July 1981 LIDS-P-1113

AUGMENTED LAGRANGIAN AND DIFFERENTIABLE EXACT PENALTY METHODS*

by

Dimitri P. Bertsekas**

*This research was condicted in the M.I.T. Laboratory for Information and Decision Systems with partial support provided by National Science Foundation Grant No. NSF/ECS 79-20834.

**Room No. 35-210, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

1. Introduction

The original proposal of an Augmented Lagrangian method by Hestenes (1969) and Powell (1969) may be viewed as a significant milestone in the recent history of the constrained optimization area. Augmented Lagrangian methods are not only practically important in their own right, but have also served as the starting point for a chain of research developments centering around the use of penalty functions, Lagrange multiplier iterations, and Newton's method for solving the system of necessary optimality conditions.

Augmented Lagrangian methods became quite popular in the early seventies but then yielded ground to algorithms based on Newton's method for solving the system of necessary optimality conditions usually referred to as recursive quadratic programming (RQP) techniques. The author believes however that Augmented Lagrangian methods will very likely maintain for a long time a significant position within the arsenal of computational methodology for constrained optimization. In fact their value may be appreciated further as interest shifts more towards large problems. I will try to outline some of the reasons for this assessment and briefly survey the state of the art of Augmented Lagrangian methods in the next section.

On the other hand there is extensive evidence that for many problems, particularly those of relatively small dimension, RQP techniques are considerable more efficient than Augmented Lagrangian methods. Locally convergent variants of these methods have been known for many years and have seen considerable use in control theory and economics. Their broad acceptance in mathematical programming practice became possible, however, only after a methodology was developed that allowed global convergence based on descent using exact penalty functions. The use of a nondifferentiable exact penalty function for this purpose was originally proposed by Pschenichny (1970), (1975). His work became widely known in the Soviet Union

but went largely unnoticed in the West where nondifferentiable exact penalty functions were independently introduced in connection with iterations based on RQP by Han (1977). The work of Powell (1978) showed how to use effectively Quasi-Newton approximations within the nondifferentiable exact penalty-RQP framework and contributed significantly to the popularization of the overall approach. There are many significant contributions in this area and they will be covered extensively in other papers in this volume. It is interesting to note that the RQP direction together with a unity stepsize does not necessarily lead to a decrease of the value of the nondifferentiable exact penalty function even arbitrarily close to a solution as noted by Maratos (1978). This is a potentially serious shortcoming since it may prevent superlinear convergence in situations where it otherwise might be expected. To bypass this difficulty it is necessary to introduce modifications in the algorithm such as those suggested by Mayne and Polak (1978) and Chamberlain et al (1979).

Recently there has been some interest in the use of differentiable exact penalty functions in connection with RQP. A class of such functions has been proposed by DiPillo and Grippo (1979). There is an interesting connection between the Newton direction for minimizing any function in the DiPillo-Grippo class and the Newton direction for solving the system of necessary optimality conditions which has been noted independently in connection with second derivative algorithms in Bertsekas (1980a) and in connection with Quasi-Newton methods in Dixon (1980). It is also interesting that a class of exact penalty functions proposed by Fletcher (1970) can be derived (and indeed expanded) via the DiPillo-Grippo class [Bertsekas (1980a)]. A further link in the chain of these developments was established in Bertsekas (1980b) where it was shown that the RQP direction based on positive definite approximations to the Hessian of the Lagrangian [in-

cluding those obtained by the formula of Powell (1978)] is a descent direction for any function in Fletcher's class arbitrarily far from a solution as long as the penalty parameter is sufficiently large. Furthermore a unity stepsize near the solution decreases the value of the penalty function, so the difficulty noted by Maratos (1978) in connection with nondifferentiable exact penalty functions does not arise. These results which will be described in Section 3, have placed differentiable exact penalty functions on an equal footing with non-differentiable ones in terms of desirable descent properties. More research should be expected in this area as evidenced by recent work by Boggs and Tolle (1981) and Han and Mangasarian (1981) reported during the meeting. We mention also a two-parameter differentiable exact penalty proposed independently by Boggs and Tolle (1980) which has also been related to Fletcher's class of penalty functions and to Newton's method for solving the system of necessary optimality conditions.

