used with $\boldsymbol{\lambda}^{(k+1)}$ is changed. Therefore it is quite consistent when devising an algorithm, to make the change to $\boldsymbol{\lambda}$ by virtue of (3.2) or (3.8) based on $\boldsymbol{\lambda}^{(k)}, S^{(k)}$, and then to choose $S^{(k+1)}$ arbitrarily larger than $S^{(k)}$. An algorithm in which both are changed simultaneously is more efficient because it avoids an evaluation of $\mathbf{x}(\boldsymbol{\lambda}, S)$ which would be needed if the changes were made separately.

An example of this is in the Powell algorithm (4.2) where if $K^{(k-1)} > K^{(k)} > K^{(k-1)}/4$ then $\boldsymbol{\lambda}$ is kept constant whilst S is increased. In fact it is possible to correct $\boldsymbol{\lambda}$ and also increase S at the same time, and this modification has been found slightly more efficient. The modified algorithm is

```
Initially set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(1)}$, $S = S^{(1)}$, $k = 0$, $K^{(0)} = \infty$.
    (i) Evaluation $\mathbf{x}(\boldsymbol{\lambda}, S)$ and $\mathbf{c} = \mathbf{c}(\mathbf{x}(\boldsymbol{\lambda}, S))$.          (4.4)
    (ii) Find $\{i\colon |c_i| \geqslant K^{(k)}/4\}$. If $\|\mathbf{c}\|_\infty \geqslant K^{(k)}$, (set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)}$, goto (v)).
```

(iii) Set $k = k+1$, $\lambda^{(k)} = \lambda$, $S^{(k)} = S$, $K^{(k)} = \|\mathbf{c}\|_\infty$, Finish if $K^{(k)} \leqslant \varepsilon$.
(iv) Set $\lambda = \lambda^{(k)} - S^{(k)}\mathbf{c}^{(k)}$, If $K^{(k)} \leqslant K^{(k-1)}/4$, goto (i).
(v) Set $\sigma_i = 10\sigma_i \; \forall \; i \in \{i\}$, goto (i).

Some results are given in Section 5 for this algorithm, and also for one in which a Newton-like correction is used in (4.4(iv)).

The algorithm (4.4) is valid only for the equality problem but a small modification enables it to be used for the inequality problem. To do this the definition of K (4.1) is changed so that

$$K = \max_i |\min(c_i(\lambda, S), \theta_i)|$$

and a similar change is made where $\|\mathbf{c}\|_\infty$ occurs in (4.4). Also the correction formula to be used must be one appropriate to the inequality problem (that is (3.4) or (3.9)). Of course it is possible to solve problems with mixed equality and inequality constraints and the modifications to do this are similar.

When using either the Powell/Hestenes formula (3.8) or the Newton-like formula based on (3.2) for changing λ inside an algorithm like (4.4), it will be noticed that with the Newton-like method S is only increased so as to force global convergence, whereas with the Powell/Hestenes formula it may be increased further to force the sufficiently rapid rate of linear convergence to λ^*. Thus it would be expected that the values of σ_i used by the Newton-like method would be smaller than those used by the Powell/Hestenes method. This has been borne out in practice. Intuitively this would appear to be good in that too large a value of S might make the minimum $\mathbf{x}(\lambda, S)$ difficult to determine due to ill-conditioning of the penalty function. However no such evidence has been forthcoming on the variety of problems considered, and indeed some advantages of having S larger have showed up. One of these of course is that the larger S is, so the more rapidly is the region reached in which the local convergence results are valid. Furthermore the basic strategy of (4.2) and (4.4) is dependent on the asymptotic result $c_i \sim \text{const}/\sigma_i$ and this is only valid for large σ_i. Although clearly the optimum values of S must be a balance between these effects, practical experience indicates that the values of S chosen when using the Powell/Hestenes correction formula in (4.4) have not caused any loss of accuracy in $\mathbf{x}(\lambda, S)$.

