A SUPERLINEARLY CONVERGENT ALGORITHM FOR MIN-MAX PROBLEMS

by

E. Polak, D. Q. Mayne, and J. E. Higgins

Memorandum No. UCB/ERL M86/103

15 November 1986

COVER PAGE

A SUPERLINEARLY CONVERGENT ALGORITHM FOR MIN-MAX PROBLEMS

by

E. Polak, D. Q. Mayne, and J. E. Higgins

Memorandum No. UCB/ERL M86/103

15 November 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

TITLE PAGE

A SUPERLINEARLY CONVERGENT ALGORITHM FOR MIN-MAX PROBLEMS

by

E. Polak, D. Q. Mayne, and J. E. Higgins

Memorandum No. UCB/ERL M86/103

15 November 1986

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

A Superlinearly Convergent Algorithm for Min-Max Problems[1]

E. Polak[2], D. Q. Mayne[3] and J. E. Higgins[2]

ABSTRACT

Algorithms for solving the problem of minimizing the maximum of a finite number of functions are proposed and analyzed. Quadratic approximations to the functions are employed in the determination of a search direction. Global convergence is proven and it is shown that a quadratic rate of convergence is obtained in the convex case, and a superlinear rate of convergence in the non-convex case.

KEY WORDS

min-max problem, non-differentiable optimization, superlinear convergence.

---

[1] Research sponsored in part by the National Science Foundation Grant ECS-8121149, Air Force Office of Scientific Research Grant 83-0361, Office of Naval Research Contract N00014-83-K-0602, The Semiconductor Research Corp. Contract SRC-82-11-008. The State of California MICRO Program and the General Electric Co.

[2]Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720, U.S.A.

[3]Department of Electrical Engineering, Imperial College of Science and Technology, London, SW7 2BT, Great Britain.

1. INTRODUCTION

We consider the problem:

$$\mathbf{P}: \quad \min_{x \in \mathbb{R}^n} \max_{j \in \underline{m}} f^j(x), \tag{1.1}$$

where the $f^j : \mathbb{R}^n \to \mathbb{R}$ have locally Lipschitz continuous second derivatives and $\underline{m}$ denotes the set $\{ 1, \ldots, m \}$. This problem arises frequently in optimization-based design, see, e.g., [8], but can also arise indirectly when, for example, a constrained optimization problem is converted into an unconstrained problem by the use of an exact penalty function, see e.g., [9]. In work on optimization-based control system design, see [12, 19, 17] for example, it has become evident that existing algorithms perform badly on problems with poor scaling. To mitigate these problems we have developed algorithms which incorporate second derivative information in the search direction calculation. The resulting search directions are well scaled, and the algorithms are, under reasonable assumptions, superlinearly convergent.

Problem **P** can be written in the equivalent form:

$$\min_{x \in \mathbb{R}^n} \psi(x), \tag{1.2}$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$ is defined by:

$$\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x). \tag{1.3}$$

Since ψ is non-differentiable, **P** is a non-differentiable optimization problem and it is not immediately evident that a superlinearly convergent algorithm can be devised. The problem can, of course, be converted into a constrained optimization problem through the employment of an additional variable. The following problem is equivalent to **P**:

$$\min_{(x,w) \in \mathbb{R}^{n+1}} \{ w \mid f^j(x) - w \le 0, j \in \underline{m} \}. \tag{1.4}$$

This formulation reveals that standard algorithms, such as the feasible directions algorithm [11], or the sequential quadratic programming algorithm [9], may be employed to solve **P**. However it is possible and preferable to employ the original formulation. In particular, the drawbacks of using a sequential quadratic programming algorithm (such as [9]) in solving (4) are (i) one must employ a curvilinear step size rule to ensure quadratic convergence, and (ii) the sequential quadratic programming algorithms are not invariant under invertible affine transformations of the parameter space. As we shall see, the algorithm that we shall describe, addresses the original formulation **P**, and does not suffer from these disadvantages.