In what follows we will restrict ourselves exclusively to the equality constrained problem

$$\text{minimize } f(x) \qquad \text{(ECP)}$$
$$\text{subject to } h(x) = 0$$

where $f: R^n \to R$ and $h: R^n \to R^m$ are assumed to be three times continuously differentiable.

Primary attention will be focused at local minima-Lagrange multiplier pairs (x^*,λ^*) satisfying the following second order sufficiency assumptions for optimality

$$\nabla_x L_o(x^*,\lambda^*) = 0, \quad h(x^*) = 0 \quad , \; x^* \in X^* \qquad \text{(S)}$$
$$z' \nabla^2_{xx} L_o(x^*,\lambda^*) z > 0, \quad \forall\, z \neq 0, \; \nabla h(x^*)' z = 0$$

where $L_o: R^{n+m} \to R$ is the (ordinary) Lagrangian function

$$L_o(x,\lambda) = f(x) + \lambda' h(x)$$

and X^* is the set given by

$$X^* = \{x \mid \nabla h(x) \text{ has rank } m\}.$$

In our notation all vectors are considered to be column vectors. A prime denotes transposition. The usual norm on the Euclidean space R^n is denoted by $|\cdot|$ [i.e., $|x| = (x'x)^{1/2}$ for all $x \in R^n$]. For a mapping $h: R^n \to R^m$, $h = (h_1,\ldots,h_m)'$ we denote by $\nabla h(x)$ the nxm matrix with columns the gradients $\nabla h_1(x),\ldots,\nabla h_m(x)$. Whenever there is danger of confusion we explicitly indicate the arguments of differentiation.

For the most part we make no attempt to state results precisely, and give complete references to individual contributions. A detailed analysis of each point made in the paper together with references may be found in the author's book "Constrained Optimization and Lagrange Multiplier Methods", Academic Press, 1982. For surveys of analytical and computational properties of Augmented Lagrangian methods we refer to Bertsekas (1976) and Rockafellar (1976).

2. Augmented Lagrangian Methods

The basic form of the Augmented Lagrangian method consists of solution of a sequence of problems of the form

$$\begin{aligned} &\text{minimize} \quad L_{c_k}(x,\lambda_k) \\ &\text{subject to } x \in R^n \end{aligned} \tag{1}$$

where for $c \geq 0$, $L_c: R^{n+m} \to R$ is the Augmented Lagrangian function

$$L_c(x,\lambda) = f(x) + \lambda' h(x) + \frac{c}{2}|h(x)|^2, \tag{2}$$

and the sequence of penalty parameters $\{c_k\}$ satisfies $0 < c_k \leq c_{k+1}$ for all k. The initial multiplier vector λ_o is given and subsequent multiplier vectors λ_k, $k \geq 1$ are generated by some updating formula such as the first order iteration

$$\lambda_{k+1} = \lambda_k + c_k\, h(x_k) \tag{3}$$

where x_k solves (perhaps approximately) problem (1). There is also a second order iteration

$$\lambda_{k+1} = \bar{\lambda}_k + \Delta\lambda_k \tag{4}$$

where $\bar{\lambda}_k = \lambda_k + c_k h(x_k)$ is the first order iterate, and $\Delta\lambda_k$ together with some vector Δx_k solves the system

$$\begin{bmatrix} H_k & \nabla h(x_k) \\ \nabla h(x_k)' & 0 \end{bmatrix} \begin{bmatrix} \Delta x_k \\ \Delta\lambda_k \end{bmatrix} = - \begin{bmatrix} \nabla_x L(x_k, \bar{\lambda}_k) \\ h(x_k) \end{bmatrix} \tag{5}$$

where H_k is either the Hessian $\nabla^2_{xx} L_o(x_k, \bar{\lambda}_k)$ of the ordinary Lagrangian function L_o evaluated at $(x_k, \bar{\lambda}_k)$, or some Quasi-Newton approximation thereof. Note that the system (5) is also the focal point of RQP methods a fact that points to the significant relations between Augmented Lagrangian methods and RQP.

