Mathematical Programming 5 (1973) 354–373. North-Holland Publishing Company

A DUAL APPROACH TO SOLVING NONLINEAR PROGRAMMING PROBLEMS BY UNCONSTRAINED OPTIMIZATION

R. Tyrrell ROCKAFELLAR *
University of Washington, Seattle, Washington, U.S.A.

Received 4 January 1973
Revised manuscript received 13 July 1973

Several recent algorithms for solving nonlinear programming problems with equality constraints have made use of an augmented "penalty" Lagrangian function, where terms involving squares of the constraint functions are added to the ordinary Lagrangian. In this paper, the corresponding penalty Lagrangian for problems with inequality constraints is described, and its relationship with the theory of duality is examined. In the convex case, the modified dual problem consists of maximizing a differentiable concave function (indirectly defined) subject to no constraints at all. It is shown that any maximizing sequence for the dual can be made to yield, in a general way, an asymptotically minimizing sequence for the primal which typically converges at least as rapidly.

1. Introduction

One of the chief methods of solving a problem of the form

(P) minimize $f_0(x)$ over $x \in X$,
subject to $f_i(x) \leqslant 0, \quad i = 1, \ldots, m,$

is to introduce penalties so as to convert it into a sequence of unconstrained (or less constrained) problems, for instance:

$$\text{minimize } f_0(x) + r \sum_{i=1}^{m} \theta(f_i(x))^2 \quad \text{over } x \in X, \tag{1.1}$$

* Supported in part by the Air Force Office of Scientific Research under grant AF-AFOSR-72-2269.

where $r \to \infty$,

$$\theta(t) = \max\{0, t\} \,. \tag{1.2}$$

This might be regarded as a "primal" approach to getting rid of constraints, although certain connections with duality are known [3]. The classical "dual" approach, based on the Kuhn–Tucker theory of Lagrange multipliers, replaces (P) by a sequence of problems of the form

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} y_i f_i(x) \quad \text{over } x \in X \,, \tag{1.3}$$

where the corresponding multiplier vectors y constitute a maximizing sequence for the ordinary dual problem (see problem (D_0) in Section 2).

The first approach suffers from well-known numerical instabilities as $r \to +\infty$. The second approach, on the other hand, while of importance in the decomposition of large-scale problems having separable functions f_i, is limited intrinsically to the convex case, and it involves dual constraints which can sometimes be awkward to handle: besides the linear constraints $y_i \geqslant 0$, the multipliers must be chosen so that the minimum in (1.3) exists, at least if the theoretical minimization step is not replaced by something more complicated. Thus the reduction to unconstrained optimization is not really complete. Furthermore, there is the difficulty that, unless f_0 is *strictly* convex, the x sequence generated by solving the sequence of problems (1.3) is not necessarily a minimizing sequence for the primal problem.

In the case of equality constraints, it has been observed that some of these difficulties can be obviated by a combined approach, where one solves a sequence of problems of the form

$$\text{minimize } f_0(x) + r \sum_{i=1}^{m} f_i(x)^2 + \sum_{i=1}^{m} y_i f_i(x) \quad \text{over } x \in X.$$

Often a minimizing sequence for the primal can be generated in this way without passing to ever larger values of r. This idea is due independently to Hestenes [9, 10], and Powell [16]; see also [8]. The Lagrangian expression in (1.4) was also introduced some years ago by Arrow and Solow [1], but in a different algorithmic context, where differential equations were used to locate saddle points. In the latter approach, in effect, x and the multipliers are modified continuously, rather than alternately. Related ideas of Fletcher [4, 5], stemming from Powell's work, involve substituting certain continuous functions of x for the multipliers in (1.4), so that the sequence is coalesced into a single unconstrained problem replacing the constrained primal. Intermediate methods, where at each stage only one step or cycle of an algorithm for unconstrained problems is applied to (1.4) before the multipliers are changed in some way, have been explored computationally by Miele and associates [12, 13, 14, 15]. Theoretical questions about the latter methods remain largely unanswered, however.

Powell and Fletcher note the desirability of extending their algorithms to cover inequality constraints. They speak of doing this essentially by first determining (somehow) which constraints are active and then proceeding as if one had equality constraints. Arrow and Solow treated inequalities by introducing nonnegative slack variables to convert them into equations. Thus, in their somewhat different context, certain linear inequality constraints are represented in the choice of the set X.

None of these authors has discussed the role that convexity might play, or the implications that this could have for various algorithms. The present paper is devoted mainly to such questions, partly on the grounds that a thorough understanding of the convex case is fundamental for the treatment of inequality-constrained problems. Applications to nonconvex problems will be described elsewhere.

2. The penalty Lagrangian

A natural generalization of (1.4) to constraints $f_i(x) \leq 0$ might be to replace the term $f_i(x)^2$ by $\theta(f_i(x))^2$. This would indeed preserve convexity, if present. In effect, one would be using the ordinary Lagrangian approach to solve (P), but with the objective $f_0(x)$ replaced by $f_0(x) + r\,\Sigma_{i=1}^{m}\theta(f_i(x))^2$, which has the same values on the set of feasible solutions. There would be a definite drawback, however: the modified objective function would generally have discontinuous second derivatives at an optimal solution, even if the functions f_i were highly differentiable. This would lead to unnecessary computational difficulties. Note also that the nonnegativity constraints on the dual variables y_i would still be present.

We adopt instead the following generalization of (2.1):

$$\text{minimize } L_r(x,y) \quad \text{over } x \in X\,, \tag{2.2}$$

where

$$L_r(x,y) = f_0(x) + (1/4r)\sum_{i=1}^{m} [\theta(y_i + 2rf_i(x))^2 - y_i^2]\,. \tag{2.3}$$

Thus the expression $r\theta(f_i(x))^2 + y_i f_i(x)$, $y_i \geqslant 0$, corresponding to the approach of the preceding paragraph, is altered to

$$(1/4r)[\theta(y_i + 2rf_i(x))^2 - y_i^2] = \begin{cases} rf_i(x)^2 + y_i f_i(x) & \text{if } f_i(x) \geqslant -y_i/2r\,, \\ -y_i^2/4r & \text{if } f_i(x) \leqslant -y_i/2r\,, \end{cases} \tag{2.4}$$

with no restriction placed on the sign of y_i.