Two observations from the numerical evidence of Section 5 suggest a further algorithm for changing S. It is noticeable that the rapid rate of convergence of the Newton-like iteration is significant in reducing the overall number of iterations when there is no difficulty in getting convergence to occur. On the other hand the values of S required to get the appropriate rate of linear convergence with the Powell/Hestenes formula, are always sufficient to force global convergence. Therefore a Newton-like algorithm has been investigated in which S is chosen so that the prediction obtained from the Powell/Hestenes formula (3.8) should reduce $K^{(k)}$ to $K^{(k)}/4$. To do this it is assumed that $\psi(\lambda)$ is quadratic, so that the Newton correction (3.2) should reduce $K^{(k)}$ to zero. Let Δ^{PH} and Δ^{N} be the corrections predicted by the Powell/Hestenes and Newton formulae respectively, and λ^{PH} and λ^{N} the corresponding prediction of the multipliers. Because of (2.8a, b), a Taylor series for $\mathbf{c}(\lambda)$, together with $\mathbf{c}(\lambda^*) = 0$ and $\lambda^N = \lambda^*$ gives

$$(N^TG^{-1}N)^{-1}\mathbf{c}(\lambda^{PH}) = \lambda^{PH} - \lambda^* = \Delta^{PH} - \Delta^{N}. \tag{4.5}$$

The right hand side of (4.5) is independent of S, so the effect of increasing S is to increase $(N^TG^{-1}N)^{-1}$ and hence to decrease $\mathbf{c}(\lambda^{PH}, S)$. For large S, $(N^TG^{-1}N)^{-1}$ can be estimated by S (see (3.7) for instance). Now S is to be found so that

$$|c_i(\lambda^{PH}, S)| \leqslant |c_i(\lambda^{(k)}, S^{(k)})|/4 = |\Delta_i^{PH}|/(4\sigma_i^{(k)})$$

by (3.8). Hence from (4.5) it follows that

$$\sigma_i \geqslant 4\sigma_i^{(k)}\left|\frac{\Delta_i^{PH}-\Delta_i^N}{\Delta_i^{PH}}\right|. \tag{4.6}$$

An algorithm based on using this formula to choose σ_i at each iteration has also been tried. In this algorithm σ_i is only increased by the factor 10 if the λ correction formula fails to improve K. The algorithm as it relates to equality constraints is

```
Initially set $\lambda = \lambda^{(1)}$, $S = S^{(1)}$, $k = 0$, $K^{(0)} = \infty$.
  (i) Evaluate $\mathbf{x}(\lambda, S)$ and $\mathbf{c} = \mathbf{c}(\mathbf{x}(\lambda, S))$.        (4.7)
  (ii) If $\|\mathbf{c}\|_\infty \geqslant K^{(k)}$, (set $\sigma_i = 10\sigma_i \,\forall\, i: |c_i| \geqslant K^{(k)}$, goto (i)).
  (iii) Set $k = k+1$, $\lambda^{(k)} = \lambda$, $S^{(k)} = S$, $K^{(k)} = \|\mathbf{c}\|_\infty$, Finish if $K^{(k)} \leqslant \varepsilon$.
  (iv) Set $\lambda$ by (3.2), increase each $\sigma_i$ if necessary so as to satisfy (4.6), goto (i).
```

The modifications to deal with inequalities are straightforward as described above, and numerical experience with the algorithm is set out in Section 5.

These ideas by no means exhaust the possibilities for a strategy for changing S, and in particular no algorithms have been tried in which the value of $\Psi(\lambda, S)$ is used. Yet this information is readily available, so further research in this direction might be fruitful.

5. Practical Experience and Discussion

In this section numerical experience gained with the algorithm (4.4) will be described, using both the Powell/Hestenes formulae (3.8, 3.9) and the Newton-like formula based on (3.2) and (3.4) to change the λ parameters. Experience with the algorithm (4.7) is also described. A general program has been written and modified for each of these algorithms, but various features are common to all. The program works with scaled constraint values, that is the user supplies a vector $\bar{\mathbf{c}} > 0$ whose magnitude is typical of that of the constraint functions $\mathbf{c}(\mathbf{x})$. The program then works with constraint functions $\mathbf{c}'(\mathbf{x})$, where $c_i' = c_i/\bar{c}_i$. The initial λ and S are set automatically by the program unless the user chooses otherwise. For instance the user might want to try the choice $\lambda = S\mathbf{c}+N^+\nabla F$ which minimizes $\|\nabla\phi\|_2$. The automatic choice of λ is $\lambda = 0$, and the choice for S is based on the following criterion. A rough estimate of the likely change ΔF in F on going to the solution is made by the user. ΔF is used to scale the other terms which occur in the penalty function, so that σ_i is set to make $\frac{1}{2}\sigma_i c_i'^2 = |\Delta F|$. A quasi-Newton method VA09A from the Harwell subroutine library is used to minimize $\Phi(\mathbf{x}, \boldsymbol{\theta}, S)$ with respect to $\mathbf{x}$, and the initial estimate of $\nabla^2\Phi$ can either be set automatically to I, or otherwise by the user. However the suggestion by Buys (1972) that the estimate be reset to the unit matrix whenever the active set is changed is not used, and in my opinion would be rather inefficient. In fact the approximation to $\nabla^2\Phi$ is carried forward from one minimization to the next, and whenever S is changed the estimate of $\nabla^2\Phi$ is changed by virtue of (3.5). This involves a rank one correction to the estimate for every σ_i which is increased.