There is an extensive literature dealing with methods for solving problem **P**. Here we briefly summarize some related work. One of the earliest approaches to solving the problem **P** is Pshenichnyi's linearization method [15], which is a phase I version of the Pironneau-Polak method of centers [10], uses first order expansions of the functions f^j and a fixed quadratic term to obtain a search direction. It is straightforward to demonstrate that, with a suitable step length calculation, any accumulation point produced will satisfy first order optimality conditions. In [15] Pshenichnyi noted that near a *Chebyshev point* (also called a *vertex minimum* and a *Haar point* by other authors) the search direction returned by his method is the same as Newton's method for equalizing the maxima, and hence it follows that if a step size of unity is accepted, then the sequence of iterates converges quadratically. In [2] this result was completed by showing that near such a point that an Armijo step size calculation returns unity. In [1] another first order method is presented, but no rate of convergence is stated. In [6] a method is presented which uses first order expansions of the f^j and a trust region to compute the step size. They note that under conditions similar to those of [2], the method is essentially Newton's method, and so obtain quadratic convergence. In [4] this method is extended by switching to and from a quasi-Newton iteration scheme under certain conditions, in an effort to relax the Chebyshev point assumption. This algorithm has the potentially undesirable property that it is not a descent algorithm, and the convergence properties are somewhat weaker than those of the original algorithm. Another trust region method for minimizing composite functions (of which max functions are a special case) is presented in [3]. The algorithm uses previous multipliers and second order information in the search direction subproblem. The main result of the paper is to show that under suitable assumptions an accumulation point exists at which first order optimality conditions hold. A local quadratic convergence result is also stated, but it is observed that the so called Maratos effect [7] could cause problems by rejecting a unit step near a solution. An active set method which uses reduced Hessian information is proposed in [18].

It is shown that under suitable assumptions any accumulation point satisfies first order optimality conditions. The algorithm is shown to have finite convergence when the f^j are quadratic with identical convex second order terms, and local 2-step superlinear convergence is obtained in certain circumstances. As in [3] it is noted that the Maratos effect could inhibit the superlinear convergence. A general discussion of first order minimax algorithms can be found in [13].

The algorithms presented in this paper employ quadratic approximations to ψ to determine suitable search directions. The first algorithm applies to problems in which the f^j are convex, and has quadratic convergence to Danskin points (points which satisfy first order necessary conditions for problem P) satisfying second order sufficiency conditions. It solves quadratic problems (i.e. each f^j is a convex quadratic) in one step, and under certain conditions the algorithm is scale invariant under invertible affine transformations of the domain. The second algorithm is a modified version for non-convex problems. It is not scale invariant, but it does converge quadratically under appropriate conditions. A third algorithm (also not scale invariant) is presented which yields superlinear convergence under less restrictive conditions. Numerical results are presented which compare the performance of the first algorithm with that of the Pshenichnyi linearization method, showing a significant decrease in the number of iterations required to solve various problems.

2. THE ALGORITHM (Convex Problems)

An obvious quadratic approximation to $\psi(\cdot)$ at x is given by:

$$\psi(x_+) - \psi(x) \approx \max_{j \in \mathbf{m}} \{\tilde{f}^j(x) + \langle \nabla f^j(x), (x_+ - x)\rangle + \frac{1}{2}\langle (x_+ - x), f^j_{xx}(x)(x_+ - x)\rangle\}, \tag{2.1}$$

where

$$\tilde{f}^j(x) \triangleq f^j(x) - \psi(x) \;\; j \in \mathbf{m}. \tag{2.2}$$

As in Newton's method for differentiable functions, a search direction for minimizing $\psi(x)$ can be obtained by minimizing (2.1) with respect to x_+, i.e, by solving

$$\min_{x_+ \in \mathbb{R}^n} \max_{j \in \mathbf{m}} \{\tilde{f}^j(x) + \langle \nabla f^j(x), x_+ - x\rangle + \frac{1}{2}\langle (x_+ - x), f^j_{xx}(x)(x_+ - x)\rangle\}$$