The convergence properties of the method are quite well understood. There are several results in the literature which state roughly that under second order sufficiency assumptions one can expect convergence of (3) or (4) from an initial multiplier λ_o which is arbitrarily far from a solution provided the penalty parameter c_k becomes eventually sufficiently high. The rate of convergence of $\{x_k, \lambda_k\}$ is typically linear if the simple first order iteration (3) is used and $\{c_k\}$ remains bounded and superlinear otherwise.

There is a large number of variations and extensions of the Augmented La-

grangian method idea. For example extensions are available to handle one-sided or two-sided inequality constraints, as well as nondifferentiable terms in the objective function or constraints. It is possible to use quadratic penalty functions for this purpose although these introduce second derivative discontinuities in the Augmented Lagrangian. An alternative that the author has found useful on several occasions and which does not suffer from this shortcoming is to use one of several possible nonquadratic penalty functions--for example an exponential function. Other variations include alternative stepsize choices in the first order iteration (3), and methods based on partial elimination of constraints. For example if in (ECP) there are additional nonnegativity constraints on x, i.e. the problem has the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && h(x) = 0, \quad x \geq 0, \end{aligned}$$

it may be more convenient to eliminate only the (presumably more difficult) constraints $h(x) = 0$ via a penalty. Minimization of $L_{c_k}(\cdot,\lambda_k)$ should then be carried out subject to the remaining constraints $x \geq 0$. This points to an important advantage of the Augmented Lagrangian method namely the flexibility it affords in changing the structure of a given problem to one that is more favorable. This can prove decisive in solution of large problems where much depends on being able to exploit the existing structure. Finally there is a rich theory associated with Augmented Lagrangian methods which revolves around duality, convexification of nonconvex problems, the proximal point algorithm, and related subjects which can play an important role in analysis of specific problems as well as provide the basis for the development of new algorithms.

Typical advantages cited in favor of the Augmented Lagrangian approach are its robustness, and its ease in programming and tuning for a given problem.

Furthermore the method is broadly applicable since it is capable of solving problems for which the second order sufficiency conditions are not satisfied (although not quite as efficiently as when these conditions are satisfied). Its disadvantages versus other competing methods are primarily in two areas. First feasibility of the generated iterates is not maintained, so if the algorithm is prematurely terminated it will not provide a feasible solution. For some problems this can be an important or even decisive drawback. The second disadvantage manifests itself primarily in small problems and is based on a comparison of the relative efficiency of the method versus RQP techniques. A substantial amount of computational evidence points to the fact that (well tuned) RQP methods require considerably fewer iterations to converge than Augmented Lagrangian methods. On the other hand each iteration of the Augmented Lagrangian method requires less overhead particularly for problems of large dimension. It is difficult to make a precise comparison since much depends on the relative cost of function and derivative evaluations for a given problem. It seems accurate to conclude however that for every type of problem there is a critical size (or dimension) above which either a first order or a second order Augmented Lagrantian method is computationally more efficient than RQP methods, and below which the situation is reversed.

3. Differentiable Exact Penalty Methods

An interesting class of differentiable exact penalty functions for (ECP) was recently introduced by DiPillo and Grippo (1979). Its basic form is

$$P_o(x,\lambda;c) = L_o(x,\lambda) + \frac{c}{2}\left|h(x)\right|^2 + \frac{1}{2}\left|M(x)\nabla_x L_o(x,\lambda)\right|^2. \qquad (6)$$

A more general version which is essentially the same as one proposed in DiPillo, Grippo, and Lampariello (1978) is given by

$$P_\tau(x,\lambda;c) = L_o(x,\lambda) + \frac{c + \tau|\lambda|^2}{2} |h(x)|^2 + \frac{1}{2} |M(x)\nabla_x L_o(x,\lambda)|^2 \quad (7)$$