The routine (and also VA09A), uses LDL^T factorizations to represent $\nabla^2\Phi$, and this contributes to the accuracy of the process. Double length computation on an IBM 370/165 computer is used in the tabulated results and the convergence criterion is that $K \leqslant 10^{-6}$.

A wide selection of test problems has been used. These are the parcel problem (PP) of Rosenbrock (1960), the problem (RS) due to Rosen & Suzuki (1965), the problem (P) due to Powell (1969), and four test problems (TP1, 2, 3 and 7) used in the comparisons carried out by Colville (1968). The features of these problems are set out in Table 1, where m_i, m_e, and m_a indicate the numbers of inequality, equality and active (at $\mathbf{x}^*$) constraints respectively. $P(A)$ and $P(B)$ etc. indicates that the same problem has been repeated with different initial S values. The criterion used for comparison is the number of times that F, $\mathbf{c}$, ∇F and $[\nabla c_i]$ together are evaluated for given $\mathbf{x}$. In fact however ∇c_i is not evaluated for any inequality constraint for which $c_i \geqslant \theta_i$, and ∇c_i for any linear constraint is set on entry to the program as it is a constant vector. Not only has the total number of evaluations on each problem been tabulated but also the number required on iterations after the first. The first minimization is the same in each case and is often the most expensive, and this can obscure the comparison.

TABLE 1

Resumé of problems and performance

Problem	Type of problem				Performance					
	n	m_i	m_e	m_a	Powell		Newton		Mod. Newton	
PP	3	7	—	1	37†	22‡	30	15	30	15
RS	4	3	—	1	57	37	36	16	35	15
P(A)	5	—	3	3	45	27	32	14	32	14
P(B)	5	—	3	3	52	36	40	24	37	21
*TP*1	5	15	—	4	51	36	40	25	39	24
*TP*2	15	20	—	11	149	21	181	53	162	34
*TP*3*(A)*	5	16	—	5	95	70	113	88	101	76
*TP*3*(B)*	5	16	—	5	64	33	94	63	64	33
*TP*7	16	32	8	13	89	73	65	49	53	37

† Total number of evaluations.
‡ Number of evaluations excepting the first minimization.

The detailed performance of the three different algorithms tested is given in Tables 2, 3 and 4. The most striking feature is that the number of outer iterations taken by methods based on the Newton-like formula is far fewer than is taken when using the Powell/Hestenes formula. This substantiates empirically the second-order convergence of these methods, even though the second derivatives $\nabla_\lambda^2\Psi(\lambda)$ are not calculated exactly. Another pointer to this fact is in the values of K on the later minimizations. For the Powell/Hestenes formula the values go down in a way which appears to be linear and the final K values are all in the range $(10^{-6}, 10^{-7})$. For the Newton-like formula the ratio of successive K values increases for increasing k, suggesting superlinear convergence, and the final K values are often much smaller than 10^{-6}. However, the

TABLE 2

Powell/Hestenes correction (3.9) *in algorithm* (4.4)