The function L_r on $X \times \mathbf{R}^m$ will be called the *penalty Lagrangian* associated with problem (P) and the (positive) parameter value r. Its virtues, despite its peculiar appearance, will be apparent below. In particular, limiting ourselves for present purposes to the convex case, we shall demonstrate that an asymptotically minimizing sequence for (P) with reasonable convergence properties can be generated by attacking a sequence of subproblems (2.2) for arbitrary fixed r, without necessarily moving more than a part of the way toward the solution of each subproblem. The subproblems are likely to be tractable, in the sense that the gradient of $L_r(x, y)$ with respect to x can be expressed simply in terms of the gradients of the functions f_i, if, for example, X is a region in $\mathbf{R}^n$ on which every f_i is differentiable. Second derivatives are likely to be better behaved in this case than in the alternative one mentioned above, as will be seen in Section 5.

The penalty Lagrangian L_r was introduced by the author in [21], where some of its properties were mentioned without proof. Recently, Arrow, Gould and Howe [2] have studied its saddle point properties in the absence of convexity, but under the restriction that $y_i \geqslant 0$. This restriction is not made here.

The formula for L_r is not as mysterious as it might seem. It can be derived by representing inequality constraints as equations in the manner of Arrow and Solow [1]. Namely, if we regard (P) as the problem of minimizing $f_0(x)$ subject to $f_i(x) + z_i = 0$ for $i = 1, \ldots, m$, where $(x, z) \in X \times \mathbf{R}^m_+$, the corresponding problem (1.4) consists of minimizing

$$f_0(x) + r \sum_{i=1}^{m} (f_i(x) + z_i)^2 + \sum_{i=1}^{m} y_i(f_i(x) + z_i) \tag{2.5}$$

over $(x, z) \in X \times \mathbf{R}_+^m$. The minimization in z can be carried out explicitly, and the residual problem is then (2.2). This derivation, since it appears to do violence to possible convexity properties of the constraint functions, hardly suggests the strong properties which L_r turns out to have for convex programming.

Of course, in problems with mixed inequality and equality constraints, the equations can be expressed as pairs of inequalities, or terms like those in (1.4) can be included in (2.2). This generalization is elementary, and so to keep the notation simpler we do not carry it out explicitly in the present discussion.

3. Convexity properties and duality

Henceforth we assume that X is a (nonempty) convex subset of a real linear space E, and that the functions $f_i : X \to \mathbf{R}$, $i = 0, 1, ..., m$, are convex. We observe that

$$\lim_{r \downarrow 0} L_r(x, y) = L_0(x, y) \qquad \text{for all } x \in X,\ y \in \mathbf{R}^m , \tag{3.1}$$

where L_0 is the ordinary Lagrangian associated with (P):

$$L_0(x, y) = \begin{cases} f_0(x) + \sum_{i=1}^{m} y_i f_i(x) & \text{if } y \geqslant 0, \\ -\infty & \text{if } y \not\geqslant 0 \ \ (y = (y_1, ..., y_m)). \end{cases} \tag{3.2}$$

According to the perturbational theory of duality (cf. [19]), L_0 corresponds to the embedding of (P) in the class of problems where for each $u = (u_1, ..., u_m) \in \mathbf{R}^m$ the function f_i is perturbed to $f_i - u_i$, $i = 1, ..., m$. The following result shows that, for $r > 0$, L_r too is a Lagrangian in the sense of the theory, and that it arises in the same way, except that f_0 is simultaneously perturbed to $f_0 + r|u|^2$.

Theorem 3.1. *For any* $r \geqslant 0$, *one has*

$$L_r(x,y) = \min \{F_r(x,u) + u \cdot y : u \in \mathbf{R}^m\}, \quad x \in X, \tag{3.3}$$

where F_r is the convex function on $X \times \mathbf{R}^m$ defined by

$$F_r(x,u) = \begin{cases} f_0(x) + r \sum_{i=1}^{m} u_i^2 & \text{if } u_i \geqslant f_i(x), \quad i = 1, \ldots, m, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.4}$$

Therefore $L_r(x,y)$ is convex in $x \in X$ and concave in $y \in Y$.

Proof. The first assertion is simple to verify. In fact, it corresponds to the observation at the end of Section 2, with $u_i = z_i + f_i(x)$. The second assertion then follows from general duality theory, cf. [19]; indeed, the partial conjugate of a convex function is always a saddle function [19, Theorem 33.1].

The *dual* of (P) corresponding to L_r is

$$(\mathrm{D}_r) \qquad \text{maximize } g_r(y) \overset{\mathrm{d}}{=} \inf_{x \in X} L_r(x,y) \quad \text{over all } y \in \mathbf{R}^m .$$

Theorem 3.2. *For every $r > 0$, the function g_r is concave and satisfies*

$$g_r(y) = \max_{z \in \mathbf{R}^m} \{g_0(z) - (1/4r)\,|z - y|^2\} . \tag{3.5}$$

Thus the dual problems (D_r) all have the same optimal solutions and supremum as the ordinary dual (D_0). Moreover (assuming $g_0 \not\equiv -\infty$), g_r is everywhere finite and continuously differentiable on $\mathbf{R}^m$. Specifically, if for a given y the infimum defining $g_r(y)$ happens to be attained at a point x (not necessarily unique), then

$$\partial g_r(y)/\partial y_i = \partial L_r(x,y)/\partial y_i \tag{3.6}$$

$$= [\theta(y_i + 2rf_i(x)) - y_i]/2r = \max\{-y_i/2r, f_i(x)\} .$$

Proof. Let

$$p_r(u) \overset{\mathrm{d}}{=} \inf_{x \in X} F_r(u,x) , \quad r \geqslant 0 . \tag{3.7}$$

Thus p_r is the "perturbation" function corresponding to the class of perturbations of (P) giving rise to L_r and (D_r). It is a convex function, since F_r is convex. Indeed,

$$p_r = p_0 + 2rq, \quad \text{where} \quad q(u) = \tfrac{1}{2}|u|^2 . \tag{3.8}$$

$$g_r(y') \leqslant L_r(x, y') \leqslant L_r(x, y) + (y' - y) \cdot \nabla_y L_r(x, y)$$

$$= g_r(y) + (y' - y) \cdot \nabla_y L_r(x, y) \qquad (3.15)$$

for all $y' \in \mathbf{R}^m$. This says that $\nabla_y L_r(x, y) \in \partial g_r(y)$, and therefore $\nabla_y L_r(x, y) = \nabla g_r(y)$ as claimed.