In (6) and (7) it is assumed that $c > 0$, $\tau \geq 0$ and $M(x)$ is an mxn twice continuously differentiable matrix function on the set $X^* = \{x \mid \nabla h(x) \text{ has rank } m\}$ such that $M(x)\nabla h(x)$ is invertible for all $x \in X^*$. For example one may choose $M(x) = \nabla h(x)'$ or $M(x) = [\nabla h(x)'\nabla h(x)]^{-1}\nabla h(x)'$. When $\tau = 0$ the function (7) is identical to the one of (6) but it seems that the presence of a positive value of τ can have a substantial beneficial effect in algorithmic applications.

The main fact concerning the function (7) is that, roughly speaking, for any value of $\tau \geq 0$, local minima-Lagrange multiplier pairs (x^*,λ^*) of (ECP) can be identified with local minima (with respect to <u>both</u> x and λ) of $P_\tau(\cdot,\cdot;c)$ provided c exceeds a certain threshold value. There is an extensive analysis that clarifies the "equivalence" just stated and quantifies the threshold level for c, but in view of space limitations we cannot go into details. It is worth to point out however that this threshold level depends on eigenvalues of certain second derivative matrices and is largely unrelated to the magnitude of Lagrange multipliers which in turn determines the corresponding threshold level for non-differentiable exact penalty functions.

There is an interesting connection between Newton-like methods for minimizing $P_\tau(\cdot,\cdot;c)$ and Newton's method for solving the (n+m)-dimensional system of necessary conditions $\nabla L_o(x,\lambda) = 0$. It can be shown that the Newton (or second-order RQP) direction

$$d_N = \nabla^2 L_o(x,\lambda)^{-1}\nabla L_o(x,\lambda)$$

can be expressed as

$$d_N = B_\tau(x,\lambda;c)\nabla P_\tau(x,\lambda;c)$$

where $B_\tau(\cdot,\cdot;c)$ is a continuous (n+m)x(n+m) matrix satisfying

$$B_\tau(x^*,\lambda^*;c) = [\nabla^2 P_\tau(x^*,\lambda^*;c)]^{-1}$$

for any local minimum-Lagrange multiplier pair (x^*,λ^*) of (ECP) satisfying the sufficiency assumptions (S). In other words the RQP direction asymptotically [near (x^*,λ^*)] approaches the Newton direction for minimizing $P_\tau(\cdot,\cdot;c)$. This is potentially interesting as it shows that the DiPillo-Grippo penalty function can serve locally [within a neighborhood of (x^*,λ^*)] as a descent function for RQP methods. However a result that is more interesting from a practical point of view is that a descent property of this type holds globally within an arbitrarily large compact subset of X^*. We will present a version of this result shortly in connection with an exact penalty function depending only on x which was introduced in Fletcher (1970).

For $x \in X^*$ consider the function

$$\hat{P}_\tau(x;c) = \min_\lambda P_\tau(x,\lambda;c).$$

Since P_τ is, for each x, a positive definite quadratic function of λ, one can carry out the minimization with respect to λ explicitly. A straightforward calculation yields the minimizing vector

$$\hat{\lambda}(x) = -[\nabla h(x)'M(x)'M(x)\nabla h(x) + \tau|h(x)|^2 I]^{-1} [h(x) + \nabla h(x)'M(x)'M(x)\nabla f(x)]$$

and the equation

$$\hat{P}_\tau(x,\lambda) = P_\tau[x,\hat{\lambda}(x);c],$$

For specific choices of M and τ this equation yields penalty functions in the class of Fletcher (1970). For example if $\tau = 0$ and $M(x) = [\nabla h(x)'\nabla h(x)]^{-1}\nabla h(x)'$ we obtain $\hat{P}_\tau(x,\lambda) = L_o[x,\lambda(x)] + \frac{c-1}{2}|h(x)|^2$ where $\lambda(x) = -[\nabla h(x)'\nabla h(x)]^{-1}\nabla h(x)'\nabla f(x)$.

