A GLOBALLY CONVERGENT METHOD FOR NONLINEAR PROGRAMMING[+]

Shih-Ping Han

TR 75-257

August 1975

Department of Computer Science
Cornell University
Ithaca, New York 14853

[+]This research was supported in part by the National Science Foundation under Grant ENG 75-10486.

A GLOBALLY CONVERGENT METHOD FOR NONLINEAR PROGRAMMING+

Shih-Ping Han

ABSTRACT

Recently developed Newton and quasi-Newton methods for nonlinear programming possess only local convergence properties. Adopting the concept of the damped Newton method in unconstrained optimization, we propose a stepsize procedure to maintain monotone decrease of an exact penalty function. In so doing, the convergence of the method is globalized.

KEYWORDS: nonlinear programming, global convergence, exact penalty function.

+This research was supported in part by the National Science Foundation under Grant ENG 75-10486.

A GLOBALLY CONVERGENT METHOD FOR NONLINEAR PROGRAMMING+

Shih-Ping Han

1. Introduction

Consider the nonlinear programming problem

$$(1.1) \qquad \begin{array}{l} \min \ f(x) \\ \text{s.t. } g(x) \leqq 0 \end{array}$$

where $f:R^n \to R$ and $g:R^n \to R^m$. A great deal of attention has been paid to extending Newton and Newton-like methods for solving (1.1). With the efforts of many authors, this attempt has recently achieved some success. One approach on this line is to generate a sequence $\{x^k\}$ converging to the desired solution by means of solving iteratively the quadratic programming problem

$$(1.2) \qquad \begin{array}{l} \min\limits_{x} \ \nabla f(x^k)^T(x-x^k) + \frac{1}{2}(x-x^k)^T H_k (x-x^k) \\ \text{s.t. } g(x^k) + \nabla g(x^k)^T(x-x^k) \leqq 0 \end{array}$$

where the $n \times n$ matrix H_k is intended to be an approximation of the Hessian of the Lagrange $L(x,u) = f(x) + u^T g(x)$. Some results on the convergence and the rate of convergence have been accomplished [1,2,3]. However, as the Newton method in unconstrained optimization, all the results are local. In this work we show that the direction generated by (1.2) turns out to be a descent direction

+This research was supported in part by the National Science Foundation under Grant ENG 75-10486.

of the exact penalty function, $\theta_r : R^n \to R$,

(1.3) $$\theta_r(x) = f(x) + r \sum_{i=1}^{m} g_i(x)_+$$

where $g_i(x)_+ = \max\{0, g_i(x)\}$ and r is a positive number. Consequently, we introduce a procedure by which stepsizes are determined to maintain monotone decrease of this function. With this stepsize procedure the method can be shown globally convergent. In this sense our approach extends the concept of the damped Newton method to the contrained optimization.

For convenience we shall restrict ourselves to problems with inequality constraints only. The inclusion of equality constraints causes no difficulties and all the results go through with minor modification.

We state the method in Section 2 and present global convergence theorems in Section 3. Some comments are in Section 4.

It is noted that all vectors are column vectors and a row vector is denoted by superscript T. The notation $||\cdot||$ denotes a vector norm and also its induced operator norm.

2. Algorithm

Before the statement of the algorithm we first define the following quadratic programming problem Q(x,H):

$$\min_p \ \nabla f(x)^T p + \frac{1}{2} p^T H p$$

$$\text{s.t.} \ \ g(x) + \nabla g(x)^T p \leqq 0$$

which can be associated with any x in R^n and any $n \times n$ matrix H.

Algorithm.

Step 1. Start with a point x^0 in R^n, an $n \times n$ matrix H_0 and a positive number r.

Step 2. Set $k=0$.

Step 3. Having x^k and H_k, find a Kuhn-Tucker point p^k of the quadratic program $Q(x^k, H_k)$.