Corollary 3.3. *The function* g_r, *unless it is identically* $-\infty$, *satisfies for all* y *and* y':

$$g_r(y) + (y' - y) \cdot \nabla g_r(y) \geqslant g_r(y')$$

$$\geqslant g_r(y) + (y' - y) \cdot \nabla g_r(y) - (1/4r)\, |y' - y|^2 \; . \qquad (3.16)$$

Proof. The first inequality is immediate from concavity. For the second inequality, we note from (3.5) the existence, for any given y, of a quadratic function of the form

$$h(y') = g_0(z) - (1/4r)\, |y' - z|^2$$

satisfying $h(y) = g_r(y)$, while $h(y') \leqslant g_r(y')$ for all y'. The two properties imply $\nabla h(y) = \nabla g_r(y)$. But since h is quadratic, we have

$$h(y') = h(y) + (y' - y) \cdot \nabla h(y) - (1/4r)\, |y' - y|^2 \; .$$

Thus $h(y')$ equals the expression on the right side of (2.3).

In view of Theorem 3.2, we shall in the rest of this paper refer simply to *dual optimal solutions* and the *dual optimal value,* since these are independent of r. Note that a dual optimal solution is necessarily a *nonnegative* vector, since $g_0(y) = -\infty$ if $y \ngeqslant 0$. (By convention, dual optimal solutions are not said to exist when the functions g_r are all identically $-\infty$.) The dual optimal value is, of course, generally less than or equal to the primal optimal value (i.e., the infimum in (P)). If they are equal, we say (P) is *normal.*

A *Kuhn–Tucker* vector for (P) relative to the Lagrangian L_r is a vector $\bar{y}$ such that

$$-\infty < \inf_{x \in X} L_r(x, \bar{y}) = \inf \;\; \text{in (P)}. \qquad (3.17)$$

This condition on $\bar{y}$ is known to hold if and only if $\bar{y}$ is an optimal

Substituting (3.3) into the definition of g_r, we therefore have

$$g_r(y) = \inf_{u \in \mathbf{R}^m} \{p_r(u) + u \cdot y\} = -p_r^*(-y) . \tag{3.9}$$

A basic formula for the conjugate of a sum of convex functions [19, Theorem 16.4] yields

$$\begin{aligned} -p_r^*(-y) &= -(p_0 + 2rq)^*(-y) = -(p_0^* \,\square\, q^*2r)(-y) \\ &= -\min_{z \in \mathbf{R}^m} \{p_0^*(-z) + 2rq^*((y-z)/2r)\} \end{aligned} \tag{3.10}$$

and consequently, since $q^* = q$,

$$g_r(y) = \max_{z \in \mathbf{R}^m} \{g_0(z) - 2rq((y-z)/2r)\} , \tag{3.11}$$

which is the same as (3.5). In particular, g_r is finite everywhere if $g_0 \not\equiv -\infty$, and hence the subgradient set $\partial g_r(y)$ is always nonempty. If $w \in \partial g_r(y)$, that is,

$$g_r(y') \leqslant g_r(y) + w \cdot (y' - y) \quad \text{for all } y' \in \mathbf{R}^m , \tag{3.12}$$

then for the point z at which the maximum in (3.5) is attained (unique by strict concavity) we have also

$$\begin{aligned} &g_0(z) - (1/4r)\,|z - y'|^2 \leqslant \\ &\leqslant g_0(z) - (1/4r)\,|z - y|^2 + w \cdot (y' - y) \quad \text{for all } y' \in \mathbf{R}^m , \end{aligned} \tag{3.13}$$

in other words,

$$(1/4r)\,|z - y'|^2 + wy' \geqslant (1/4r)\,|z - y|^2 + wy \quad \text{for all } y' \in \mathbf{R}^m , \tag{3.14}$$

implying $w = (z - y)/2r$. Thus $\partial g_r(y)$ consists of a single vector for each y, and we may conclude, since g_r is concave, that this vector is the gradient [19, Theorem 25.1] and depends continuously on y [19, Theorem 25.5]. Now suppose the infimum defining $g_r(y)$ is attained at a certain x: $L_r(x, y) = g_r(y)$. Since L_r is concave as well as differentiable in the second argument, we have

$$g_r(y') \leqslant L_r(x, y') \leqslant L_r(x, y) + (y'-y) \cdot \nabla_y L_r(x, y)$$

$$= g_r(y) + (y'-y) \cdot \nabla_y L_r(x, y) \qquad (3.15)$$

for all $y' \in \mathbf{R}^m$. This says that $\nabla_y L_r(x, y) \in \partial g_r(y)$, and therefore $\nabla_y L_r(x, y) = \nabla g_r(y)$ as claimed.

Corollary 3.3. *The function* g_r, *unless it is identically* $-\infty$, *satisfies for all* y *and* y':

$$g_r(y) + (y'-y) \cdot \nabla g_r(y) \geqslant g_r(y')$$

$$\geqslant g_r(y) + (y'-y) \cdot \nabla g_r(y) - (1/4r)\, |y'-y|^2 \,. \qquad (3.16)$$

Proof. The first inequality is immediate from concavity. For the second inequality, we note from (3.5) the existence, for any given y, of a quadratic function of the form

$$h(y') = g_0(z) - (1/4r)\, |y'-z|^2$$

satisfying $h(y) = g_r(y)$, while $h(y') \leqslant g_r(y')$ for all y'. The two properties imply $\nabla h(y) = \nabla g_r(y)$. But since h is quadratic, we have

$$h(y') = h(y) + (y'-y) \cdot \nabla h(y) - (1/4r)\, |y'-y|^2 \,.$$

Thus $h(y')$ equals the expression on the right side of (2.3).

