MONOTONE OPERATORS AND AUGMENTED LAGRANGIAN METHODS IN NONLINEAR PROGRAMMING

R. T. Rockafellar[1]

ABSTRACT

The Hestenes-Powell method of multipliers in convex programming is modified to obtain a superior global convergence property under a stopping criterion that is easier to implement. The convergence results are obtained from the theory of the proximal point algorithm for solving $0 \in T(z)$ when T is a maximal monotone operator. An extension is made to an algorithm for solving variational inequalities with explicit constraint functions.

[1]Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under grant number 77-3204 at the University of Washington, Seattle.

ISBN 0-12-468660-3

1. INTRODUCTION

Let X be a nonempty, closed, convex subset of a Hilbert (or Euclidean) space H, and let $f_i : H \to R$ be a differentiable convex function for $i = 0,1,\ldots,m$. We shall be concerned with the problem

$$\text{(P)} \qquad \text{minimize } f_0(x) \text{ subject to } x \in X,\ f_i(x) \leq 0 \text{ for } i = 1,\ldots,m .$$

The ordinary dual of (P) is

$$\text{(D)} \qquad \text{maximize } \inf_x \{f_0(x) + \sum_{i=1}^m y_i f_i(x)\} \text{ subject to } 0 \leq y = (y_1,\ldots,y_m) \in R^m .$$

It will be assumed in what follows that (P) has at least one optimal solution characterized by the Kuhn-Tucker conditions. Then min (P) = max (D) of course, and the pairs (x,y) satisfying the Kuhn-Tucker conditions are precisely the ones such that x is optimal for (P) and y is optimal for (D).

In recent years there has been much interest in computational methods for (P) (and its nonconvex version) based on the augmented Lagrangian, which is the expression $L(x,y,c)$ defined for all $x \in X$, $y \in R^m$, and parameter values $c > 0$ by

$$L(x,y,c) = f_0(x) + \sum_{i=1}^m \begin{cases} y_i f_i(x) + \frac{c}{2} f_i(x)^2 & \text{if } y_i + cf_i(x) \geq 0 \\ -\frac{1}{2c} y_i^2 & \text{if } y_i + cf_i(x) \leq 0 . \end{cases}$$

This is convex in x, concave in y, and continuously differentiable in all arguments. Its saddle points (for arbitrary fixed c) are the Kuhn-Tucker pairs (x,y) for (P) and (D). If each f_i happens to be continuously twice differentiable, then so is L in all arguments, except on the hypersurfaces $y_i + cf_i(x) = 0$. Anyway, the first derivatives of L are everywhere Lipschitz continuous with one-sided directional derivatives.

For more discussion of the properties of the augmented Lagrangian, see [1], [2]. The most recent survey of the "multiplier methods" based on the augmented Lagrangian is that of Bertsekas

[3]. Many extensions and modifications of the multiplier method have been explored since it was originally suggested independently by Hestenes and Powell in 1968. In essence, all are aimed at replacing the constrained problem (P) by a sequence of unconstrained, or more simply constrained problems that can be solved efficiently by the very powerful algorithms now known for that special case. They resemble penalty methods in this respect, but they generally are better behaved than penalty methods in their rate of convergence and numerical stability (cf. [3]).

The present article, while concerned only with convex problems, will treat a new kind of modification which produces some very favorable properties and also admits a generalization from convex programming to the solution of variational inequalities with explicit constraints. The following scheme will be called the <u>proximal multiplier method</u>.

$$\text{parameters:}\quad \mu > 0,\ 0 < c_k \nearrow c_\infty \leq \infty$$
$$\text{initial guess:}\quad (x^0, y^0)$$
$$F_k(x) \triangleq L(x, y^k, c_k) + \frac{\mu^2}{2c_k}|x - x^k|^2 \text{ on } R^n$$
$$x^{k+1} \approx \arg\min_{x \in X} F_k(x)$$
$$y_i^{k+1} = \max\{0, y_i^k + c_k f_i(x^{k+1})\} \text{ for } i = 1, \ldots, m .$$

Note that the function F_k which must be minimized over X at each iteration is differentiable and convex, in fact strongly convex with modulus μ^2/c_k: for all x, x', one has

$$F_k(x') \geq F_k(x) + (x' - x) \cdot \nabla F_k(x) + \frac{\mu^2}{2c_k}|x' - x|^2 .$$

The sense in which x^{k+1} is an approximate minimizer depends on the choice of stopping rule for the minimization step.

The usual multiplier method corresponds to $\mu = 0$ (no automatic strong convexity). The modified method was introduced in [4] with $\mu = 1$ and shown to have two theoretical advantages besides the strong convexity. The sequence $\{x^k\}$ has better properties, and global convergence can be obtained under a more easily implementable stopping rule in terms of the magnitude of $\nabla F_k(x^{k+1})$. For the usual multiplier rule, the magnitude of $F_k(x^{k+1}) - \inf_X F_k$ must be monitored if global convergence (i.e. from any starting point) is to be ensured, although local convergence (i.e. from a starting point "sufficiently" close to being optimal) has been established by Polyak and Tretyakov [5] and Bertsekas [6], [7], in terms of $\nabla F_k(x^{k+1})$ in the case where $X = H = R^n$ and the strong second-order optimality conditions are satisfied.