$$= \min_{x_+ \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \sum_{j=1}^{m} \langle \mu^j \nabla f^j(x), (x_+ - x) \rangle + \frac{1}{2} \langle (x_+ - x), l_{xx}(x,\mu)(x_+ - x) \rangle \right\}, \tag{2.3}$$

where

$$\Sigma \triangleq \left\{ \mu \in \mathbb{R}^m \mid \mu \geq 0, \sum_{j=1}^{m} \mu^j = 1 \right\}, \tag{2.4}$$

$$l(x,\mu) \triangleq \langle \mu, f(x) \rangle, \tag{2.5}$$

$$f(x) \triangleq \left[f^1(x), \cdots, f^m(x) \right]^T. \tag{2.6}$$

One of the ways of solving this search direction problem is to apply to its dual a gradient projection method or constrained Newton method [14]. We define the *optimality function* $\theta(\cdot)$ and the *search direction map* $H(\cdot)$ as follows:

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in \mathbf{m}} \left\{ \tilde{f}^j(x) + \langle \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, f^j_{xx}(x) h \rangle \right\}$$

$$= \min_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, l_{xx}(x,\mu) h \rangle \right\}, \tag{2.7}$$

$$H(x) \triangleq \operatorname*{argmin}_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, l_{xx}(x,\mu) h \rangle \right\}. \tag{2.8}$$

The algorithm for solving **P** (where each f^j is convex) may now be stated.

Algorithm 2.1.

Data: $x_0 \in \mathbb{R}^n$, $\alpha, \beta \in (0,1)$, $S \triangleq \{ 1, \beta, \beta^2, \ldots \}$.

Step 0: Set i = 0.

Step 1: Compute $\theta(x_i)$, and $h_i \in H(x_i)$.

Step 2: Compute the step size

$$\lambda_i \triangleq \max\{ \lambda \in S \mid \psi(x_i + \lambda h_i) - \psi(x_i) \leq \lambda \alpha \theta(x_i) \}. \tag{2.9}$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$.

Step 4: Replace i by $i+1$ and go to Step 1. ■

To establish global convergence we make the following assumptions:

Assumption 2.1: The functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \underline{m}$, are twice locally Lipschitz continuously differentiable. ■

Assumption 2.2: There exist $i \in \underline{m}$ and $\eta > 0$ such that

$$\langle h, f^i_{xx}(x)h \rangle \geq \eta \|h\|^2 \qquad \forall\, x, h \in \mathbb{R}^n. \tag{2.10}$$

■

Assumption 2.3: Each function is convex, i.e.

$$\langle h, f^i_{xx}(x)h \rangle \geq 0 \qquad \forall\, x, h \in \mathbb{R}^n, \forall\, i \in \underline{m}. \tag{2.11}$$

■

The main result of this section is that any accumulation point generated by Algorithm 2.1 satisfies a necessary condition of optimality for P. Assumption 2.2 ensures that the optimality function $\theta(\cdot)$ is continuous, and that the search direction map $H(\cdot)$ is u.s.c. Assumption 2.3 is sufficient to prove that the resulting directions yield descent directions for $\psi(\cdot)$. A few preliminary results are necessary. First, we introduce a notation for open balls: $B(x,\rho) \triangleq \{ x' \mid \|x' - x\| < \rho \}$.

Lemma 2.1: Suppose Assumptions 2.1, 2.3 are satisfied, then the function $\theta : \mathbb{R}^n \to \mathbb{R}$ is continuous, and the search direction map $H : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is u.s.c.

Proof: By Assumptions 2.1, 2.2, it is clear that for all bounded sets $K \subset \mathbb{R}^n$ there exists a $\delta > 0$ such that for all $x \in K$ the definitions of $\theta(\cdot)$ and $H(\cdot)$ may be replaced by:

$$\theta(x) \triangleq \min_{h \in B(0,\delta)} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, l_{xx}(x,\mu) h \rangle \right\}, \tag{2.12}$$

$$H(x) \triangleq \operatorname*{argmin}_{h \in B(0,\delta)} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, l_{xx}(x,\mu) h \rangle \right\}. \tag{2.13}$$

Hence by Proposition 3.1 and Corollary 3.1 of [13], $\theta(\cdot)$ is continuous, and by Proposition 3.2 of same, $H(\cdot)$ is u.s.c. ■

Let $d\psi(x;h)$ denote the directional derivative of $\psi(\cdot)$ at x in the direction h.