In view of Theorem 3.2, we shall in the rest of this paper refer simply to *dual optimal solutions* and the *dual optimal value*, since these are independent of r. Note that a dual optimal solution is necessarily a *nonnegative* vector, since $g_0(y) = -\infty$ if $y \ngeqslant 0$. (By convention, dual optimal solutions are not said to exist when the functions g_r are all identically $-\infty$.) The dual optimal value is, of course, generally less than or equal to the primal optimal value (i.e., the infimum in (P)). If they are equal, we say (P) is *normal*.

A *Kuhn–Tucker* vector for (P) relative to the Lagrangian L_r is a vector $\bar{y}$ such that

$$-\infty < \inf_{x \in X} L_r(x, \bar{y}) = \inf \quad \text{in (P)}. \qquad (3.17)$$

This condition on $\bar{y}$ is known to hold if and only if $\bar{y}$ is an optimal solution to (D_r) and (P) is normal [19, p. 317–318]. It follows that this concept is likewise independent of r. Moreover, a pair $(\bar{x}, \bar{y})$ is a saddle point for L_r if and only if $\bar{x}$ is an optimal solution to (P) and $\bar{y}$ is a Kuhn–Tucker vector [19, p. 319]. Summarizing, we therefore can state:

Corollary 3.4. *Relative to the Lagrangians* L_r, $r \geq 0$, *one has the same Kuhn–Tucker vectors and saddle points. Thus* $(\bar{x}, \bar{y})$ *is a saddle point of* L_r *if and only if the ordinary Kuhn–Tucker conditions are satisfied:*

(i) $\bar{y}_i \geq 0$, $f_i(\bar{x}) \leq 0$, $\bar{y}_i f_i(\bar{x}) = 0$ *for* $i = 1, \ldots, m$;

(ii) $\bar{x}$ *minimizes* $f_0 + \Sigma_{i=1}^m \bar{y}_i f_i$ *over* X.

Conditions for normality and the existence of Kuhn–Tucker vectors are given for example in [6, 7, 11, 12, 18, 19, 20]. In particular, normality holds if a Kuhn–Tucker vector exists, or if the following compactness condition is satisfied (in some locally convex topology): X is closed, every f_i is lower semicontinuous, and the set

$$\{x \in X: f_0(x) \leq \alpha,\ f_1(x) \leq \epsilon, \ldots, f_m(x) \leq \epsilon\}$$

is (weakly) compact for some real α greater than the infimum in (P) and some $\epsilon > 0$. A Kuhn–Tucker vector exists if, say, there is a strictly feasible solution to (P) (i.e., the Slater condition is satisfied) and the infimum in (P) is not $-\infty$; this can be weakened when some of the constraints are linear.

Corollary 3.4 emphasizes properties which L_r for $r > 0$ has in common with the ordinary Lagrangian L_0. The next result, however, gives a strong property whose well-known absence for L_0 has been a serious impediment to computational approaches based on duality.

Theorem 3.5. *Assume that* (P) *is normal, and that* $\bar{y}$ *is an arbitrary dual optimal solution. Let* $r > 0$. *Then* $\bar{x}$ *is an optimal solution to* (P) *if and only if* $\bar{x}$ *minimizes* $L_r(x, \bar{y})$ *over all* $x \in X$.

Proof. The hypothesis implies that $\bar{y}$ is a Kuhn–Tucker vector, as remarked above. Hence $\bar{x}$ is an optimal solution to (P) if and only if $(\bar{x}, \bar{y})$ is a saddle point of L_r. Certainly, the latter implies that $\bar{x}$ minimizes $L_r(\cdot, \bar{y})$ over X, so the "only if" part of the theorem is true. For the other part, we observe from Theorem 3.2 that if $\bar{x}$ minimizes $L_r(\cdot, \bar{y})$ over X, we have $\nabla_y L_r(\bar{x}, \bar{y}) = \nabla g_r(\bar{y}) = 0$. Thus $\bar{y}$ maximizes the (concave) function $L_r(\bar{x}, y)$ over all $y \in \mathbf{R}^m$. This shows that $(\bar{x}, \bar{y})$ is a saddle point of L_r, and consequently, $\bar{x}$ is an optimal solution.

The reason why the characterization in Theorem 3.5 fails for $r = 0$ is that the set of points where $L_0(\cdot, \bar{y})$ attains its minimum may include besides the optimal solutions to (P), various points which are not even feasible.

4. Computational reduction to unconstrained optimization

Theorems 3.2 and 3.5 yield a dual method of computation of optimal solutions to (P), which may be stated as follows in its ideal form, assuming that (P) possesses an optimal solution and a Kuhn–Tucker vector (see above). Fixing $r > 0$, we somehow determine one of the points $\bar{y}$ at which the continuously differentiable concave function g_r attains its maximum over $\mathbf{R}^m$ (no constraints). Next, we determine any point $\bar{x}$ minimizing the convex function $L_r(\cdot, \bar{y})$ over X. (This is a nice kind of unconstrained problem if, for example, $X = \mathbf{R}^n$ and the functions f_i are differentiable.) Then $\bar{x}$ is an optimal solution to (P), and $\bar{y}$ is a vector of multipliers satisfying, together with $\bar{x}$, the Kuhn–Tucker conditions.

The essential difficulty in this ideal scheme is the fact that, computationally speaking, the maximization and minimization cannot be carried out exactly. But if $\bar{y}$ and $\bar{x}$ are determined only approximately as described, can we be sure that $\bar{x}$ is "approximately" an optimal solution to (P)?

Another complication arises because the function g_r to be maximized is not expressed as "directly" as might be desired. While it is true we can calculate $g_r(y)$ and $\nabla g_r(y)$ for any y by determining an x which minimizes $L_r(\cdot, y)$ over X (cf. Theorem 3.2), this is not a "cheap" operation. It precludes the application to g_r of an algorithm which requires extremely many evaluations, for example, in order to maximize g_r along a line segment. Furthermore, we are again faced with our inability to determine a minimizing x (or even the minimum value) exactly.