Numerical experiments have disclosed, however, that the proximal multiplier method with $\mu = 1$ moves rather slowly in initial stages in comparison with the usual multiplier method, despite its ultimate convergence properties. The reason for this appears to be that, when c_k is too low, the quadratic term in $F_k(x)$ dominates and does not allow the Lagrangian term to have a strong enough effect in the selection of x^{k+1} . On the other hand, when c_k is too high, the "penalty" aspects of the augmented Lagrangian are too strong, and the prime advantage over penalty methods gets lost.

The introduction here of the factor μ restores flexibility in allowing the role of the quadratic term to be damped while c_k is still reasonably low. The same type of convergence results as for $\mu > 0$ will be demonstrated for arbitrarily small $\mu > 0$, although the algorithm tends to resemble the usual multiplier method more and more as $\mu \searrow 0$.

The fact that the multiplier method can be approximated in this sense by an algorithm possessing global convergence under a rule involving $\nabla F_k(x^{k+1})$ is interesting for applications to the solution of variational inequalities. As explained below in § 5, these can be handled in the same theoretical framework, essentially by replacing the gradient mapping ∇f_0 by a more general "monotone" mapping. The minimization step equivalent to finding an "approximate" solution to the equation $\nabla F_k(x) = 0$, becomes a matter of solving a more general (but "nice") equation

$A_k(x) = 0$. Since there is no longer any minimization, a stopping rule in terms of $F_k(x^{k+1}) - \inf_X F_k$ is a dead end, but one in terms of $\nabla F_k(x^{k+1})$ can be adapted by substituting $A_k(x^{k+1})$.

2. MONOTONE OPERATORS AND VARIATIONAL INEQUALITIES

Many problems can be reduced to the model: find z satisfying $0 \in T(z)$, where T is a multifunction (set-valued mapping) from a Hilbert space H into itself. Typically T is some "operator" involving subgradients, normal vectors, etc. In the infinite-dimensional case, differential operators and boundary conditions may also be involved.

The convex programming problem can be reduced to this model in three basic ways. Solutions to (P) itself are characterized by

$$0 \in T_P(x), \quad T_P = \partial f , \tag{1}$$

where ∂f is the subgradient multifunction associated with the (closed proper convex) essential objective function in (P),

$$f(x) = \begin{cases} f_0(x) & \text{if } x \text{ is feasible,} \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

Solutions to (D) are characterized by

$$0 \in T_D(y), \quad T_D = -\partial g , \tag{3}$$

where g is the (closed proper concave) essential objective function in (D),

$$g(y) = \begin{cases} \inf\limits_{x \in X} \{f_0(x) + \sum\limits_{i=1}^{m} y_i f_i(x)\} & \text{if } y \geq 0 , \\ -\infty & \text{otherwise.} \end{cases} \tag{4}$$

Finally, optimal pairs (x,y) are characterized by the Kuhn-Tucker conditions as solutions to

$$(0,0) \in T_S(x,y) , \tag{5}$$

where

$$T_S(x,y) = \{(v,u) \in H \times R^m \mid \nabla f_0(x) + \sum_{i=1}^{m} y_i \nabla f_i(x) = v , \; f_i(x) + u_i \leq 0 \text{ and } y_i[f_i(x) + u_i] = 0\} \tag{6}$$

if $x \in X$ and $y \geq 0$, but $T_S(x,y) = \phi$ otherwise.

Another class of problems is the following: given a mapping $A : H \to H$ (single valued) and a nonempty closed convex set $C \subset H$, find a point x such that

(7) $$-A(x) \in N_C(x) \text{ (normal cone to C at x) },$$

where

(8) $$N_C(x) = \begin{cases} \{-w \in H \mid (x'-x)\cdot w \geq 0 \text{ for all } x' \in C\} & \text{if } x \in C \\ \phi & \text{if } x \notin C \end{cases},$$

$$= \{-w \in H \mid \min_{x' \in C} w \cdot x' \text{ is attained at } x' = x\} .$$

This is called a variational inequality in view of its expression directly in terms of the inequalities in (8) and because it reduces when A is the gradient of a convex function f_0 to the condition for the minimum of f_0 relative to C. It reduces when C is the whole space H to the equation $A(x) = 0$.

The variational inequality (7) can be expressed in the form

(9) $$0 \in T_{VP}(x),\ T_{VP} = A + N_C .$$

However, a structured representation involving multipiers is also available when

(10) $$C = \{x \in X \mid f_i(x) \leq 0 \text{ for } i = 1,\dots,m\} .$$

Suppose the extended Slater condition is fulfilled: X is polyhedral and there is an $\tilde{x} \in C$ satisfying strictly all the inequalities for which f_i is not affine. (The polyhedral property means that X can be expressed by a finite system of linear inequalities. It could be replaced by the condition $\tilde{x} \in \text{int } X$.) In this case it is known (from the existence of Kuhn-Tucker characterizations of the minima in (8)) that

(11) $$N_C(x) = \begin{cases} \{-w \in H \mid \exists y_i \geq 0 \text{ with } y_i f_i(x) = 0 \text{ and} \\ \qquad -[w + \sum_{i=1}^{m} y_i \nabla f_i(x)] \in N_X(x)\} & \text{if } x \in C , \\ \phi \text{ if } x \notin C . \end{cases}$$

Then solving (7) is equivalent to finding a pair (x,y) such that

(12) $$(0,0) \in T_{VS}(x,y) ,$$

where

(13) $$T_{VS}(x,y) = \{(v,u) \in H\times R^m \mid v-[A(x) + \sum_{i=1}^{m} y_i \nabla f_i(x)] \in N_X(x) ,$$
$$f_i(x) + u_i \leq 0 \text{ and } y_i[f_i(x)+u_i] = 0\}$$

if $x \in X$ and $y \geq 0$, but $T_{VS}(x,y) = \phi$ otherwise.