Lemma 2.2: Suppose Assumptions 2.1-2.3 are satisfied. Then the following are true:

$$\theta(x) \geq d\psi(x;h) \quad \forall\, h \in H(x); \tag{2.14}$$

$$\theta(x) = 0 \text{ if and only if } 0 \in \partial\psi(x). \tag{2.15}$$

Proof: (i) Let the active index set be defined by

$$A(x) \triangleq \{\, j \in \underline{m} \mid f^j(x) = \psi(x) \,\}. \tag{2.16}$$

Then for all $h \in H(x)$ we have that

$$\theta(x) \geq \max_{j \in A(x)} \langle \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, f^j_{xx}(x) h \rangle. \tag{2.17}$$

Because of (2.7), this results in

$$\theta(x) \geq \max_{j \in A(x)} \langle \nabla f^j(x), h \rangle = d\psi(x;h). \tag{2.18}$$

Hence (2.14) is true.

(ii) ($\Leftarrow = =$) First, $0 \in \partial\psi(x)$ holds if and only if $d\psi(x;h) \geq 0$ for all $h \in \mathbb{R}^n$. Hence it follows from (2.14) that if $0 \in \partial\psi(x)$ and $h \in H(x)$, then $0 \leq d\psi(x;h) \leq \theta(x) \leq 0$, which implies that $\theta(x) = 0$.

($= = \Rightarrow$) Note that if $\theta(x) = 0$ then for all $h \in \mathbb{R}^n$,

$$\max_{j \in \underline{m}} f^j(x) - \psi(x) + \langle \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, f^j_{xx}(x) h \rangle \geq 0. \tag{2.19}$$

Consequently, for some $\rho > 0$ and any $h \in B(0,\rho)$, we have

$$\max_{j \in A(x)} \langle \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, f^j_{xx}(x) h \rangle \geq 0. \tag{2.20}$$

Hence there exists a $L > 0$ such that for any $h \in B(0,\rho)$,

$$L\|h\|^2 + \max_{j \in A(x)} \langle \nabla f^j(x), h \rangle \geq 0. \tag{2.21}$$

Therefore for all $\lambda \in [0,1]$ and $h \in B(0,\rho)$,

$$\lambda^2 L\|h\|^2 + \lambda \max_{j \in A(x)} \langle \nabla f^j(x), h \rangle \geq 0, \tag{2.22}$$

from which we may conclude that $d\psi(x;h) \geq 0$ for all $h \in B(0,\rho)$. Since $d\psi(x;\cdot)$ is positive homogeneous, it follows that $d\psi(x;h) \geq 0$ for all $h \in \mathbb{R}^n$, and hence that $0 \in \partial\psi(x)$.

$\blacksquare$

Theorem 2.3: Suppose Assumptions 2.1-2.3 are satisfied. Then any accumulation point $\hat{x}$ of an infinite sequence $\{ x_i \}_{i=0}^{\infty}$, generated by Algorithm 2.1, satisfies $0 \in \partial\psi(\hat{x})$.