Step 4. Set $x^{k+1} = x^k + \lambda_k p^k$ with λ_k satisfying

$$\theta_r(x^{k+1}) = \min_{0 \leq \lambda \leq \alpha} \theta_r(x^k + \lambda p^k)$$

where θ_r is defined in (1.3) and α is a fixed number with $0 < \alpha < \infty$.

Step 5. Update H_k by some scheme, set $k = k+1$ and go to Step 3. □

Remarks:

1. It has been shown [3] that when $\{H_k\}$ are generated by some well-known quasi-Newton updates such as the DFP update, the algorithm without the stepsize procedure converges locally with a superlinear rate.

2. The function θ_r is nondifferentiable at some transient surfaces. Since efficient methods for one dimensional minimization of such nondifferentiable functions are available [4], the algorithm is computational implementable.

3. From a different viewpoint the algorithm can also be considered as a descent method for finding a minimum point of the function θ_r. It is noted that for minimizing nondifferentiable functions like θ_r the steepest descent method, even with the exact line-search, may fail to work. The cause of the failure is that the generated sequence may jam into a corner. We refer to [5, pp. 75] for an example. However, in our case, the direction generated by solving $Q(x^k, H_k)$, though not the steepest, is adequate enough to avoid the jamming situation.

3. Global Convergence

For establishing global convergence theorems the concept of a directional derivative and some of its properties are needed. Recall that a directional derivative of a real-valued function h at a point x in the direction p is defined as

$$D_p h(x) = \lim_{t \to 0_+} \frac{h(x + tp) - h(x)}{t} .$$

Clearly, if $D_p h(x) < 0$ then we have $h(x + tp) < h(x)$ for all sufficiently small but nonzero t. The existence of directional derivatives for the function θ_r is insured by the following lemma. We will not give the proof for this lemma but refer to Dem'yanov [5] for a more general and detailed discussion on this result.

Lemma 3.1: If $h_i (i=1,\ldots k)$ are continuously differentiable functions from R^n into R and $\Phi(x) = \max_i \{h_i(x)\}$, then for any direction p the directional derivative $D_p\Phi(x)$ exists and

$$D_p\Phi(x) = \max_{i\in I(x)} \{\nabla h_i(x)^T p\}$$

where $I(x) = \{i: h_i(x) = \Phi(x)\}$. □

Theorem 3.2: Let f and g_i $(i=1,\ldots m)$ be continuously differentiable at x and H be a positive definite $n\times n$ matrix. If (p,u) is a Kuhn-Tucker pair of $Q(x,H)$ with $p\neq 0$ and $||u||_\infty \leqq r$, then $D_p\theta_r(x) < 0$.

Proof: Let $\{I = \{i:g_i(x) > 0\}$, $\bar{I} = \{i:g_i(x) = 0\}$ and $\hat{I} = \{i:g_i(x) < 0\}$. By Lemma 3.1 we have that

$$(3.1) \qquad D_p\theta_r(x) = \nabla f(x)^T p + r \sum_{i\in I} \nabla g_i(x)^T p + r \sum_{i\in\bar{I}} (\nabla g_i(x)^T p)_+ .$$

Since (p,u) is a Kuhn-Tucker pair of $Q(x,H)$, we have that for $i = 1,\ldots m$,

$$(3.2) \quad g_i(x) + \nabla g_i(x)^T p \leqq 0 .$$

Thus,

$$\sum_{i\in\bar{I}} (\nabla g_i(x)^T p)_+ = 0 .$$

Hence by taking $u_i(g_i(x) + \nabla g_i(x)^T p) = 0$ into account we obtain

$$D_p\theta_r(x) = \nabla f(x)^T p + \sum_{i=1}^{m} u_i \nabla g_i(x)^T p + \sum_{i=1}^{m} u_i g_i(x) + r \sum_{i \in I} \nabla g_i(x)^T p.$$