Variational inequalities were first studied extensively in the mid 1960's by F. E. Browder, J. Lions and others, under the principal interpretation that A is some kind of integral-differential operator. This is why it is important to allow for an infinite-dimensional space H in the theory, although computation might typically proceed by a series of reductions to finite-dimensional subspaces. The use of Lagrange multipliers in such variational inequalities was first put forward by Rockafellar [8].

One of the most valuable notions that emerged from the theory of variational inequalities was that of a maximal monotone operator. A multifunction $T: H \to H$ is said to be monotone if

$$w_0 \in T(z_0),\ w_1 \in T(z_1) \Rightarrow (z_0-z_1)\cdot(w_0-w_1) \geq 0 .$$

It is maximal monotone if it is monotone and its graph $\{(z,w) | w \in T(z)\}$ is not properly included in the graph of any other monotone operator.

The subdifferentials $T = \partial h$ of the closed proper convex functions h on H are important examples of maximal monotone operators [9], [10]. Another example of special interest below is

(14) $$T_\ell(x,y) = (\partial_x \ell(x,y), -\partial_y \ell(x,y)) ,$$

where ℓ is a "closed" convex-concave function on $H\times H'$ (product of two Hilbert spaces) [11]. A single-valued monotone mapping defined on all of H is maximal monotone if and only if it is continuous from the norm topology to the weak topology (cf. [12]). (When H is finite-dimensional, the two topologies reduce to the usual one.) If $T(x) = Mx + b$ where M is a continuous linear transformation, then T is maximal monotone if and only if M is positive semidefinite (not necessarily symmetric):

$$w \cdot Mw \geq 0 \text{ for all } w \in H . \tag{15}$$

More generally, if T is a single-valued differentiable mapping, it is maximal monotone if and only if its derivative (i.e. its Jacobian) at every point is positive semidefinite in the sense of (15).

The operators T_P and T_D above are maximal monotone because they are of the form $T = \partial h$, while T_S is maximal monotone because it is of the form (14) for ℓ the ordinary Lagrangian function in the convex programming problem.

Proposition 1

Suppose $A : H \to H$ is a single-valued(everywhere defined) monotone operator which is continuous from the norm topology to the weak topology. Then the operators T_{VP} and T_{VS} are maximal monotone.

Proof

The operator $N_C : x \to N_C(x)$ is maximal monotone, because it is the subdifferential of the indicator function of C. On the other hand, A is maximal monotone (as already noted in the remarks above. To obtain the conclusion about T_{VP}, we need only apply the fact that the sum of two maximal monotone operators is maximal monotone if the effective domain of one has a point interior to the effective domain of the other [8, Theorem 1]. (The effective domain of a multifunction is the set of points where it is nonempty-valued.) The conclusion about T_{VP} follows in the same way from the representation $T_{VP} = T_0 + T_1$ where T_0 is the special case of T_S with $f_0 \equiv 0$ and $T_1 : (x,y) \to (A(x),0)$.

The extended Slater condition is therefore not needed for T_{VS} to be maximal monotone, but merely as a sufficient condition for the original variational inequality to be equivalent to finding a pair (x,y) that satisfies

$$(16)\quad \begin{array}{l} x\in X \text{ and } f_i(x) \le 0,\ y_i \ge 0,\ y_i f_i(x) = 0 \text{ for } i=1,\ldots,m\ , \\ -[A(x) + y_1 \nabla f_1(x) + \ldots + y_m \nabla f_m(x)] \in N_X(x)\ . \end{array}$$

The latter could just as well be adopted as the real problem of interest when C is given by explicit constraints.

3. PROXIMAL POINT ALGORITHM FOR MONOTONE OPERATORS

We have reduced a number of problems to the model

(17) find z satisfying $0 \in T(z)$,
where $T: H \to H$ is a given maximal monotone opeator.

A fundamental algorithm for this problem has been developed in [13] using the fact that for arbitrary $c > 0$ the operator $(I+cT)^{-1}$ is a single-valued and nonexpansive. It is called the proximal point algorithm.

parameters: $0 < c_k \nearrow c_\infty \leq \infty$, $\varepsilon_k > 0$, $\sum_{k=0}^{\infty} \varepsilon_k < \infty$
initial point: z^0
$z^{k+1} \approx (I+c_k T)^{-1}(z^k)$
stopping criterion:
$|z^{k+1} - (I+c_k T)^{-1}(z^k)| \leq \varepsilon_k \max\{1, |z^{k+1}-z^k|\}$.

It may seem that the stopping criterion requires explicit knowledge of the mapping $(I+c_k T)^{-1}$, but this is not the case in a number of applications where convenient estimates are available.