Proof: Suppose $x_i \xrightarrow{K} \hat{x}$ where K is some infinite subset of $\mathbb{N}$ and that $0 \notin \partial\psi(\hat{x})$ (equivalently that $\theta(\hat{x}) < 0$). Let $\{ h_i \}_{i=0}^{\infty}$ be the sequence of search directions produced by Algorithm 2.1. As in Lemma 2.1 we note that on bounded sets, the search directions are bounded and, without loss of generality, we assume that $h_i \xrightarrow{K} \hat{h}$. Since $H(\cdot)$ is u.s.c., we have $\hat{h} \in H(\hat{x})$. Since $\theta(\hat{x}) < 0$, $d\psi(\hat{x};\hat{h}) < 0$. Hence by Lemma 2.2 (ii) and by definition of the directional derivative, there exists a $\hat{\lambda} \in S$ such that

$$\psi(\hat{x} + \hat{\lambda}\hat{h}) - \psi(\hat{x}) < \hat{\lambda}\alpha d\psi(\hat{x};\hat{h}) \le \hat{\lambda}\alpha\theta(\hat{x}). \tag{2.23}$$

By continuity of $\psi(\cdot)$ and $\theta(\cdot)$, there exists an $i_0 \in \mathbb{N}$ such that for all $i \ge i_0$, $i \in K$,

$$\psi(x_i + \hat{\lambda}h_i) - \psi(x_i) < \hat{\lambda}\alpha\theta(x_i). \tag{2.24}$$

Therefore $\lambda_i \ge \hat{\lambda}$ for all $i \ge i_0$, $i \in K$. Furthermore, since $\theta(\hat{x}) < 0$, we may assume, without loss of generality, that $\theta(x_i) \le \hat{\delta} < 0$, for all $i \ge i_0$, $i \in K$. Therefore we obtain that

$$\psi(x_{i+1}) - \psi(x_i) \le \hat{\lambda}\alpha\hat{\delta} \quad \forall\, i \ge i_0,\, i \in K. \tag{2.25}$$

Combining (2.25) with the fact that the sequence $\{ \psi(x_i) \}_{i=0}^{\infty}$ is monotonically decreasing, we conclude that $\psi(x_i) \to -\infty$, which contradicts the continuity of $\psi(\cdot)$ at $\hat{x}$. $\blacksquare$

Clearly, $0 \in \partial\psi(\hat{x})$ if and only if there exists a $\hat{\mu} \in \mathbb{R}^m$ satisfying

$$\sum_{j=1}^{m} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{2.26}$$

$$\hat{\mu} \ge 0, \tag{2.27}$$

$$\sum_{j=1}^{m} \hat{\mu}^j = 1, \tag{2.28}$$

$$\hat{\mu}^j [\psi(\hat{x}) - f^j(\hat{x})] = 0 \quad \forall\, j \in \underline{m}. \tag{2.29}$$

Hence, to honor John Danskin who pioneered many results in the theory of minimax, we introduce the following:

Definition 2.1: A point $(\hat{x},\hat{\mu})$ satisfying conditions (2.26) to (2.29) will be referred to as a *Danskin point* for problem **P**. ∎

To demonstrate scale invariance of Algorithm 2.1, we require an additional hypothesis:

Assumption 2.4: The search direction set $H(\cdot)$ is a singleton for all $x \in \mathbb{R}^n$. ∎

Theorem 2.4: Suppose Assumptions 2.1-2.4 are satisfied, and that $L:\mathbb{R}^n \to \mathbb{R}^n$ is an invertible affine transformation. For $j \in \underline{m}$, let $g^j \triangleq f^j \circ L$. Let $y_0 \in \mathbb{R}^n$ be arbitrary and let $x_0 = L(y_0)$. Suppose that Algorithm 2.1 is applied to the functions $\psi_g(x) \triangleq \max_{j \in \underline{m}} g^j(x)$ and $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$, respectively, from the initial points be y_0 and x_0, respectively, and suppose that it produces the sequences $\{ y_n \}_{i=0}^{\infty}$ and $\{ x_n \}_{i=0}^{\infty}$. Then $x_i = L(y_i)$ for all $i \in \mathbb{N}$.