The penalty function $\hat{P}_\tau(x;c)$ also has nice (and global) descent properties in connection with directions generated by RQP techniques as shown in the following proposition [Bertsekas (1980b),(1982)]:

Proposition 1: Let X be a compact subset of the set $X^* = \{x \mid \nabla h(x) \text{ has rank } m\}$, let H be a bounded set of symmetric nxn matrices, and let $\underline{b}$, $\overline{b}$ be two positive scalars. There exist scalars $\overline{c} > 0$ and $w > 0$ (depending on X, H, $\underline{b}$, $\overline{b}$) such that for every $x \in X$ and every matrix $H \in H$ satisfying

$$\underline{b}\,|z|^2 \leq z'Hz \leq \overline{b}\,|z|^2 \quad , \quad \forall\, z \in R^n \text{ with } \nabla h(x)'z = 0,$$

the solution $(\Delta x, \lambda)$ of the system

$$\begin{bmatrix} H & \nabla h(x) \\ \nabla h(x)' & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \lambda \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ h(x) \end{bmatrix}$$

exists, is unique, and satisfies for all $c \geq \overline{c}$

$$\nabla \hat{P}_\tau(x;c)'\Delta x \leq -w\,|\nabla \hat{P}_\tau(x;c)|^2.$$

Proposition 1 shows that the algorithm

$$x_{k+1} = x_k + \alpha_k \Delta x_k \tag{8}$$

where Δx_k together with some vector λ_{k+1} is obtained by solution of a system of the form

$$\begin{bmatrix} H_k & \nabla h(x_k) \\ \nabla h(x_k)' & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \lambda \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ h(x_k) \end{bmatrix} \tag{9}$$

has global convergence properties provided c is chosen sufficiently large. The following proposition clarifies its superlinear rate of convergence properties. Let us consider the case where α_k is chosen by the Armijo rule with unity initial stepsize, i.e. $\alpha_k = \beta^{m_k}$ where m_k is the first nonnegative integer m satisfying

$$\hat{P}_\tau(x_k;c) - \hat{P}_\tau(x_k+\beta^m \Delta x_k;c) \geq -\sigma\beta^m \nabla \hat{P}_\tau(x_k;c)'\Delta x_k \tag{10}$$

and $\sigma \in (0, \frac{1}{2})$.

Proposition 2: Let x^* be a local minimum of (ECP) which together with a Lagrange multiplier λ^* satisfies the sufficiency assumptions (S). Assume that the algo-

rithm (8)-(10) generates a sequence $\{x_k\}$ converging to x^* and that the sequence $\{H_k\}$ in (9) is bounded and satisfies

$$\frac{\Delta x_k'[H_k - \nabla^2_{xx} L_o(x^*,\lambda^*)]Z^*}{|\Delta x_k|} \to 0$$

where Z^* is an $n \times (n-m)$ matrix the columns of which form a basis for the tangent plane $T^* = \{z \mid \nabla h(x^*)'z = 0\}$. Then:

a) There exists an index $\bar{k}$ such that for all $k \geq \bar{k}$ the stepsize α_k equals unity.

b) The rate of congergence of $\{|x_k - x^*|\}$ is (Q) superlinear.

The conditions of Proposition 2 are always obtained if $H_k = \nabla^2_{xx} L_o(x_k,\lambda_k)$, and usually in practice if H_k is generated by the variable metric formula of Powell (1978).

It is not possible at present to provide a comparison between RQP techniques that use differentiable and nondifferentiable exact penalty functions for descent. Both types of methods behave identically sufficiently close to a solution where the superlinear convergence property takes effect. Far from a solution their behavior can be quite different and furthermore the threshold values for the penalty parameter in both methods can differ greatly on a given problem (these values can have a substantial influence on algorithmic behavior when far from a solution). Methods based on differentiable exact penalty functions require more overhead per iteration in view of the fact that they involve more complex expressions [although not as much overhead as may appear at first sight--see Bertsekas (1982)], and their extensions available at present to deal with inequality constraints are not very "clean". On the other hand they have the theoretical advantage (which may translate into a practical advantage) that they do not require modifications to induce superlinear convergence.