The main result about the proximal point algorithm is the following. Suppose at least one solution to (17) exists (as can be guaranteed by conditions discussed in [13], for instance). Then z^k converges (in the weak topology of H) to a particular solution z^∞ (even though there may be more than one solution!). Moreover

(18) $$\limsup_{k\to\infty} \frac{|z^{k+1}-z^\infty|}{|z^k-z^\infty|} \leq \frac{(a/c_\infty)}{[1+(a/c_\infty)^2]^{\frac{1}{2}}} \leq 1 \quad ,$$

where $a \in [0,\infty]$ is a certain number associated with the problem, namely

(19) $$\begin{aligned} a &= \text{Lipschitz constant for } T^{-1} \text{ at } 0 \\ &= \lim_{\delta \searrow 0} \sup_{\substack{z \in T^{-1}(w) \\ \bar{z} \in T^{-1}(0) \\ |w| \leq \delta}} \frac{|z-\bar{z}|}{|w|} \leq \infty . \end{aligned}$$

(In (18), the convention is adopted that the right hand side is 1 if $a = \infty$, but 0 if $a < \infty$ and $c_\infty = \infty$. The ratio on the left side is ∞ if just the denominator vanishes, but 0 if both the numerator and denominator vanish. The latter rule is involved also in (19).)

Observe in particular that if $a < \infty$ the convergence is linear with a modulus that can be forced as close to zero as desired by choosing c_∞ high enough; for $c_\infty = \infty$, the convergence is superlinear. It is of interest therefore to have some feeling for whether the constant a can be expected to be finite. This can be provided in some applications in terms of the kind of optimality conditions satisfied by the solution, but the following general result may be cited.

Proposition 2

If H is finite-dimensional, then the set of points belonging to

(20) $$\text{range } T = \{w \mid T^{-1}(w) \neq \phi\}$$

and at which the Lipschitz constant for T^{-1} is infinite is negligible (i.e. of Lebesgue measure zero).

Proof

Mignot [14] has shown that in the finite-dimensional case a maximal monotone operator is actually single-valued and differentiable at almost every interior point of its effective domain. At such points it has in particular a finite Lipschitz constant. Applying this fact to T^{-1}, whose maximal monotonicity follows trivially from that of T, we see that T^{-1} has a finite Lipschitz constant at almost every interior point of the set (20). But this set is also known to be virtually convex, which in the finite dimensional case means that it differs from a closed convex set only in the possible omission of certain relative boundary points [16, 17]. The noninterior points therefore form a set of measure 0, and the result follows.

To interpret Proposition 2, think of the basic problem in parametric form: find z such that $\bar{w} \in T(z)$, where $\bar{w}$ is a given vector. This amounts to considering simultaneously all the maximal monotone operators of the form $T - \bar{w}$. The choices of $\bar{w}$ for which the problem has a solution are those in (20). Thus the problems for which a solution exists, but the corresponding constant a is not finite, form a negligible set.

4. APPLICATION TO CONVEX PROGRAMMING

When the proximal point algorithm is applied to the maximal monotone operator T_P one obtains a method of the form

$$x^{k+1} \approx \arg\min_{x\in X} \{f_0(x) + \frac{1}{2c_k} |x-x^k|^2 \big| f_i(x) \le 0,\ i=1,\dots,m\}$$

This has its interesting aspects, discussed in [4], but more important are the cases of T_D and T_S, which reduce respectively to the usual multiplier method and the proximal multiplier method with $\mu = 1$.

The main results obtained in this way for the usual multiplier method (the algorithm in §1 for $\mu = 0$) are the following (see [4, Theorems 4 and 5]). Suppose the stopping rule is

$$F_k(x^{k+1}) - \inf_{x\in X} F_k(x) \le \frac{\delta_k^2}{2c_k} \max\{1, |y^{k+1}-y^k|^2\}, \tag{21}$$

where y^{k+1} is the multiplier vector defined in the algorithm as a function of y^k and x^{k+1}, and $\delta_k > 0$, $\sum_{k=0}^{\infty} \delta_k < \infty$. Then (under the assumptions about the problem stated in §1) the sequence $\{x^k\}$ is asymptotically minimizing for (P) in the sense that

$$x^k \in X,\ f_0(x^k) \to \min(P),\ \max\{0, f_i(x^k)\} \to 0 \text{ for } i=1,\dots,m,$$

while $\{y^k\}$ converges to some optimal solution y^∞ to (D). Furthermore

$$\limsup_{k\to\infty} \frac{|y^{k+1}-y^\infty|}{|y^k-y^\infty|} \le \frac{(a_0/c_\infty)}{[1+(a_0/c_\infty)^2]^{\frac{1}{2}}}, \tag{22}$$

where a_0 is the Lipschitz constant for T_D^{-1} at 0. (This expression uses the conventions explained after (18), (19).)

A stricter stopping rule is needed to obtain better convergence properties of $\{x^k\}$. Let

(23) $\quad \operatorname{proj} \nabla F_k(x)$ = projection of $\nabla F_k(x)$ on the closed tangent cone to X at x.

(If X is the whole space H, this is just $\nabla F_k(x)$ itself.)

If one invokes (21) simultaneously with

(24) $$|\text{proj } \nabla F_k(x^{k+1})| \le \frac{\varepsilon_k}{c_k} \max \{1, |y^{k+1}-y^k|\} ,$$

where $0 < \varepsilon_k \to 0$, and if the Lipschitz constant a_1 associated with T_S^{-1} at $(0,0)$ happens to be finite, then $\{x^k\}$ converges (in the norm topology of H) to a solution x^∞ to (P) and

(25) $$\limsup_{k\to\infty} \frac{|x^{k+1}-x^\infty|}{|y^{k+1}-y^k|} \le a_1/c_\infty .$$

(In (25), the conventions explained for (18) are likewise in effect.)