Proof: It is sufficient to show that $x_1 = L(y_1)$, since the rest of the proof then follows by induction. Let A and b be such that $L(y) = Ay + b$. By assumption A is invertible. Let h_g and h_f be the search directions produced by Algorithm 2.1 at y_0, x_0, respectively. Next we show that $h_f = Ah_g$. For $j \in \underline{m}$, let $\bar{g}^j(y) \triangleq g^j(y) - \psi_g(y)$. Then

$$h_g = \operatorname*{argmin}_{h \in \mathbb{R}^n} \max_{i \in \underline{m}} \left\{ \bar{g}^j(y_0) + \langle \nabla g^j(y_0), h \rangle + \frac{1}{2} \langle h, g^j_{xx}(y_0) h \rangle \right\}$$

$$= \operatorname*{argmin}_{h \in \mathbb{R}^n} \max_{j \in \underline{m}} \left\{ \bar{f}^j(L(y_0)) + \langle A^T \nabla f^j(L(y_0)), h \rangle + \frac{1}{2} \langle h, A^T f^j_{xx}(L(y_0)) A h \rangle \right\}$$

$$= A^{-1} \operatorname*{argmin}_{Ah \in \mathbb{R}^n} \max_{i \in \underline{m}} \left\{ \bar{f}^j(x_0) + \langle \nabla f^j(x_0), Ah \rangle + \frac{1}{2} \langle Ah, f^j_{xx}(x_0) Ah \rangle \right\}. \tag{2.30}$$

Because of Assumption 2.4, this implies that $h_f = Ah_g$. Finally, since $L(y_0 + \lambda h_g) = x_0 + \lambda h_f$, it is clear that the same step size is returned in both cases. Hence $x_1 = L(y_1)$. ∎

3. QUADRATIC CONVERGENCE (Convex Problems)

To establish superlinear convergence we require additional assumptions.

Assumption 3.1: Strict complementarity holds at any Danskin point $(\hat{x},\hat{\mu})$, i.e., $f^j(\hat{x}) = \psi(\hat{x})$ implies $\hat{\mu}^j > 0$. ■

Assumption 3.2: At any Danskin point $(\hat{x},\hat{\mu})$ second order sufficiency conditions are satisfied, i.e., there exists a $m_0 > 0$ such that

$$\langle h, l_{xx}(\hat{x},\hat{\mu})h\rangle \geq m_0 \|h\|^2, \tag{3.1}$$

for all $h \in \mathbb{R}^n$ satisfying

$$\langle \nabla f^j(\hat{x}), h\rangle = 0\,, \qquad \forall\, j \in A(\hat{x}), \tag{3.2}$$

where $A(\cdot)$ is the active function set defined by (2.16). ■

There are two main results in this section, the first shows that if algorithm 2.1 is started sufficiently close to a Danskin point and the step size in step 2 is set to 1, then the sequence of iterates converges quadratically to the Danskin point. The second result shows that near a Danskin point satisfying Assumptions 3.1-3.2, the step size λ_i becomes 1. Combining these results with Theorem 2.2 shows that if the sequence of points generated by the algorithm has an accumulation point; then, in fact, it converges quadratically to this point. Observe that Assumption 3.1 implies that at a Danskin point the active gradients are affinely independent (*a set of vectors* $\{ p_i \}$ *are said to be affinely independent if and only if the set of vectors* $\{ (1, p_i) \}$ *are linearly independent*). We now establish that Assumption 3.2 is, in fact, a second-order sufficiency condition.

Proposition 3.1: Let $(\hat{x},\hat{\mu})$ be a Danskin point for problem **P** at which Assumptions 2.1, 3.1 and 3.2 hold. Then $\hat{x}$ is a local minimizer for problem **P**.