Minimization	1	2	3	4	5	6	7	8	9	10
PP	15 0·15	† 8 0·014	4 $0{\cdot}1\times10^{-2}$	4 $0{\cdot}1\times10^{-3}$	3 $0{\cdot}1\times10^{-4}$	3 $0{\cdot}8\times10^{-6}$	← No. of evaluations ← K			
RS	20 0·17	7 0·033	5 0·017	6 $0{\cdot}3\times10^{-2}$	5 $0{\cdot}5\times10^{-3}$	4 $0{\cdot}8\times10^{-4}$	4 $0{\cdot}1\times10^{-4}$	3 $0{\cdot}2\times10^{-5}$	3 $0{\cdot}3\times10^{-6}$	
P(*A*)	18 0·067	7 0·011	4 $0{\cdot}2\times10^{-2}$	4 $0{\cdot}3\times10^{-3}$	4 $0{\cdot}5\times10^{-4}$	3 $0{\cdot}8\times10^{-5}$	3 $0{\cdot}1\times10^{-5}$	2 $0{\cdot}2\times10^{-6}$		
P(*B*)	16 0·84	7 0·60	8 0·20	7 0·014	5 $0{\cdot}2\times10^{-2}$	4 $0{\cdot}1\times10^{-3}$	3 $0{\cdot}7\times10^{-5}$	2 $0{\cdot}4\times10^{-6}$		
*TP*1	15 0·45	6 0·10	16 0·032	4 $0{\cdot}2\times10^{-2}$	4 $0{\cdot}7\times10^{-4}$	3 $0{\cdot}3\times10^{-5}$	3 $0{\cdot}1\times10^{-6}$			
*TP*2	128 0·28	15 $0{\cdot}5\times10^{-4}$	6 $0{\cdot}5\times10^{-6}$							
*TP*3(*A*)	25 0·95	5 0·48	13 0·34	19 0·20	7 0·22	8 0·039	6 $0{\cdot}4\times10^{-2}$	5 $0{\cdot}1\times10^{-3}$	4 $0{\cdot}6\times10^{-5}$	3 $0{\cdot}3\times10^{-6}$
*TP*3(*B*)	31 0·27	11 0·11	7 0·014	5 $0{\cdot}6\times10^{-3}$	4 $0{\cdot}4\times10^{-4}$	3 $0{\cdot}3\times10^{-5}$	3 $0{\cdot}2\times10^{-6}$			
*TP*7	16 0·96	10 0·75	11 0·30	9 0·15	12 0·067	12 0·010	7 $0{\cdot}4\times10^{-2}$	5 $0{\cdot}9\times10^{-4}$	4 $0{\cdot}7\times10^{-5}$	3 $0{\cdot}6\times10^{-6}$

† A vertical rule indicates the stage at which a correct active set is established.

Table 3

Newton-like correction based on (3.4) *in algorithm* (4.4)

Minimization	1	2	3	4	5	6	7	8	9	10
PP	15 0·15	│† 8 │ $0{\cdot}4\times10^{-2}$	4 $0{\cdot}7\times10^{-5}$	3 $0{\cdot}5\times10^{-9}$						
RS	20 0·17	│ 7 │ $0{\cdot}4\times10^{-2}$	5 $0{\cdot}5\times10^{-4}$	4 $0{\cdot}8\times10^{-6}$						
P(*A*)	│ 18 │ 0·067	7 $0{\cdot}2\times10^{-2}$	4 $0{\cdot}1\times10^{-4}$	3 $0{\cdot}1\times10^{-7}$						
P(*B*)	│ 16 │ 0·84	9 0·26	7 0·011	5 $0{\cdot}7\times10^{-4}$	3 $0{\cdot}6\times10^{-6}$					
*TP*1	15 0·45	18 0·024	│ 4 │ $0{\cdot}1\times10^{-3}$	3 $0{\cdot}1\times10^{-6}$						
*TP*2	128 0·28	16 0·050	│ 19 │ $0{\cdot}2\times10^{-2}$	9 $0{\cdot}2\times10^{-3}$	5 $0{\cdot}5\times10^{-5}$	4 $0{\cdot}9\times10^{-7}$				
*TP*3(*A*)	25 0·95	35 2·15	24 0·32	│ 11 │ 0·12	10 $0{\cdot}3\times10^{-3}$	5 $0{\cdot}5\times10^{-5}$	3 $0{\cdot}1\times10^{-7}$			
*TP*3(*B*)	31 0·27	38 1·09	13 0·040	│ 5 │ $0{\cdot}5\times10^{-3}$	4 $0{\cdot}2\times10^{-5}$	3 $0{\cdot}1\times10^{-10}$				
*TP*7	16 0·96	15 0·30	│ 12 │ 0·051	8 $0{\cdot}2\times10^{-2}$	6 $0{\cdot}2\times10^{-3}$	5 $0{\cdot}6\times10^{-5}$	3 $0{\cdot}1\times10^{-6}$			

† A vertical rule indicates the stage at which a correct active set is established.