However, this situation is not so hopeless as might be thought. Suppose for a given y that we have an $x \in X$ and an $\alpha \geqslant 0$ such that x minimizes $L_r(\cdot, y)$ over X to within α, that is,

$$L_r(x, y) - g_r(y) \leqslant \alpha \,. \tag{4.1}$$

Then generalizing slightly the argument at the end of the proof of Theorem 3.2, we see that the affine function

$$h(y') = L_r(x, y) + (y' - y) \cdot \nabla_y L(x, y) \tag{4.2}$$

satisfies

$$g_r(y') \leqslant L_r(x, y') \leqslant h(y') \quad \text{for all } y' \in \mathbf{R}^m \,, \tag{4.3}$$

$$g_r(y) \leqslant L_r(x, y) = h(y) \leqslant g_r(y) + \alpha \,. \tag{4.4}$$

(In particular, $\nabla_y L(x, y)$ is a so-called α-subgradient of g_r at y.) What is needed is an algorithm to maximize g_r which efficiently uses information of just this type, taking advantage of such theoretical properties as follow from Theorems 3.1 and 3.2.

A cutting plane algorithm could be used, for instance, but other procedures which take advantage of the "curvature" of the functions $L_r(x, \cdot)$ which majorize g_r are also possible. A procedure corresponding to a generalization of a method of Hestenes and Powell will be described elsewhere [21]. In this paper, we focus attention not on a particular algorithm, but on the general way that a method based on such information can be used to solve (P).

By a *maximizing sequence* for the dual problem (D_r), we shall mean of course a sequence $\{y^k\}$ in $\mathbf{R}^m$ such that $g_r(y^k) \to \sup g_r$. A sequence $\{x^k\}$ in X will be called *asymptotically feasible* for (P) if

$$\limsup_{k \to \infty} \{f_i(x^k)\} \leqslant 0 \,, \qquad i = 1, \ldots, m \,. \tag{4.5}$$

The infimum of the quantity $\limsup_{k \to \infty} \{f_0(x^k)\}$ over all such sequences is the *asymptotic optimal value* in (P), and an asymptotically feasible sequence $\{x^k\}$ for which this infimum is attained is an *asymptotically minimizing* sequence for (P). As is well known, the asymptotic optimal value in (P) equals the dual optimal value if the latter is not $-\infty$, or if there exist asymptotically feasible sequences at all (e.g., [7, 11, 19]). Thus, in particular, if (P) is normal and possesses feasible solutions, a sequence $\{x_k\}$ in X is asymptotically minimizing if and only if (4.5) holds and

$$\lim_{k \to \infty} f_0(x^k) = \text{optimal value in (P)}. \tag{4.6}$$

If, more specifically, (P) satisfies the compactness condition sufficient for normality, which was mentioned following Corollary 3.4, it is obvious that an asymptotically minimizing sequence $\{x_k\}$ must be relatively (weakly) compact, with all of its (weak) cluster points actually optimal solutions to (P).

Theorem 4.1. *Suppose, the asymptotic optimal value in* (P) *is finite. Let* $\{y^k\}$ *be a bounded maximizing sequence for* (D_r), $r > 0$, *and for each* k, *let* $x^k \in X$ *satisfy*

$$L_r(x^k, y^k) - \inf L_r(\cdot, y^k) = L_r(x^k, y^k) - g_r(y^k) \leqslant \alpha_k \ , \quad (4.7)$$

where α_k *tends to* 0. *Then* $\{x^k\}$ *is an asymptotically minimizing sequence for* (P).

The proof of Theorem 4.1 involves the following estimates.

Lemma 4.2. $r|\nabla g_r(y^k)|^2 \leqslant (\sup g_r) - g_r(y_k)$.

Proof. From Corollary 3.3, we have

$$\begin{aligned}\sup g_r &\geqslant \max_{y' \in \mathbf{R}^m} \{g_r(y^k) + (y' - y^k)\cdot \nabla g_r(y^k) - (1/4r)\,|y' - y^k|^2\} \\ &= g_r(y^k) + \max_{u \in \mathbf{R}^m} \{u \cdot \nabla g_r(y^k) - (1/4r)|u|^2\} \ ,\end{aligned}$$

and the latter maximum turns out to be $r|\nabla g_r(y^k)|^2$.

Lemma 4.3. *Condition* (4.7) *implies*

$$r|\nabla_y L_r(x^k, y^k) - \nabla g_r(y^k)|^2 \leqslant \alpha_k \ .$$

Proof. Using Corollary 3.3 and the convexity of $L_r(x^k, \cdot)$, we have for every $w \in \mathbf{R}^m$

$$L_r(x^k, y^k) + (w - y^k)\cdot \nabla_y L_r(x^k, y^k) \geqslant L_r(x^k, w) \geqslant g_r(w)$$

$$\geqslant g_r(y^k) + (w - y^k)\cdot \nabla g_r(y^k) - (1/4r)\,|w - y^k|^2 \ ,$$

and hence

$$L_r(x^k, y^k) - g_r(y^k) \geq \sup_{u \in \mathbf{R}^m} \{u \cdot (\nabla g_r(y^k) - \nabla_y L_r(x^k, y^k)) - (1/4r)|u|^2\}$$

$$= r|\nabla g_r(y^k) - \nabla_y L_r(x^k, y^k)|^2 . \tag{4.8}$$

Proof of Theorem 4.1. Theorem 3.1 asserts that

$$L_r(x^k, y) = \min_{u \in \mathbf{R}^m} \{F_0(x^k, u) + u \cdot y + r|u|^2\} , \tag{4.9}$$

where F_0 is given by (3.4) for $r = 0$. For $y = y^k$, the minimum is attained at a unique point which we shall denote by u^k. We thus have

$$L_r(x^k, y) \leq F_0(x^k, u^k) + u^k \cdot y + r|u^k|^2 \text{ for all } y \in \mathbf{R}^m , \tag{4.10}$$

where equality holds when $y = y^k$. This implies

$$u^k = \nabla_y L_r(x^k, y^k) , \tag{4.11}$$

and therefore $u^k \to 0$ by Lemmas 4.2 and 4.3. Since by hypothesis

$$\lim_{k\to\infty} L_r(x^k, y^k) = \lim_{k\to\infty} g_r(y^k) = \sup g_r , \tag{4.12}$$

while the sequence $\{y^k\}$ is bounded, we see that

$$\lim_{k\to\infty} F_0(x^k, u^k) = \lim_{k\to\infty} \{L_r(x^k, y^k) - u^k \cdot y^k - r|u^k|^2\}$$

$$= \text{dual optimal value.} \tag{4.13}$$

In view of the definition of F_0 and the remarks preceding the theorem, this means that $\{x^k\}$ is an asymptotically minimizing sequence.