By the Kuhn-Tucker equality

$$\nabla f(x) + \nabla g(x) u + \tfrac{1}{2}(H + H^T)p = 0$$

and by observing that

$$\sum_{i \in \bar{I} \cup \hat{I}} u_i g_i(x) \leqq 0 ,$$

we have

$$D_p\theta_r(x) \leqq - \tfrac{1}{2}p^T(H+H^T)p + \sum_{i \in I} (u_i g_i(x) + r\nabla g_i(x)^T p)$$

$$\leqq - \tfrac{1}{2}p^T(H+H^T)p + \sum_{i \in I} (u_i - r) g_i(x) \qquad \text{(by 3.2)}$$

$$< 0 .$$

(Since H is positive definite and $||u||_\infty < r$) □

Before establishing the global convergence theorems we need a Lemma concerning the perturbation of quadratic programs. The proof can be found in [6].

Lemma 3.3: Let x' minimize $q(x) = \frac{1}{2}x^T H x + b^T x$ over $S = \{x : Ax \leqq a\}$ and $\bar{x}'$ minimize $\bar{q}(x) = \frac{1}{2}x^T \bar{H} x + \bar{b}^T x$ over

$\bar{S} = \{x:\bar{A}x \leqq \bar{a}\}$, where A and $\bar{A}$ are m×n matrices, H and $\bar{H}$ are n×n matrices, a and $\bar{a}$ are in R^m and b and $\bar{b}$ are in R^n. If H is positive definite and $S^o = \{x: Ax < a\} \neq \phi$, then for any fixed norm $||\cdot||$ there exist positive numbers c and $\bar{\varepsilon}$ such that $||\bar{x} - \bar{x}'|| \leqq c\varepsilon$ whenever $\varepsilon \leqq \bar{\varepsilon}$ and $\varepsilon = \max\{||H-\bar{H}||, ||A-\bar{A}||, ||a-\bar{a}||, ||b-\bar{b}||\}$. □

<u>Theorem 3.4</u>: If f and g_i (i=1,...m) are continuously differentiable and the following conditions are satisfied

(i) There exist two positive numbers α and β such that $\alpha x^T x \leqq x^T H_k x \leqq \beta x^T x$ for each k and any x in R^n.

(ii) For each k there exists a Kuhn-Tucker point of $Q(x^k, H_k)$ with a Lagrange multiplier vector bounded by r in ∞-norm.

Then the sequence $\{x^k\}$ generated by the algorithm either terminates at a Kuhn-Tucker point of (1.1) or any accumulation point $\bar{x}$ with $S^o(\bar{x}) = \{p: g(x) + \nabla g(x)^T p < 0\} \neq \phi$ is a Kuhn-Tucker point of (1.1).

<u>Proof</u>: By assumption (ii) we have (p^k, u^k) which is a Kuhn-Tucker pair of $Q(x^k, H_k)$ with $||u^k||_\infty \leqq r$. If $p^k = 0$ then (x^k, u^k) satisfies the Kuhn-Tucker conditions of (1.1) and the sequence terminates at the Kuhn-Tucker point x^k of (1.1). Suppose $p^k \neq 0$ for each k, from Theorem 3.2 and the way we choose x^{k+1} it follows that x^{k+1} exists and