TABLE 4

Newton-like correction based on (3.4) *in algorithm* (4.7)

Minimization	1	2	3	4	5	6	7	8	9	10
PP	15 0·15	│† 8 $0·4\times10^{-2}$	4 $0·7\times10^{-5}$	3 $0·5\times10^{-9}$						
RS	20 0·17	│ 7 $0·4\times10^{-2}$	5 $0·8\times10^{-5}$	3 $0·4\times10^{-7}$						
P(*A*)	│ 18 0·067	7 $0·2\times10^{-2}$	4 $0·2\times10^{-4}$	3 $0·2\times10^{-6}$						
P(*B*)	│ 16 0·84	8 0·10	6 $0·9\times10^{-3}$	4 $0·2\times10^{-4}$	3 $0·6\times10^{-10}$					
*TP*1	15 0·45	17 $0·7\times10^{-2}$	│ 4 $0·5\times10^{-4}$	3 $0·3\times10^{-6}$						
*TP*2	128 0·28	16 0·050	│ 8 $0·3\times10^{-2}$	6 $0·3\times10^{-4}$	4 $0·9\times10^{-6}$					
*TP*3(*A*)	25 0·95	29 0·23	27 0·049	│ 12 $0·6\times10^{-3}$	5 $0·3\times10^{-4}$	3 $0·3\times10^{-7}$				
*TP*3(*B*)	31 0·27	│ 8 0·062	17 $0·2\times10^{-2}$	5 $0·2\times10^{-4}$	3 $0·2\times10^{-7}$					
*TP*7	16 0·96	14 0·15	│ 10 0·083	6 $0·6\times10^{-3}$	4 $0·7\times10^{-5}$	3 $0·4\times10^{-7}$				

† A vertical rule indicates the stage at which a correct active set is established.

difference in number of evaluations is not as severe as this discrepancy in minimizations might suggest, because the successive minima take fewer evaluations to compute. This is presumably because each starting approximation is closer due to a smaller change being made to the λ parameters. Another feature of interest is that the correct active set for an inequality problem is usually established quickly by the λ iteration. Incidentally it is instructive that it is not worth trying to extrapolate these methods by estimating a starting value of x for $\Phi(\mathbf{x}, \boldsymbol{\lambda}^{(k+1)})$ from information taken at the solution $\mathbf{x}(\boldsymbol{\lambda}^{(k)})$. It is merely necessary to choose this starting value as $\mathbf{x}(\boldsymbol{\lambda}^{(k)})$, because the first step of the quasi-Newton method applied to $\Phi(\mathbf{x}, \boldsymbol{\lambda}^{(k+1)})$ will move $\mathbf{x}^{(1)}$ in the direction of the extrapolated minimum, assuming that an updated estimate of $\nabla^2\Phi(\mathbf{x}, \boldsymbol{\lambda}^{(k+1)})$ has been used.

In examining the problems individually, it is noticeable that when solving *TP*1 and *TP*7 the effect of estimating second derivatives of $\nabla^2\Phi$ for the first minimization leads to a particularly good number of evaluations for that minimization. The problem *TP*3(*A*) has a poor estimate of the likely change ΔF in $F(x)$ and so the σ_i are estimated very much on the small side. This causes a slow rate of convergence until larger values are obtained. The effect is particularly noticeable with the Newton-like iteration in (4.4). Increasing the initial σ_i by 10 (*TP*3(*B*)) improved matters considerably. However for problem *TP*7 the initial S is adequate to ensure convergence and here the advantage of the Newton-like iteration is most apparent. The problem *TP*2 is anomalous in that the Powell/Hestenes formula gives multipliers correct to three figures after one minimization. In view of the results on other problems it seems likely that some special effect may be at work, perhaps on account of *TP*2 being the dual of the linearly constrained problem *TP*1.