The next result sharpens a special case of Theorem 4.1.

Theorem 4.4. *Suppose that* (P) *has a strictly feasible solution (Slater condition), and that the optimal value in* (P) *is finite. Let* $\{y^k\}$ *be any maximizing sequence for* (D_r), $r > 0$, *and let* x^k *and* α_k *satisfy* (4.7), *where* $\alpha_k \to 0$. *Then* $\{x^k\}$ *is asymptotically minimizing, and there exists a sequence* $\{\bar{x}^k\}$ *of feasible solutions to* (P) *such that* $\bar{x}^k - x^k \to 0$ *and*

$$\lim_{k\to\infty} f_0(\bar{x}^k) = \lim_{k\to\infty} f_0(x^k) = \mu \;, \tag{4.14}$$

where μ *is both the optimal value and the asymptotic optimal value in* (P). *Moreover, the sequence* $\{y^k\}$ *is bounded, and all of its cluster points are Kuhn–Tucker vectors for* (P).

Proof. The condition that (P) have a strictly feasible solution is equivalent to the condition that $p_r(u) < +\infty$ for all u in some neighborhood of 0, where p_r is the (convex) perturbation function given by (3.7). The optimal value in (P) is $p_r(0)$. Thus our hypothesis implies p_r is a proper convex function which is finite in a neighborhood of 0. It then follows from the conjugacy relation (3.9) that the level sets $\{y: g_r(y) \geq \alpha\}$ are all compact, and $\max g_r = p_r(0)$ [19, Corollary 14.22]. Thus (P) is normal, $\{y^k\}$ is bounded, and every cluster point of $\{y^k\}$, being an optimal solution to (D_r), is a Kuhn–Tucker vector for (P). The boundedness of $\{y^k\}$ allows us to conclude from Theorem 4.1 that $\{x^k\}$ is asymptotically minimizing. It remains only to construct feasible solutions x^k to (P) such that the left equality of (4.14) holds. For this we employ a standard device, cf. [3, p. 107]. Let $\bar{x}$ be a strictly feasible solution and define

$$\bar{x}^k = (1-\lambda_k)\, x^k + \lambda_k \bar{x} \in X \;,$$

where

$$0 \leq \lambda_k = \max_{i=1,\ldots,m} \left\{ \frac{\theta(f_i(x^k))}{\theta(f_i(x^k)) - f_i(\bar{x})} \right\} < 1 \;.$$

Then for $i = 1, \ldots, m$ we have

$$\begin{aligned} f_i(\bar{x}^k) &\leq (1-\lambda_k)\, f_i(x^k) + \lambda_k f_i(\bar{x}) \\ &\leq (1-\lambda_k)\, \theta(f_i(x^k)) + \lambda_k f_i(\bar{x}) \\ &= \theta(f_i(x^k)) - \lambda_k\, [\theta(f_i(x^k)) - f_i(\bar{x})] \leq 0 \;, \end{aligned}$$

so that $\bar{x}^k$ is feasible. Furthermore, $\lambda_k \to 0$, so that

$$\limsup_{k\to\infty} f_0(\bar{x}^k) \leq \lim_{k\to\infty} [(1-\lambda_k) f_0(x^k) + \lambda_k f_0(\bar{x})]$$
$$= \mu \leq \liminf_{k\to\infty} f_0(x^k) .$$

This completes the proof.

5. Second derivatives and rates of convergence

In this section it is supposed that $X \subset \mathbf{R}^n$, and that the functions f_i are continuously twice differentiable in a neighborhood of a certain point $\bar{x} \in \operatorname{int} X$ which is an optimal solution to (P). We denote by $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_m)$ a multiplier vector satisfying with $\bar{x}$ the Kuhn–Tucker conditions. We assume that $\bar{x}$ and y also satisfy the extra conditions which are well known to be sufficient, if we were dealing with the non-convex case, for an isolated local constrained minimum at x [3, p. 30]:

(i) $\bar{y}_i \neq 0$ for $i \in I$, where I is the set of indices i such that $f_i(\bar{x}) = 0$.

(ii) The vectors $\nabla f_i(\bar{x})$, $i \in I$, are linearly independent.

(iii) $z \cdot H(\bar{x}, \bar{y}) z > 0$ for every nonzero $z \in \mathbf{R}^n$ satisfying $z \cdot \nabla f_i(\bar{x}) = 0$ for all $i \in I$, where

$$H(x, y) = \nabla^2 f_0(x) + \sum_{i \in I} y_i \nabla^2 f_i(x) . \tag{5.1}$$

(The inequality $z \cdot H(x, y) z \geq 0$ holds at all events for all x, y and z, due to the convexity of the functions f_i.)

These assumptions imply, of course, by convexity that $\bar{x}$ is the *unique* optimal solution to (P), and that $\bar{y}$ (being the only Kuhn–Tucker vector, in view of (ii)) is the *unique* optimal solution to the dual problems (D_r). Our aim is to use the second derivative information to analyze the speed of convergence to $\bar{x}$ and $\bar{y}$ inherent in the scheme of Theorem 4.1.