(3.3) $$\theta_r(x^{k+1}) < \theta_r(x^k).$$

Let $\bar{x}$ be an accumulation point of $\{x^k\}$ with $S^o(\bar{x}) \neq \phi$. Without loss of generality we can assume $x^k \to \bar{x}$ and $H_k \to \bar{H}$. The existence of $\bar{H}$ follows from assumption (i); furthermore, $\bar{H}$ is positive definite. It follows from $S^0(\bar{x}) \neq \phi$ and the positive definiteness of $\bar{H}$ that $Q(\bar{x},\bar{H})$ has a unique Kuhn-Tucker point $\bar{p}$. If $\bar{p} = 0$ then $\bar{x}$ is a Kuhn-Tucker point of (1.1) and the theorem follows. Suppose $\bar{p} \neq 0$. By Lemma 3.3 we have $p^k \to \bar{p}$. Since $\{u^k\}$ is uniformly bounded, there exists an accumulation point $\bar{u}$ of $\{u^k\}$. From $x^k \to \bar{x}$, $p^k \to \bar{p}$ and the continuity of gradients of f and g, it follows that $\bar{u}$ is a Lagrange multiplier vector of $Q(\bar{x},\bar{H})$ and $||\bar{u}||_\infty \leqq r$. Let $\bar{\lambda} \in [0,\alpha]$ be chosen such that

$$\theta_r(\bar{x} + \bar{\lambda}\bar{p}) = \min_{0 \leqq \lambda \leqq \alpha} \theta_r(\bar{x} + \lambda\bar{p}).$$

By Theorem 3.2 we have

$$\theta_r(\bar{x} + \bar{\lambda}\bar{p}) < \theta_r(\bar{x}).$$

Since $x^k + \bar{\lambda}p^k \to \bar{x} + \bar{\lambda}\bar{p}$, it follows that for sufficiently large k we have

$$\theta_r(x^k + \bar{\lambda}p^k) < \theta_r(\bar{x}). \tag{3.4}$$

However, by the monotone decrease of $\{\theta_r(x^k)\}$ and the choice of x^{k+1} we have

$$\begin{aligned}\theta_r(\bar{x}) &< \theta_r(x^{k+1}) \\ &\leqq \theta_r(x^k + \bar{\lambda}p^k)\end{aligned}$$

which contradicts (3.4). Hence $\bar{p} = 0$ and $\bar{x}$ is a Kuhn-Tucker point of (1.1). □

Assumption (ii) of Theorem 3.4 is not restrictive as it might appear. In the rest of this section we will give a sufficient condition which ensures the satisfaction of this assumption. First we introduce the following lemma.

Lemma 3.5: Let f and g_i $(i=1,\dots m)$ be continuously differentiable and the following conditions be satisfied

(i) g_i's are convex.

(ii) $X^0 = \{x: g(x) < 0\} \neq \phi$.

(iii) $\alpha y^T y \leqq y^T H y \leqq \beta y^T y$ for some positive numbers α and β and for any y in R^n.

Then for any compact set $U \subset R^n$ there exist $r > 0$ such that if u in R^m is a Lagrange multiplier vector of quadratic program $Q(x,H)$ with x in U then $||u|| \leqq r$ where $||\cdot||$ is any prescribed norm.

Proof: We can assume that H is symmetric. If not, we can replace it by $\frac{1}{2}(H+H^T)$ without affecting the results.

From assumption (ii) of this Lemma there exists at least one point, say $\hat{x}$, in X^0. Let

$$(3.5) \qquad \eta = \max_i\{-g_i(\hat{x})\}$$

and

$$(3.6) \qquad \xi = \max_x\{||x-\hat{x}||_2 : x \in U\}.$$

We further assume that λ is an upper bound of $||\nabla f(x)||_2$ on U and also an upper bound of $||H||_2$ and $||H^{-1}||_2$.

From assumptions it follows that a Kuhn-Tucker point p of Q(x,H) exists and is unique. Let u be a Lagrange multiplier vector of Q(x,H) and $\bar{p} = \hat{x}-x$. By the convexity of g_i's we have that for i=1,...m

(3.7) $\quad g_i(x) + \nabla g_i(x)^T\bar{p} \leqq g_i(\hat{x}) < 0$.

Hence $\bar{p}$ is a feasible point of Q(x,H) and from the Kuhn-Tucker saddle point theorem [7] it follows that

(3.8) $\quad \nabla f(x)^T p + \frac{1}{2}p^T Hp \leqq \nabla f(x)^T\bar{p} + \frac{1}{2}\bar{p}^T H\bar{p} + \sum_{j=1}^{m} u_i(g_i(x) + \nabla g_i(x)^T\bar{p})$.