Overall the best method is that of algorithm (4.7). More detail is given about this algorithm in Table 5 which relates to how the individual λ_i and σ_i are changed on the difficult problem *TP*3(*A*). This problem is one with 8 two-sided constraints and if these are written in the order given by Colville (1968), then the lower bounds have been numbered 1–8 and the upper bounds 9–16. Thus the constraints 1, 2, 3, 9, 10, 11 are quadratic and the remainder are simple bounds. The active constraints A^* are $\{3, 4, 5, 9, 15\}$, but constraints 6, 8 and 16 are also involved on early iterations. The quantities $\lambda_i^{(k)}$, $\sigma_i^{(k)}$ and $\min(c_i^{(k)}, \theta_i^{(k)})$ are tabulated for these constraints. The initial choice of $\boldsymbol{\lambda}$ is zero, and of S is $40I$, and the constraint functions have all been scaled by dividing by 5. Apart from the ultimate rapid convergence of $\lambda_i^{(k)}$ to λ_i^* and $\min(c_i^{(k)}, \theta_i^{(k)})$ to zero, it is also interesting to see how the current active set changes on the early iterations, and how the multipliers are not necessarily changed when $\min(c_i^{(k)}, \theta_i^{(k)})$ is not zero (in contrast to the Powell/Hestenes iteration (3.9)). In fact the multipliers which are changed on any iteration are only those which are active in the subproblem (3.4). The changes in the $\sigma_i^{(k)}$ are also interesting in that the $\sigma_i^{(k)}$ are increased rapidly by (4.6) to begin with, and then remain constant. Also because the use of arbitrary factors of 10 is avoided, the $\sigma_i^{(k)}$ tend to be scaled amongst themselves rather better than with the other methods. This method never fails to reduce $K^{(k)}$ and the worst iteration over all the problems is the one in which $K^{(k)}$ is reduced from 0·15 to 0·083. A subroutine VF01A/AD which implements this method is available in the Harwell subroutine library and those interested should contact the subroutine librarian.

When comparing this method against other penalty or barrier functions it is found that the new function has a number of good properties which are not found together in any other penalty function. One of these is good conditioning of Φ due to the fact that no singularities are introduced in the penalty term, and that it is not necessary to make the parameters $\sigma_i \to \infty$ in order to force local convergence. Once S has been made sufficiently large, convergence of the λ iteration occurs at a rapid rate, and numerical experience suggests that high accuracy can be obtained in very few minimizations. Furthermore because the hessian matrix $\nabla^2\Phi$ can be carried forward from one iteration to the next, and updated when necessary, the computational effort required

TABLE 5

The algorithm (4.7) *applied to problem TP3(A)*

Con- \ k straint	1	2	3	4	5	6
3	0†	2008	252·9	806·8	809·31	809·425
	40	1877	3009	3009	3009	3009
	−0·789	0·134	−0·039	$-0{\cdot}18\times10^{-3}$	$-0{\cdot}76\times10^{-5}$	$-0{\cdot}10\times10^{-7}$
4	0	2·59	112·6	51·22	48·800	48·9275
	40	152·4	810·8	810·8	810·8	810·8
	−0·272	−0·062	0·016	$0{\cdot}57\times10^{-3}$	$-0{\cdot}31\times10^{-4}$	$0{\cdot}30\times10^{-7}$
5	0	0	659·6	75·65	84·337	84·3231
	40	40	863·3	863·3	863·3	863·3
	0	−0·060	0·013	$-0{\cdot}20\times10^{-3}$	$0{\cdot}32\times10^{-6}$	$0{\cdot}99\times10^{-8}$
	0	0	No			
6	40	160	further			
	−0·952	0	change			
8	0	0	No			
	40	160	further			
	−0·237	0	change			
9	0	0	255·1	417·9	402·23	403·270
	40	40	8689	8689	8689	8689
	0	−0·231	0·049	$0{\cdot}33\times10^{-3}$	$-0{\cdot}77\times10^{-6}$	$0{\cdot}17\times10^{-7}$
15	0	88·4	0	26·78	26·627	26·6392
	40	1953	6126	6126	6126	6126
	−0·033	$0{\cdot}51\times10^{-2}$	$-0{\cdot}87\times10^{-3}$	$0{\cdot}47\times10^{-5}$	$-0{\cdot}40\times10^{-6}$	$0{\cdot}26\times10^{-9}$
16	0	0	0	No		
	40	40	160	further		
	0	−0·153	0	change		

† Note: the three entries in each position are respectively $\lambda_i^{(k)}$, $\sigma_i^{(k)}$ and the resulting value of $\min(c_i^{(k)}, \theta_i^{(k)})$.

for the successive minimizations goes down rapidly. Most important of all for a penalty or barrier function is that it is very easy to program the method by incorporating an established quasi-Newton minimization routine into the program. With a barrier function, difficult decisions have to be taken about how to define the barrier function in the infeasible region, and it is not easy to avoid having to modify the minimization routine. Furthermore the linear search in the quasi-Newton subroutine is usually based on a cubic interpolation and is unsuitable for functions with singularities. In the Powell/Hestenes/Rockafellar penalty function however the function is defined for all **x** and the cubic linear search is also adequate. Finally there is no need to supply an initial feasible point to start off the whole process.