Theorem 5.1. *Under the assumptions above, there exist for arbitrary $r > 0$ and $\beta > 0$ numbers $\epsilon > 0$ and $\alpha > 0$ such that for all y and x satisfying*

$$\sup g_r - g_r(y) \leq \epsilon , \tag{5.2}$$

$$L_r(x, y) - \inf L_r(\cdot, y) \leq \alpha , \tag{5.3}$$

the following properties hold:

(i) $|y - \bar{y}| \leqslant \beta$, $|x - \bar{x}| \leqslant \beta$ *and* $x \in \operatorname{int} X$;

(ii) $L_r(x, y)$ *is continuously twice differentiable with respect to* x *and* y, *in fact*

$$L_r(x,y) = f_0(x) + \sum_{i \in I} [y_i f_i(x) + r f_i(x)^2] - (1/4r) \sum_{i \notin I} y_i^2 \,, \tag{5.4}$$

where I *is the index set corresponding to the active constraints at* $\bar{x}$ *as above;*

(iii) *the Hessian matrix* $\nabla_x^2 L_r(x,y)$ *is positive definite;*

(iv) *the convex function* $L_r(\cdot, y)$ *attains its minimum over* X *at a unique point* $\xi(y)$ *whose coordinates are continuously differentiable functions of* y *satisfying*

$$\partial \xi(y)/\partial y_i = \begin{cases} -\nabla_x^2 L_r(\xi(y), y)^{-1} \nabla f_i(\xi(y)) & \text{if } i \in I\,, \\ 0 & \text{if } i \notin I; \end{cases} \tag{5.5}$$

(v) $g_r(y)$ *is twice continuously differentiable, and the Hessian matrix* $\nabla^2 g_r(y)$ *is negative definite and satisfies*

$$w \cdot \nabla^2 g_r(y) w = -[A(y)w] \cdot \nabla_x^2 L_r(\xi(y), y)^{-1} [A(y)w] - (1/2r) \sum_{i \notin I} w_i^2 \,, \tag{5.6}$$

where the matrix $A(y)$ *is defined by*

$$A(y)w = \sum_{i \in I} w_i \nabla f_i(\xi(y)) \,. \tag{5.7}$$

Proof. We observe first that (5.4) does hold in some neighborhood of $(\bar{x}, \bar{y})$, because of the complementary slackness conditions. Passing to a smaller neighborhood if necessary to ensure differentiability, we thus obtain the formulas

$$\nabla_x L_r(x, y) = \nabla f_0(x) + \sum_{i \in I} (y_i + 2r f_i(x)) \nabla f_i(x) \,, \tag{5.8}$$

$$z \cdot \nabla_x^2 L_r(x, y) z = z \cdot H(x, y) z + 2r \sum_{i \in I} [z \cdot \nabla f_i(x)]^2 \,, \tag{5.9}$$

with $H(x, y)$ as in (5.1). It is clear from (5.9) and our assumption (iii) at the beginning of this section that $\nabla_x^2 L_r(x, y)$ is positive definite when $(x, y) = (\bar{x}, \bar{y})$, hence by continuity it is also positive definite for all (x, y) in some neighborhood of $(\bar{x}, \bar{y})$ lying within the neighborhood already considered. The implicit function theorem therefore allows us to solve the equation $\nabla_x L_r(x, y) = 0$ for x in terms of y, at least locally around $\bar{y}$. Denoting the solution by $\xi(y)$, we have

$$\nabla_x L_r(\xi(y), y) = 0 , \tag{5.10}$$

where (5.5) holds. In particular, since $\xi(y)$ is continuous in y and $\xi(\bar{y}) = \bar{x}$, the matrix $\nabla_x^2 L_r(\xi(y), y)$ is positive definite for y sufficiently near $\bar{y}$, and hence the convex function $L_r(\cdot, \bar{y})$ attains its minimum over X uniquely at $\xi(y)$. Summarizing then, we have shown the existence of a neighborhood of $(\bar{x}, \bar{y})$ on which properties (ii), (iii) and (iv) hold. We can choose this neighborhood to imply (i) and to be for the form $U \times V$, where U is a compact convex neighborhood of $\bar{x}$.

Now since $L_r(\cdot, \bar{y})$ is a continuous function attaining its minimum value $g_r(\bar{y})$ uniquely at $\bar{x}$, there is an $\alpha > 0$ such that

$$\min\{L_r(x, \bar{y}) : x \in \text{bdry}\, U\} > g_r(\bar{y}) + \alpha . \tag{5.11}$$

By the continuity of L_r and g_r, the same inequality must hold if $\bar{y}$ is replaced by any $y \in V_0$, where V_0 is a certain neighborhood of $\bar{y}$ in V. The convexity of $L_r(x, y)$ in x then implies that

$$\{x : L_r(x, y) \leqslant g_r(y) + \alpha\} \subset U \quad \text{if } y \in V_0 . \tag{5.12}$$

We now observe that, since g_r is a continuous concave function attaining its maximum uniquely at $\bar{y}$, there is an $\epsilon > 0$ such that (5.2) implies $y \in V_0$ [19, Theorem 27.2]. Thus we have found ϵ and α such that (i), (ii), (iii) and (iv) hold when x and y satisfy (5.2) and (5.3). To conclude the proof, it remains only to note that (v) follows from (iii) and (iv), using the relation

$$\nabla g_r(y) = \nabla_y L_r(\xi(y), y) , \tag{5.13}$$

which is valid by Theorem 3.2.

Corollary 5.2. *Under the assumptions at the beginning of this section, if sequences* $\{y^k\}$, $\{x^k\}$ *and* $\{\alpha_k\}$ *are generated as in Theorem* 4.1 *one has*

$$\lim_{k\to\infty} x^k = \bar{x} , \quad \lim_{k\to\infty} y^k = \bar{y} . \tag{5.14}$$

Moreover, there exist positive constants a, b_1, b_2, c_1 *and* c_2 *such that the following estimates hold for all indices* k *sufficiently large (with* $\xi(y)$ *as defined in Theorem* 5.1*):*

$$a|x^k - \xi(y^k)|^2 \leqslant \alpha_k \,, \tag{5.15}$$

$$b_1|y^k - \bar{y}|_I \leqslant |\xi(y^k) - \bar{x}| \leqslant b_2|y^k - \bar{y}|_I \,, \tag{5.16}$$

where $|y|_I^2 = \Sigma_{i \in I} y_i^2$ *, and*

$$c_1|y^k - \bar{y}|^2 \leqslant g_r(\bar{y}) - g_r(y^k) \leqslant c_2|y^k - \bar{y}|^2 \,. \tag{5.17}$$

Proof. For any $\epsilon > 0$ and $\alpha > 0$, we eventually have $g_r(y^k) \geqslant \sup g_r - \epsilon$ and $L_r(x^k, y^k) \leqslant \inf L_r(\cdot, y^k) + \alpha$. The first assertion is therefore valid, because of the arbitrariness of β in Theorem 5.1. The estimates are obvious from the differential information in the theorem.