Thus by (3.8), (3.7) and (3.5) we have

(3.9) $\quad \eta||u||_1 \leqq \nabla f(x)^T\bar{p} + \frac{1}{2}\bar{p}^T H\bar{p} - \nabla f(x)^T p - \frac{1}{2}p^T Hp$.

Now consider the dual problem of Q(x,H)

$$\max_{v \in R^m} \; -(\nabla f(x) + \nabla g(x)v)^T H^{-1}(\nabla f(x) + \nabla g(x)v) + v^T g(x)$$

$$\text{s.t.} \qquad v \geqq 0 .$$

Since $v = 0$ is dual feasible, by the Dorn's duality theorem [7] we have

(3.10) $\quad \nabla f(x)^T p + \frac{1}{2}p^T Hp \geqq -\nabla f(x)^T H^{-1}\nabla f(x)$.

From (3.9) and (3.10) it follows that

$$||u||_1 \leq \frac{1}{\eta}(\nabla f(x)^T\bar{p} + \frac{1}{2}\bar{p}^T H\bar{p} + \nabla f(x)^T H^{-1}\nabla f(x))$$

$$\leq \frac{1}{\eta}(\lambda\xi + \frac{1}{2}\lambda\xi^2 + \frac{1}{2}\lambda\xi^2)$$

$$\leq \frac{\lambda\xi}{\eta}(1+\xi),$$

which by the equivalence of norms implies the desired result. □

We also need the following Lemma on the compactness of some level sets.

Lemma 3.6: If $X = \{x: g(x) \leq 0\}$ is compact and g_i's are lower semi-continuous and convex, then $X_c = \{x: \sum_{i=1}^{m} g_i(x)_+ \leq c\}$ is compact for any finite real number c.

Proof: Define $\Phi(x) = \sum_{i=1}^{m} g_i(x)_+$. By the lower semi-continuity of g_i's, the function Φ is closed and convex. Since $X_0 = X$ is compact, it follows from [8, Lemma 4.1.14, pp. 139] that X_c is compact for any finite c. □

A global convergence theorem is given below.

Theorem 3.7: Let f and g_i (i=1,...m) be continuously differentiable and the following conditions hold

(i) f is bounded below

(ii) g_i's are convex

(iii) $X = \{x: g(x) \leq 0\}$ is compact and $X^o = \{x: g(x) < 0\} \neq \phi$

(iv) There exists positive numbers α and β such that $\alpha x^T x \leqq x^T H_k x \leqq \beta x^T x$ for each k and for any x in R^n.

Then for any starting point x^0 there exists a positive number $\bar{r}$ such that if $r \geqq \max\{\bar{r},1\}$ then the sequence $\{x^k\}$ generated from the algorithm either terminates at a Kuhn-Tucker point of (1.1) or any accumulation point of this sequence is a Kuhn-Tucker point of (1.1).

<u>Proof</u>: It is evident that the sequence exists. By (ii) and (iii) we also have that for any x in R^n the set $S^o(x) = \{p: g(x) + \nabla g(x)^T p < 0\} \neq \phi$. Therefore, we need only to prove that assumption (ii) of Theorem 3.4 holds.

Let x^0 be a given starting point and f be bounded below by $-\alpha$. Define

$$c = f(x^o) + \sum_{i=1}^{m} g_i(x^0)_+ + \alpha .$$

By Lemma 3.6 the set $X_c = \{x: \sum_{i=1}^{m} g_i(x)_+ \leqq c\}$ is compact. Hence it follows from Lemma 3.5 that there exists an $\bar{r} > 0$ such that it $x \in X_c$ and $\alpha y^T y \leqq y^T H y \leqq \beta y^T y$ for any y in R^n then a Lagrange multiplier vector u of $Q(x,H)$ exists and $||u||_\infty < \bar{r}$. Therefore it is only necessary to show that $x^k \in X_c$ for each k. It is clear that $x^0 \in X_c$. Assume $x^k \in X_c$ and $\theta_r(x^k) \leqq \theta_r(x^0)$, then we have $||u^k||_\infty \leqq \bar{r}$. Hence by Theorem 3.2 and the choice of x^{k+1} in the algorithm we have