The corresponding results for the proximal multiplier method will depend on the operator $T_\mu : H\times R^m \to H\times R^m$ defined by

(26) $$T_\mu(\xi,y) = \{(\eta,u) \mid (\mu\eta,u) \in T_S(\mu^{-1}\xi,y)\}$$

for $\mu > 0$, as well as on the associated value

(27) $$a_\mu = \text{Lipschitz constant for } T_\mu^{-1} \text{ at } (0,0) .$$

Note that for $\mu = 1$ these reduce to T_S and a_1 (as already defined above).

Proposition 3

For every $\mu > 0$, T_μ is a maximal monotone operator and $a_\mu \ge a_0$ (the Lipschitz constant for T_D^{-1} at 0). Moreover a_μ is nondecreasing as a function of μ , and if it is finite for one $\mu > 0$ it is finite for all $\mu > 0$. If the space H is finite-dimensional, then $a_\mu \searrow a_0$ as $\mu \searrow 0$.

Proof

If $(\eta,u) \in T_\mu(\xi,y)$ and $(\bar\eta,\bar u) \in T_\mu(\bar\xi,\bar y)$, then $(\mu\eta,u) \in T_S(\mu^{-1}\xi,y)$ and $(\mu\bar\eta,\bar u) \in T_S(\mu^{-1}\bar\xi,\bar y)$ by definition, and the monotonicity of T_S implies

(28) $$\begin{aligned} 0 &\le [(\mu\bar\eta,\bar u)-(\mu\eta,u)]\cdot[(\mu^{-1}\bar\xi,\bar y)-(\mu^{-1}\xi,y)] \\ &= \mu\mu^{-1}(\bar\eta-\eta)(\bar\xi-\xi) + (\bar u-u)\,(\bar y-y) \\ &= [(\bar\eta,\bar u)-(\eta,u)]\cdot[(\bar\xi,\bar y)-(\xi,y)] . \end{aligned}$$

Thus T_μ is monotone. Now suppose $(\bar{\eta},\bar{u}) \in T(\bar{\xi},\bar{y})$, where T is a monotone operator whose graph includes the graph of T_μ. Then in particular (28) holds whenever $(\eta,u) \in T_\mu(\xi,\eta)$; hence

$$0 \leq [(\mu\bar{\eta},\bar{u})-(v,u)]\cdot[(\mu^{-1}\bar{\xi},\bar{y})-(x,y)]$$

whenever $(v,u) \in T_S(x,y)$. In other words, one could add $(\mu^{-1}\bar{\xi},\bar{y},\mu\bar{\eta},\bar{u})$ to the graph of T_S without destroying its monotonicity. But T_S is maximal monotone, so this implies $(\mu\bar{\eta},\bar{u}) \in T_S(\mu^{-1}\bar{\xi},\bar{y})$. Thus $(\bar{\mu},\bar{u}) \in T_\mu(\bar{\xi},\bar{y})$. This shows that T_μ is maximal too. By definition

$$a = \lim_{\delta \searrow 0} \sup_{\substack{(\xi,y) \in T_\mu^{-1}(\eta,u) \\ (\bar{\xi},\bar{y}) \in T_\mu^{-1}(0,0) \\ |(\eta,u)| \leq \delta}} \frac{|(\xi,y)-(\bar{\xi},\bar{y})|}{|(\eta,u)|},$$

$$= \lim_{\delta \searrow 0} \sup_{\substack{(x,y) \in T_S^{-1}(v,u) \\ (\bar{x},\bar{y}) \in T_S^{-1}(0,0) \\ |(\mu^{-1}v,u)| \leq \delta}} \frac{|(\mu x,y)-(\mu\bar{x},\bar{y})|}{|(\mu^{-1}v,u)|}.$$

The ratio can also be expressed as

$$\frac{[\mu^2|x-\bar{x}|^2+|y-\bar{y}|^2]^{\frac{1}{2}}}{[\mu^{-2}|v|^2 + |u|^2]^{\frac{1}{2}}},$$

which clearly is nondecreasing as a function of μ. Hence a_μ is nondecreasing in μ. A change in μ merely amounts to a change to equivalent norms and thus cannot affect the finiteness of a_μ.

In the case where a_μ is finite for $\mu > 0$, T_S^{-1} has a finite Lipschitz constant at $(0,0)$. Then $T_S^{-1}(0,0)$ consists of a single element $(\bar{x},\bar{y})$. Furthermore, $(0,0)$ is an interior point of the rnage of T_S [12, Theorem 1] and

(29) $$T_D^{-1}(u) = \{y \mid \exists x \text{ with } (x,y) \in T_S^{-1}(0,y)\}$$

[4, Proposition 1]. Thus

$$a_0 = \lim_{\delta \searrow 0} \sup_{\substack{(x,y) \in T_S^{-1}(0,u) \\ (\bar{x},\bar{y}) \in T_S^{-1}(0,0) \\ |(0,v)| \le \delta}} \frac{|(0,y)-(0,\bar{y})|}{|(0,u)|}$$

and in particular $a_0 \le a_\mu$ for all $\mu > 0$.