Corollary 5.3. *Suppose in Corollary* 5.2 *that the numbers* α_k *satisfy for some* $q > 0$.

$$\alpha_k \leqslant q[\sup g_r - g_r(y^k)] \quad \textit{for all sufficiently large } k. \tag{5.18}$$

Then there is a constant $s > 0$ *such that*

$$|x^k - \bar{x}| \leqslant s|y^k - \bar{y}| \quad \textit{for all sufficiently large } k. \tag{5.19}$$

Proof. Using the estimates in Corollary 1, we have

$$\alpha|x^k - \xi(y^k)|^2 \leqslant q[g_r(\bar{y}) - g_r(y)] \leqslant c_2 q|y^k - \bar{y}|^2 \tag{5.20}$$

and consequently

$$\begin{aligned} |x^k - x| &\leqslant |x^k - \xi(y^k)| + |\xi(y^k) - \bar{x}| \\ &\leqslant [(c_2 q/a)^{1/2} + b_2]|y^k - \bar{y}| \,. \end{aligned} \tag{5.21}$$

Remarks. These results show that when the scheme in Theorem 4.1 is applied to a "twice differentiable" problem (P), one can typically expect $\{y^k\}$ to converge to a Kuhn–Tucker vector $\bar{y}$ and $\{x^k\}$ to converge to an optimal solution $\bar{x}$, moreover with the convergence of $\{x^k\}$

"at least as rapid" as the convergence of $\{y^k\}$, provided only that the numbers α_k decrease fast enough. Thus, whatever algorithm one applies to generate a maximizing sequence for the dual problem (D_r), one can hope to generate correspondingly good convergence toward a solution of (P). Naturally, this does not exclude the possibility that, by generating $\{y^k\}$ and $\{x^k\}$ by a special method, even sharper convergence properties of $\{x^k\}$ might be guaranteed than would follow just from the properties of $\{y^k\}$.

We emphasize again that these results carry over in the obvious way if (affine) equality constraints are explicitly introduced into the model.

References

[1] K.J. Arrow and R.M. Solow, "Gradient methods for constrained maxima, with weakened assumptions", in: *Studies in linear and nonlinear programming*, Eds. K. Arrow, L. Hurwicz and H. Uzawa (Stanford Univ. Press, Stanford, Calif., 1958).

[2] K.J. Arrow, F.J. Gould and S.M. Howe, "A general saddle point result for constrained optimization", Institute of Statistics Mimeo Series No. 774, Dept. of Statistics, Univ. of North Carolina, Chapel Hill, N.C., (1971).

[3] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, New York, 1968).

[4] R. Fletcher, "A class of methods for nonlinear programming with termination and convergence properties", in: *Integer and nonlinear programming*, Ed. J. Abadie (North-Holland, Amsterdam, 1970) pp. 157–175.

[5] R. Fletcher and Shirley A. Lill, "A class of methods for nonlinear programming, II: Computational experience", in: *Nonlinear programming*, Eds. J.B. Rosen, O.L. Mangasarian and K. Ritter (Academic Press, New York, 1971) pp. 67–92.

[6] E.G. Golshtein, *The Theory of Duality in Mathematical Programming and its Applications*, Nauka (1971) (in Russian).

[7] E.G. Golshtein, *Theory of Convex Programming*, AMS Translation Series (1972).

[8] P.C. Haarhoff and J.D. Buys, "A new method for the optimization of a nonlinear function subject to nonlinear constraints", *Computer Journal* 13 (1970) 178–184.

[9] M.R. Hestenes, "Multiplier and gradient methods", in: *Computing methods in optimization problems* – 2, Eds. L.A. Zadeh, L.W. Neustadt, A.V. Balakrishnan (Academic Press, New York, 1969) pp. 143–164.

[10] M.R. Hestenes, "Multiplier and gradient methods", *Journal of Optimization Theory and Applications* 4 (1969) 303–320.

[11] J.L. Joly and P.-J. Laurent, "Stability and duality in convex minimization problems", *Revue Française d'Informatique et de Recherche Opérationelle* R-2 (1971) 3–42.

[12] P.J. Laurent, *Approximation et Optimisation* (Hermann, Paris, 1972).

[13] A. Miele, E.E. Cragg, R.R. Iver and A.V. Levy, "Use of the augmented penalty function in mathematical programming problems, part I", *Journal of Optimization Theory and Applications* 8 (1971) 115–130.

[14] A. Miele, E.E. Cragg and A.V. Levy, "Use of the augmented penalty function in mathematical programming problems", part II,. *Journal of Optimization Theory and Applications* 8 (1971) 131–153.

[15] A. Miele, P.E. Moseley and E.E. Cragg, "A modification of the method of multipliers for mathematical programming problems", in: *Techniques of optimization,* Ed. A.V. Balakrishnan (Academic Press, New York, 1972) pp. 247–260.

[16] A. Miele, P.E. Moseley, A.V. Levy and G.M. Coggins, "On the method of multipliers for mathematical programming problems", *Journal of Optimization Theory and Applications* 10 (1972) 1–33.

[17] M.J.D. Powell, "A method for nonlinear constraints in minimization problems", in: *Optimization,* Ed. R. Fletcher (Academic Press, New York, 1969) pp. 283–298.

[18] R.T. Rockafellar, "Convex functions and duality in optimization problems and dynamics", in: *Mathematical systems theory and economics,* I, Eds. H.W. Kuhn and G.P. Szegö (Springer, Berlin, 1969) pp. 117–141.

[19] R.T. Rockafellar, *Convex Analysis* (Princeton Univ. Press, Princeton, N.J., 1970).

[20] R.T. Rockafellar, "Ordinary convex programs without a duality gap", *Journal of Optimization Theory and Applications* 7 (1971) 143–148.

[21] R.T. Rockafellar, "New applications of duality in convex programming", written version of talk presented at the Seventh International Symposium on Mathematical Programming, The Hague, 1970, and elsewhere; in: *Proceedings of the fourth conference on probability theory,* Brasov, Romania, 1971 (Editura Academici Republicii Socialiste Romania, Bucharest, 1973) pp. 73–81.

[22] R.T. Rockafellar, "The multiplier method of Hestenes and Powell applied to convex programming", *Journal of Optimization Theory and Applications* 12 (6) (1973).