Proof: Suppose $\hat{x}$ is not a local minimizer for **P**. Then there exists a sequence $\{ x_i \}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$ and

$$\psi(x_i) < \psi(\hat{x}) \quad \forall\, i. \tag{3.3}$$

Let $\delta x_i \triangleq x_i - \hat{x}$ and $u_i \triangleq \delta x_i / \|\delta x_i\|$. Without loss of generality, we may assume that $u_i \to \hat{u}$. Then we must have that

$$f^j(x_i) - f^j(\hat{x}) = \langle \nabla f^j(\hat{x} + s_i^j \delta x_i), \delta x_i \rangle < 0\,, \tag{3.4}$$

for all $j \in A(\hat{x})$, $i \in \mathbb{N}$ and some $s_i^j \in (0,1)$. Hence, in the limit we must have that

$$\langle \nabla f^j(\hat{x}), \hat{u} \rangle \leq 0 \quad \forall\, j \in A(\hat{x})\,. \tag{3.5}$$

(i) Suppose $\hat{u}$ satisfies $\langle \nabla f^j(\hat{x}), \hat{u} \rangle = 0$ for all $j \in A(\hat{x})$. By Assumption 3.2, we have

$$\langle \hat{u}, l_{xx}(\hat{x}, \hat{\mu}) \hat{u} \rangle \geq m_0 > 0\ . \tag{3.6}$$

Then

$$\begin{aligned}
\psi(x_i) - \psi(\hat{x}) &= \max_{\mu \in \Sigma} \sum_{j=1}^{m} \mu^j [f^j(x_i) - \psi(\hat{x})] \\
&= \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j [f^j(\hat{x}) - \psi(\hat{x})] + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(\hat{x}), \delta x_i \rangle + \int_0^1 (1-s) \langle \delta x_i, l_{xx}(\hat{x} + s\delta x_i, \mu) \delta x_i \rangle ds \right\} \\
&\geq \sum_{j=1}^{m} \hat{\mu}^j [f^j(\hat{x}) - \psi(\hat{x})] + \langle \sum_{j=1}^{m} \hat{\mu}^j \nabla f^j(\hat{x}), \delta x_i \rangle + \int_0^1 (1-s) \langle \delta x_i, l_{xx}(\hat{x} + s\delta x_i, \mu) \delta x_i \rangle ds\,.
\end{aligned} \tag{3.7}$$

The first and second terms are zero by assumption. Hence

$$\psi(x_i) - \psi(\hat{x}) \geq \frac{m_0}{2} \|\delta x_i\|^2, \tag{3.8}$$

for all i sufficiently large, which is a contradiction.

(ii) Suppose $\langle \nabla f^k(\hat{x}), \hat{u} \rangle < 0$ for some $k \in A(\hat{x})$. Then

$$\sum_{j \in A(\hat{x})} \hat{\mu}^j \nabla f^j(\hat{x}) = 0\,, \tag{3.9}$$

$$\hat{\mu}^k > 0\,, \tag{3.10}$$

$$\langle \nabla f^k(\hat{x}), \hat{u} \rangle < 0\,, \tag{3.11}$$

imply that

$$0 > \langle \nabla f^k(\hat{x}), \hat{u} \rangle = - \sum_{\substack{j \in A(\hat{x}) \\ j \neq k}} \langle (\hat{\mu}^j / \hat{\mu}^k) \nabla f^j(\hat{x}), \hat{u} \rangle \geq 0\,, \tag{3.12}$$

which contradicts (3.5). Hence $\hat{x}$ is a local minimizer for P.

■

The proof of local quadratic convergence follows the ideas in [16]. It differs in that our proof deals directly with the max function formulation, whereas the proof in [16] centers around a constrained differentiable problem. A version of the Implicit Function Theorem given in [5] is utilized, and is restated here for convenience. In fact, the statement of the Implicit Function Theorem in [5] can be strengthened slightly without any change in the proof, and this modification is included below.

Theorem 3.2: ([5, Theorem 1, section 4.XVII]) Suppose that (i) $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ is a neighborhood of the point $(\hat{x},\hat{y})$, (ii) the function $g:\Omega \to \mathbb{R}^n$ is continuous at $(\hat{x},\hat{y})$, and (iii) the following hold

$$g(\hat{x},\hat{y}) = 0; \tag{3.13}$$

$$g_x(\cdot,\cdot) \text{ exists on } \Omega \text{ and is continuous at } (\hat{x},\hat{y}); \tag{3.14}$$