On the other hand, suppose a is a number such that $a < a_\mu < \infty$ for all $\mu > 0$. Assuming finite-dimensionality, we shall show that

$$a_0 = \inf_{\mu>0} a_\mu = \lim_{\mu \searrow 0} a_\mu .$$

Fix any $\delta > 0$ small enough that the set

(30) $\{(x,y) \mid \exists (v,u) \text{ with } |(v,u)| \le \delta \text{ and } (x,y) \in T_S^{-1}(v,u)\}$

is bounded (as is possible because T_S^{-1} has a finite Lipschitz constant a_1 at $(0,0)$ by assumption). For $\mu = 1/j$, $j=1,2,\dots$, there are elements $(x_j,y_j) \in T_S^{-1}(v_j,u_j)$ such that $|(jv_j,u_j)| \le \delta$ and

(31) $$|(j^{-1}x_j,y_j)-(j^{-1}\bar{x},\bar{y})| \ge a|(jv_j,u_j)| .$$

Then in particular $|(v_j,u_j)| \le \delta$, so the sequence $\{(x_j,y_j)\}$ belongs to the bounded set (30). Extracting subsequences if necessary, we can suppose that $(v_j,u_j) \to (\tilde{v},\tilde{u})$ and $(x_j,y_j) \to (\tilde{x},\tilde{y})$. Actually $\tilde{v} = 0$, since $j|v_j| \le \delta$ for all j. Then $(\tilde{x},\tilde{y}) \in T_S^{-1}(0,\tilde{u})$ because the graph of T_S (and hence of T_S^{-1}) is closed by maximality. Taking the limit in (31), we see that

(32) $$|(0,\tilde{y})-(0,\bar{y})| \ge a|(0,\tilde{u})| .$$

Thus for each $\delta > 0$ sufficiently small there exist $\tilde{y}$ and $\tilde{u}$ such that $(\tilde{x},\tilde{y}) \in T_S^{-1}(0,\tilde{u})$ for some $\tilde{x}$, and (32) holds. This shows that $a \ge a_0$ and completes the proof.

Theorem 1

Suppose the proximal multiplier method for arbitrary $\mu > 0$ is executed with the stopping rule

(33) $$|\mathrm{proj}\ \nabla F_k(x^{k+1})| \le \frac{\varepsilon_k}{c_k} \max\ \{1, |(x^{k+1}-x^k, y^{k+1}-y^k)|_\mu\} ,$$

where $\varepsilon_k > 0$, $\sum_{k=0}^{\infty} \varepsilon_k < \infty$, and

(34) $$|(x,y)|_\mu = [\mu^2|x|^2 + |y|^2]^{1/2} .$$

Then x^k converges (in the weak topology of H) to a solution x^∞ to (P), y^k converges to a solution y^∞ to (D), and

(35) $$\limsup_{k\to\infty} \frac{|(x^{k+1},y^{k+1})-(x^\infty,y^\infty)|_\mu}{|(x^k,y^k)-(x^\infty,y^\infty)|_\mu} \le \frac{(a_\mu/c_\infty)}{[1+(a_\mu/c_\infty)^2]^{1/2}}$$

Proof

For $\mu = 1$, this was established in [4, §5] by applying the theory of the proximal point algorithm to T_S . The general case is obtained in the same way from T_μ . As a matter of fact, this simply amounts to a change of variables $\xi = \mu x$ in the convex programming problem: at each step one minimizes the function $\Phi_k(\xi) = F_k(\mu^{-1}\xi)$, whose gradient is $\nabla\Phi_k(\xi) = \mu^{-1}\nabla F_k(\mu^{-1}\xi)$, on the set $\Xi = \mu X$. The stopping rule in terms of ξ^{k+1} is

$$|\mathrm{proj}\ \nabla\Phi_k(\xi^{k+1})| \le \frac{\varepsilon'_k}{c_k} \max\ \{1, |(\xi^{k+1},y^{k+1})-(\xi^k,y^k)|\} ,$$

for values $\varepsilon'_k > 0$ with $\sum_{k=0}^{\infty} \varepsilon'_k < \infty$. In terms of $x^{k+1} = \mu^{-1}\xi^{k+1}$, this translates into

$$\mu^{-1}\ |\ \mathrm{proj}\ \nabla F_k(x^{k+1})| \le \frac{\varepsilon'_k}{c_k} \max\ \{1, |(\mu x^{k+1},y^{k+1})-(\mu x^k,y^k)|\}$$

and becomes the rule (32) upon setting $\varepsilon_k = \mu\varepsilon'_k$.

Remark 1

As $\mu \searrow 0$, this stopping rule is transformed into the rule (24) used (in part) in the usual multiplier method, while in the finite-dimensional case at least, the result (35) is transformed into (22) (since $a_\mu \searrow a_0$ by Proposition 3).

Remark 2

Proposition 2 shows that a_μ can be expected to be finite. Actually a_1 , and hence a_μ for all $\mu > 0$, is finite in particular if the strong second-order optimality conditions hold for the problem [4, Proposition 2], the solution $\bar{x}$ being interior to X. Spingarn, in his thesis [15], has extended the statement of the second-order optimality conditions to allow "active" sets X of a certain class and has proved in a well developed sense that these conditions are satisfied by almost all problems with sufficiently differentiable functions f_i. Thus the modulus on the right side of (35) is less than 1 except for a "negligible set of problems".

5. APPLICATION TO VARIATIONAL INEQUALITIES

The proximal multiplier method may now be extended easily to the problem of determining a pair (x,y) which satisfies a variational inequality expressed in the form (16), where A *is a single-valued monotone operator which is continuous from the norm topology to the weak topology of* H. The idea is clear from the fact that the algorithm can be described entirely in terms of the gradient mapping ∇f_0 rather than the values of the objective function f_0.