Osborne & Ryan (1972) give a method which also adapts the Powell/Hestenes penalty function to solve inequality constraint problems. Theirs is a hybrid method in which a barrier function is used to get an estimate of the likely active set so that the Powell/Hestenes function can be used, treating this set as equalities. They compare their method against more conventional types of barrier function on a number of problems including the problems *TP*1, 2, 3 used here. These results enable a general comparison amongst the penalty and barrier functions to be made. Osborne and Ryan work to an accuracy of 10^{-8}, so in comparing their results with those in this paper a small adjustment should be made. Assuming that 1 extra evaluation per minimization and also one extra minimization would be required on to the totals of Table 4 to achieve the slightly higher accuracy, the comparison is shown in Table 6. These results show a measurable bias in favour of the new penalty function.

TABLE 6

Comparison of penalty and barrier functions

Problem	Newton-like method (4.7)	Osborne & Ryan	Barrier function	Extrapolated B.F.
*TP*1	47	167	225	177
*TP*2	172	229	440	245
*TP*3(*B*)	73	107	173	123

So far the emphasis has been on the advantages of the new penalty function and it is advisable to consider what the disadvantages are if any. One possible disadvantage is that the presence of discontinuities in the second derivative of the penalty function might cause slow convergence of the quasi-Newton subroutine. An experiment has been conducted to test this hypothesis. The first minimization of $\Phi(\mathbf{x}, \boldsymbol{\theta}, S)$ for *TP*2 was repeated, designating the constraints known to be active at the minimum as equalities. This removes the discontinuities for these constraints and should lead to faster convergence under the hypothesis. In fact three more evaluations were required. The run in which the discontinuities were present was also checked to see whether the discontinuities were active in the sense that points either side of them were being taken, and this was certainly true. Also the results from the other test problems are by no means unduly large for the size of problem involved. Therefore I have no evidence to support the hypothesis that these discontinuities at all retard convergence.

Another possible disadvantage of the penalty function is that if S is not chosen large enough, the local minimum of $\Phi(\mathbf{x}, \boldsymbol{\theta}, S)$ at $\mathbf{x}^*$ may not exist, and even if it does, Φ may be unbounded below elsewhere (for example, minimize $-x^3$ subject to $x \leqslant 1$, for which $\Phi \sim -x^3+\frac{1}{2}\sigma x^2$ for large x). However there are various ways to get round this, for instance by increasing S, or by replacing $F(\mathbf{x})$ by $\exp(F(\mathbf{x}))$. A related disadvantage is that the method does not appear to handle problems with inequality constraints which cut out regions in which $F(\mathbf{x})$ is not defined. However I have yet to see a problem for which it is not possible to define a smooth continuation of $F(\mathbf{x})$ into the infeasible region. Such a device enables the method to be applied to this type of problem. One related advantage of the method is that it is not necessary for the initial $\mathbf{x}$ approximation to be feasible. However it is interesting to consider what happens when the problem has no feasible point. In this case $K^{(k)}$ is monotonic decreasing but not to zero. Therefore in the equality problem there will be subsequences of $\{k\}$ for which $\mathbf{c}(\boldsymbol{\lambda}^{(k)}, S^{(k)}) \to \mathbf{c}'$ where $\mathbf{c}'$ is a fixed vector, and by Theorem 3 $\mathbf{x}(\boldsymbol{\lambda}^{(k)}, S^{(k)})$ will tend to the point $\mathbf{x}'$ which minimizes $F(\mathbf{x})$ subject to $\mathbf{c}(\mathbf{x}) = \mathbf{c}'$. If in fact $\mathbf{c}^{(k)} \to \mathbf{c}'$ for the sequence $\{k\}$ itself, then it follows in addition that $\mathbf{x}(\boldsymbol{\lambda}^{(k)}, S^{(k)}) \to \mathbf{x}'$.