References

Bertsekas, D.P., (1976), "Multiplier Methods: A Survey", Automatica, Vol. 12, pp. 133-145.

Bertsekas, D.P., (1980a), "Enlarging the Region of Congergence of Newton's Method for Constrained Optimization", LIDS Report R-985, M.I.T., Cambridge, Mass. (to appear in J.O.T.A.).

Bertsekas, D.P., (1980b), "Variable Metric Methods for Constrained Optimization Based on Differentiable Exact Penalty Functions", Proc. of Eigtheenth Allerton Conference on Communication, Control and Computing, Allerton Park, Ill., pp. 584-593.

Bertsekas, D.P., (1982), Constrained Optimization and Lagrange Multiplier Methods, Academic Press, N.Y.

Boggs, P.T., and Tolle, J.W., (1980), "Augmented Lagrangians which are Quadratic in the Multiplier", J.O.T.A., Vol. 31, pp. 17-26.

Boggs, P.T., and Tolle, J.W., (1981), "Merit Functions for Nonlinear Programming Problems", Operations Research and Systems Analysis Report, Univ. of North Carolina, Chapel Hill.

Chamberlain, R.M., Lemarechal, C., Pedersen, H.C., and Powell, M.J.D., (1979), "The Watchdog Technique for Forcing Convergence in Algorithms for Constrained Optimization", Presented at the Tenth International Symposium on Mathematical Programming, Montreal.

DiPillo, G and Grippo, L., (1979), "A New Class of Augmented Lagrangians in Nonlinear Programming", SIAM J. on Control and Optimization, Vol. 17, pp. 618-628.

DiPillo, G., Grippo, L., and Lampariello, F., (1979), "A Method for Solving Equality Constrained Optimization Problems by Unconstrained Minimization", Proc. 9th IFIP Conference on Optimization Techniques, Warsaw, Poland.

Dixon, L.C.W., (1980), "On the Convergence Properties of Variable Metric Recursive Quadratic Programming Methods", Numerical Optimization Centre Report No. 110, The Hatfield Polytechnic, Hatfield, England

Fletcher, R., (1970), "A Class of Methods for Nonlinear Programming with Termination and Convergence Propeties", in Integer and Nonlinear Programming (J. Abadie, ed.), North-Holland, Amsterdam.

Han, S.-P., (1977), "A Globally Convergenct Method for Nonlinear Programming", J.O.T.A., Vol. 22, pp. 297-309.

Han, S.-P., and Mangasarian, O.L., (1981), "A Dual Differentiable Exact Penalty Function", Computer Sciences Tech. Report #434, University of Wisconsin.

Maratos, N., (1978), "Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems", Ph.D. Thesis, Imperial College of Science and Technology, University of London.

Mayne, D.Q., and Polak, E., (1978), "A Superlinearly Convergent Algorithm for Constrained Optimization Problems", Research Report 78-52, Department of Computing and Control, Imperial College of Science and Technology, University of London.

Powell, M.J.D., (1978), "Algorithms for Nonlinear Constraints that Use Lagrangian Functions", Math. Programming, Vol. 14, pp. 224-248.

Pschenichny, B.N., (1970), "Algorithms for the General Problem of Mathematical Programming", Kibernetica, pp. 120-125 (Translated in Cybernetics, 1974).

Pschenichny, B.N., and Danilin, Y.M., (1975), Numerical Methods in Extremal Problems, M.I.R. Publishers, Moscow (English Translation 1978).

Rockafellar, R.T., (1976), "Solving a Nonlinear Programming Problem by Way of a Dual Problem", Symposia Mathematica, Vol. XXVII, pp. 135-160.