Minimizing F_k over X "approximately" is the same as finding an approximate solution to the variational inequality

$$-\nabla F_k(x) \in N_X(x) .$$

It is easy to see that actually

$$|\mathrm{proj}\ \nabla F_k(x)| = \mathrm{dist}(0, \nabla F_k(x) + N_X(x)).$$

But

$$\nabla F_k(x) = \nabla f_0(x) + \sum_{i=1}^{m} \max\,\{0, y_i^k + c_k f_i(x)\}\ \nabla f_i(x) + \frac{\mu^2}{c_k}(x - x^k) .$$

The procedure makes sense, therefore, if one simply replaces $\nabla f_0(x)$ by the more general expression $A(x)$. This yields the *proximal multiplier method for variational inequalities*:

parameters: $\mu > 0,\ 0 < c_k \quad c_\infty \le \infty$,

$\varepsilon_k > 0,\ \sum_{k=0}^{\infty} \varepsilon_k < \infty$.

initial guess: (x^0, y^0) .

$$A_k(x) \triangleq A(x) + \sum_{i=1}^{m} \max\,\{0, y_i^k + c_k f_i(x)\}\ \nabla f_i(x) + \frac{\mu^2}{c_k}(x - x^k) ,$$

$x^{k+1} \approx$ solution to $-A_k(x) \in N_X(x)$

$y_i^{k+1} = \max\,\{0, y_i^k + c_k f_i(x^{k+1})\}$ for $i=1,\ldots,m$

stopping rule:

$$|\mathrm{proj}\ A_k(x^{k+1})| \le \frac{\varepsilon_k}{c_k} \max\,\{1, |(x^{k+1}, y^{k+1}) - (x^k, y^k)|_\mu\} ,$$

where $|\cdot|_\mu$ is the norm defined in (34). The solution of the simpler variational inequality at each iteration is assisted by the following fact.

Proposition 4

The mapping A_k is maximal monotone (single-valued), in fact strongly monotone with modulus μ^2/c_k, in the sense that for all x, x',

$$[x'-x]\cdot[A_k(x')-A_k(x)] \geq \frac{\mu^2}{c_k} |x'-x|^2 .$$

Proof

It has been observed that the function F_k is strongly convex with modulus μ^2/c_k (and hence ∇F_k is strongly monotone with the same modulus), even if f_0 is the constant 0. Taking $f_0 \equiv 0$, we can write $A_k = A + \nabla F_k$. Thus A_k is the sum of two maximal monotone operators and therefore is maximal monotone by [8, Theorem 1]. The strong monotonicity of ∇F_k is inherited by A_k.

The definitions of T_μ and a_μ for $\mu > 0$ are extended to the present case simply by substitution T_{VS} for T_S in (26) and (27).

Proposition 5

Under these extended definitions, it remains true that T_μ is a maximal monotone operator, and a_μ is a nondecreasing function of $\mu > 0$ which is finite for all μ if finite for one μ.

Proof

The corresponding arguments for Proposition 3 depended only on the maximal monotonicity of T_S and hence carry over to T_{VS} by virtue of Proposition 2.

Theorem 2

In the proximal multiplier method for variational inequalities, under the assumption that (16) has at least one solution, one always has $x^k \to x^\infty$ (weak topology of H) and $y^k \to y^\infty$, where (x^∞, y^∞) is a particular solution to (16). Moreover

$$\limsup_{k\to\infty} \frac{|(x^{k+1},y^{k+1})-(x^\infty,y^\infty)|_\mu}{|(x^k,y^k)-(x^\infty,y^\infty)|_\mu} \le \frac{(a_\mu/c_\infty)}{[1+(a_\mu/c_\infty)^2]} .$$

Proof

This is obtained by applying the proximal point algorithm to the maximal monotone operator T_μ and making the change of variables $\xi = \mu x$. It suffices to demonstrate that for $\xi^{k+1} = \mu x^{k+1}$ and $\varepsilon_k' = \mu^{-1}\varepsilon_k$ the present stopping condition implies

$$(36)\qquad |(\xi^{k+1},y^{k+1})-(I+c_kT_\mu)^{-1}(\xi^k,y^k)| \le \varepsilon_k' \max\{1,|(\xi^{k+1},y^{k+1})-(\xi^k,y^k)|\} .$$

We have by definition

$$(37)\quad T_\mu(\xi,y) = \{(\eta,u)\,|\,\mu\eta-[A(\mu^{-1}\xi) + \sum_{i=1}^m y_i\nabla f_i(\mu^{-1}\xi)] \in N_X(\mu^{-1}\xi) ,\ f_i(\mu^{-1}\xi)+u_i \le 0,\ y_i[f_i(\mu^{-1}\xi)+u_i] = 0\}$$

if $\mu^{-1}\xi \in X$, $y\geq 0$, and $T_\mu(\xi,y) = 0$ otherwise. On the other hand, we have by definition $y^{k+1} \ge 0$ and

$$A_k(x^{k+1}) - \frac{\mu^2}{c_k}(x^{k+1}-x^k) = A(x^{k+1}) + \sum_{i=1}^m y_i^{k+1}\nabla f_i(x^{k+1}) ,$$

or equivalently

$$(38)\qquad A_k(\mu^{-1}\xi^{k+1}) - \frac{\mu}{c_k}(\xi^{k+1}-\xi^k) = A(\mu^{-1}\xi^{k+1}) + \sum_{i=1}^m y_i^{k+1}\nabla f_i(\mu^{-1}\xi^{k+1}) .$$