A further disadvantage of the penalty function, common to many approaches to the non-linear programming problem, is that the theory given here calls for an exact local minimum of $\Phi(\mathbf{x}, \boldsymbol{\theta}, S)$ with respect to $\mathbf{x}$ to be found. Of course in general this cannot be done in a finite number of operations. Now it would be expected that if the theory were modified to remove this necessity, then to get an equivalent rate of convergence would require increasing accuracy in $\mathbf{x}(\boldsymbol{\lambda}, S)$ to be obtained as $\boldsymbol{\lambda} \to \boldsymbol{\lambda}^*$. In practice however most of the effort in unconstrained minimization goes into locating the neighbourhood of the minimum in which fast convergence occurs, and very little into achieving that fast convergence. Thus it costs little in practice to assume that high accuracy in all local minima is obtained, and in VF01A/AD the same accuracy required in the final solution is asked for in all local minima.

In view of this discussion, my opinion is that the disadvantages of the Powell/Hestenes/Rockafellar penalty function are negligible, and that the advantages are strong, especially the lack of numerical difficulties and the ease of using the unconstrained minimization routine. Nonetheless to solve a non-linear problem by transforming it to a sequence of non-linear problems should not be optimum, and I think that ultimately algorithms which vary both $\mathbf{x}$ and $\boldsymbol{\lambda}$ together will prove superior. This is already true as regards local convergence as some results about Solver-like methods which will appear in the thesis of Jackson (1974) show. However a good way of forcing global convergence for such methods is not yet clear.

Finally some possible extensions of the idea are discussed, and in particular the situation *vis-à-vis* problems in which some of the inequality constraints are linear. For small to medium problems there is not much to be gained by trying to take any special account of this feature, other than to use the fact that ∇c is constant and can be set in advance. However for large problems with many active linear and possibly sparse constraints, it is worthwhile looking to construct a penalty function from the non-linear constraints only, to be minimized by a method which maintains feasibility with regard to the linear constraints. In this case it will be important for maximum efficiency to choose a method such as that of Buckley (1973) which only keeps a second derivative approximation in an n–p dimensional space where p is the number of active linear constraints.

Another interesting modification of the method is for problems with two-sided constraints like $a_i \leqslant c_i(\mathbf{x}) \leqslant b_i$. At the moment these must be written as two separate constraints. However this is wasteful in storage space and also there are some implicit restrictions on the $\boldsymbol{\theta}$ parameters which are not taken account of when separating the two inequalities. A method which keeps the constraints together would be preferable, and would just be an extension of the current method in respect of the way equality constraints are treated, for these are two-sided constraints for which $a_i = b_i$.

References

Arrow, K. J., Gould, F. J. & Howe, S. M. 1973 *Math. Programming* **5**, 225–234.

Buckley, A. 1973 *AERE Report TP* 544.

Buys, J. D. 1972 *Dual algorithms for constrained optimization problems*. Ph.D. Thesis, University of Leiden.

Colville, A. R. 1968 *IBM New York Scientific Center Report* 320–2949.

Fiacco, A. V. & McCormick, G. P. 1968 *Non-linear programming: sequential unconstrained minimization techniques*. New York: Wiley.

Fletcher, R. & Jackson, M. P. 1974 *J. Inst. Maths Applics* **14**, 159–174.

Hestenes, M. R. 1966 *Calculus of variations and optimal control problems*. New York: Wiley.

Hestenes, M. R. 1969 *J. Opt. Theory, Applics* **4**, 303–320.

Householder, A. S. 1974 *The theory of matrices in numerical analysis*. New York: Blaisdell.

Jackson, M. P. 1974 *Ph.D. Thesis*. University of Oxford.

Mangasarian, O. L. 1973 *Comp. Sci. Tech. Report* 174. University of Wisconsin.

Osborne, M. R. & Ryan, D. M. 1972 In *Numerical methods for non-linear optimization* (ed. F. A. Lootsma). London: Academic Press. (See chapter 28.)

Powell, M. J. D. 1969 In *Optimization* (ed. R. Fletcher) London: Academic Press. (See chapter 19.)

Rockafellar, R. T. 1973*a* *J. Opt. Theory Applics* **12**, 553–562.

Rockafellar, R. T. 1973*b* *Math. Programming* **5**, 354–373.

Rockafellar, R. T. 1974 *S.I.A.M. J. Control* **12**, 268–285.

Rosen, J. B. & Suzuki, S. 1965 *Communs ACM* **8**, 113.

Rosenbrock, H. H. 1960 *Comput. J.* **3**, 175–184.