$$f(x^{k+1}) + r\sum_{i=1}^{m} g_i(x^{k+1})_+ \leqq f(x^k) + r\sum_{i=1}^{m} g_i(x^k)_+$$

$$\leqq f(x^0) + r\sum_{i=1}^{m} g_i(x^0)_+ .$$

Thus,

$$\sum_{i=1}^{m} g_i(x^{k+1})_+ \leqq \frac{1}{r}(f(x^0) - f(x^{k+1})) + \sum_{i=1}^{m} g_i(x^0)_+$$

$$\leqq f(x^0) + \alpha + \sum_{i=1}^{m} g_i(x^0)_+$$

$$\leqq c .$$

Therefore $x^{k+1} \in X_c$ and the proof is completed. □

If we further assume f to be strictly convex, then (1.1) has an unique Kuhn-Tucker point which is actually its optimal solution. Therefore, we have the following result.

Corollary 3.8: Let all the assumptions of Theorem 3.7 hold. If, furthermore, f is strictly convex then the sequence $\{x^k\}$ generated by the algorithm exists and converges to the optimal solution of (1.1). □

4. Comments:

Some comments are given below.

(1) A different way to generate the direction p^k is to solve the dual problem of $Q(x^k, H_k)$

$$\min_{u \in R^m} \frac{1}{2}(\nabla f(x^k) + \nabla g(x^k)u)^T H_k^{-1}(\nabla f(x^k) + \nabla g(x^k)u) - g(x)^T u$$

$$\text{s.t.} \quad u \geqq 0$$

and, with u^k as its solution, set $p^k = -H_k^{-1}(\nabla f(x^k) + \nabla g(x^k)u^k)$. All the results in this work are also valid in this case.

(2) It has been shown [3] that when the DFP update is used to generate the matrices and $\lambda_k = 1$, the method converges locally with a superlinear rate. For this reason, we suggest that the stepsize procedure be discarded when the points are close to the desired solution and the matrices are good approximation to the Hessian of the Lagrangian.

(3) An approximate line-search is desirable. Since the function θ_r is nondifferentiable, some well-known line-search procedures such as Armijo's and Goldstein's no longer work. It is of practical value to develop a workable one.

References

1. ROBINSON, S.M., Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear-Programming Algorithms, Mathematical Programming, Vol. 7, No. 1, pp. 1-16, 1974.

2. GARCIA-PALOMARES, U.M. and MANGASARIAN, O.L., Superlinearly Convergent Quasi-Newton Algorithms for Nonlinearly Constrained Optimization Problems, Computer Science Technical Report #195, University of Wisconsin, Madison, 1974.

3. HAN, S.P., Superlinearly Convergent Variable Metric Algorithms for General Nonlinear Programming Problems, Computer Science Technical Report 75-233, Cornell University, 1975.

4. KOWALIK, J. and OSBORNE, M.R., Methods for Unconstrained Optimization Problems, American Elsevier, New York, 1968.

5. DEM'YANOV, V.F. and MALOZEMOV, V.N., Introduction to Minimax, John Wiley & Sons, New York, 1974.

6. DANIEL, J.M., Stability of the Solution of Definite Quadratic Programs, Mathematical Programming, Vol. 5, No. 1, pp. 41-53, 1973.

7. MANGASARIAN, O.L., Nonlinear Programming, McGraw Hill, New York, 1969.

8. STOER, J. and WITZGALL, C., Convexity and Optimization in Finite Dimensions I, Springer-Verlag, Berlin, 1970.