Let $u = (y^k-y^{k+1})/c_k$, so that

$$f_i(\mu^{-1}\xi^{k+1}) + u_i = \frac{1}{c_k}[y_i^k + c_kf_i(x^{k+1}) - y^{k+1}] .$$

Then

$$(39)\qquad f_i(\mu^{-1}\xi^{k+1}) + u_i \le 0,\ y_i^{k+1}[f_i(\mu^{-1}\xi^{k+1}) + u_i] = 0 .$$

Referring from (38) and (39) back to (37), we see that for arbitrary

$$(40)\qquad w \in [A_k(x^{k+1}) + N_X(x^{k+1})]$$

it is true that

$$(\mu^{-1}w+c_k^{-1}(\xi^k-\xi^{k+1}),\ c_k^{-1}(y^k-y^{k+1})) \in T_\mu(\xi^{k+1},y^{k+1}) \ .$$

The latter is the same as

$$(\mu^{-1}c_k w+\xi^k,y^k) \in (I+c_k T_\mu)(\xi^{k+1},y^{k+1}) \ ,$$

so that

$$(\xi^{k+1},y^{k+1}) = (I+c_k T_\mu)^{-1}(\mu^{-1}c_k w+\xi^k,y^k) \ . \tag{41}$$

But the operator $(I+c_k T_\mu)^{-1}$ is nonexpansive since T_μ is monotone (cf. [13, Prop. 1]). Therefore (41) implies

$$|(\xi^{k+1},y^{k+1})-(I+c_k T_\mu)^{-1}(\xi^k,y^k)$$
$$\leq |(\mu^{-1}c_k w+\xi^k,y^k)-(\xi^k,y^k)| = \mu^{-1}c_k|w| \ .$$

This holds for all w satisfying (40), so

$$|(\xi^{k+1},y^{k+1})-(I+c_k T_\mu)^{-1}(\xi^k,y^k)| \leq \mu^{-1}c_k \operatorname{dist}\{0,A_k(x^{k+1})+N_X(x^{k+1})\}$$
$$= \mu^{-1}c_k|\operatorname{proj} A_k(x^{k+1})| \ .$$

It follows that the stopping criterion in the present algorithm, namely

$$|\operatorname{proj} A_k(x^{k+1})| \leq \frac{\varepsilon_k}{c_k} \max\{1,|(x^{k+1},y^{k+1})-(x^k,y^k)|_\mu\}$$
$$= \frac{\varepsilon_k'}{c_k} \max\{1,|(\xi^{k+1},y^{k+1})-(\xi^k,y^k)|\} \ ,$$

does imply (36) as claimed.

Remark

Proposition 2 has the interpretation in this context that for "almost all" variational inequality problems the constants a_μ will be finite.

REFERENCES

[1] R. T. Rockafellar, "Augmented Lagrange multiplier functions and duality in nonlinear programming", SIAM J. Control 12 (1974), 268-285.

[2] R. T. Rockafellar, "Solving a nonlinear programming problem by way of a dual problem", Symposia Mathematica XIX (1976), 135-160.

[3] D. P. Bertsekas, "Multiplier methods: a survey", Automatica 12 (1976), 133-145.

[4] R. T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming", Math. of Operations Research 1 (1976), 97-116.

[5] B. T. Polyak and N. V. Tretyakov, "The method of penalty bounds for constrained extremum problems", Zh. Vych. Mat. i Mat. Fiz. 13 (1973), 34-46 = U.S.S.R. Computational Math. and Math. Phys. 13 (1973), 42-58.

[6] D. P. Bertsekas, "On penalty and multiplier methods for constrained minimization", SIAM J. Control 14 (1976), 216-235.

[7] D. P. Bertsekas, "Convergence rate of penalty and multiplier methods", Proc. IEEE Conf. on Decision and Control, San Diego, Cal. (1973), 260-264.

[8] R. T. Rockafellar, "On the maximality of sums of nonlinear monotone operators", Trans. Amer. Math. Soc. 149 (1970), 75-88.

[9] G. J. Minty, "On the monotonicity of the gradient of a convex function", Pacific J. Math. 14 (1964), 243-247.

[10] R. T. Rockafellar, "On the maximal monotonicity of subdifferential mappings", Pacific J. Math. 33 (1970), 209-216.

[11] R. T. Rockafellar, "Monotone operators associated with saddle functions and minimax problems", in Nonlinear Functional Analysis, Part 1 (F. E. Browder, ed.), Symposia in Pure Math. 18, Amer. Math. Soc., 1970, 397-407.

[12] R. T. Rockafellar, "Local boundedness of nonlinear monotone operators", Michigan Math. J. 16 (1969), 397-407.

[13] R. T. Rockafellar, "Monotone operators and the proximal point algorithm", SIAM J. Control Opt. 14 (1976), 877-898.

[14] F. Mignot, "Un théorème de differentiabilité; application aux projections et au contrôle dans les inéquations", Seminaires sur les Equations aux Derivées Partielles, 1973-1974, Université Paris VI; see J. Math. Analysis Appl. 1977.

[15] J. Spingarn, Generic Conditions for Optimality in Constrained Minimization Problems, Thesis, Univ. of Washington, 1977.

[16] G. J. Minty, "On the maximal domain of a 'monotone' function", Michigan Math. J. 8 (1961), 135-137.

[17] R. T. Rockafellar, "On the virtual convexity of the domain and range of a nonlinear maximal monotone operator", Math. Ann. 185 (1970), 81-90.