$$g_x(\hat{x},\hat{y}) \text{ is invertible.} \tag{3.15}$$

Then there exist neighborhoods $N \subset \mathbb{R}^n$ and $M \subset \mathbb{R}^m$ of $\hat{x}$ and $\hat{y}$ respectively, such that $N \times M \subset \Omega$, and a function $Z: M \to N$ with the following properties:

$$g(Z(y),y) = 0 \quad \forall\, y \in M; \tag{3.16}$$

$$Z(\hat{y}) = \hat{x}; \tag{3.17}$$

$$Z(\cdot) \text{ is continuous at } \hat{y}; \tag{3.18}$$

$$\forall\, y \in M,\ Z(y) \text{ is the unique zero of } g(\cdot,y) \text{ in } N. \tag{3.19}$$

For convenience we adopt the notation $z = (x,\mu) \in \mathbb{R}^{n+m}$. To continue, we observe that if, with $h = (x_+ - x)$, $(x_+ - x), \mu_+)$ is a Danskin point for the search direction problem (2.8), then

$$\sum_{j=1}^{m} \mu_+^j \{ \nabla f^j(x) + f^j_{xx}(x)(x_+ - x) \} = 0, \tag{3.20}$$

$$\mu_+ \geq 0, \tag{3.21}$$

$$\sum_{j=1}^{m} \mu_+^j = 1, \tag{3.22}$$

$$\mu_+^j \{ \theta(x) - [f^j(x) - \psi(x) + \langle \nabla f^j(x), x_+ - x\rangle + \frac{1}{2} \langle (x_+ - x), f^j_{xx}(x)(x_+ - x) \rangle] \} = 0, \ \forall\, j \in \underline{m}. \tag{3.23}$$

Furthermore, summing (3.23) over $j \in \underline{m}$, we obtain that

$$\theta(x) = \sum_{j=1}^{m} [f^j(x) - \psi(x) + \langle \nabla f^j(x), x_+ - x\rangle + \frac{1}{2}\langle (x_+ - x), f^j_{xx}(x)(x_+ - x)\rangle] \,. \tag{3.24}$$

Condition (3.23) is equivalent to:

$$\mu^j_+\{ [f^k(x) - \psi(x) + \langle \nabla f^k(x), (x_+ - x)\rangle + \frac{1}{2}\langle (x_+ - x), f^k_{xx}(x)(x_+ - x)\rangle] -$$

$$[f^j(x) - \psi(x) + \langle \nabla f^j(x), (x_+ - x)\rangle + \frac{1}{2}\langle (x_+ - x), f^j_{xx}(x)(x_+ - x)\rangle] \} \leq 0\,,\ \forall\, j, k \in \underline{m}\,. \tag{3.25}$$

In addition it should be clear that if the convexity Assumption 2.3 is satisfied, then conditions 3.20-3.23 are also sufficient for $x_+ - x$ to solve the search direction problem.

Suppose that $\hat{z}$ is a Danskin point for problem P at which Assumption 2.1 holds. Without loss of generality, we may assume that $A(\hat{x}) = \{ 1,\ldots,r \}$. We now define the function $g : \mathbb{R}^{n+m}\times\mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ as follows:

$$g(z_1, z_2) \triangleq \begin{bmatrix} \sum_{j=1}^{m} \mu^j_1\{\nabla f^j(x_2) + f^j_{xx}(x_2)(x_1 - x_2)\} \\ \mu^1_1[\bar{f}^{r,1}(x_2) + \langle \nabla \bar{f}^{r,1}(x_2), x_1 - x_2\rangle + \frac{1}{2}\langle (x_1 - x_2), \bar{f}^{r,1}_{xx}(x_2)(x_1 - x_2)\rangle] \\ \vdots \\ \mu^{r-1}_1[\bar{f}^{r,r-1}(x_2) + \langle \nabla \bar{f}^{r,r-1}(x_2), x_1 - x_2\rangle + \frac{1}{2}\langle (x_1 - x_2), \bar{f}^{r,r-1}_{xx}(x_2)(x_1 - x_2)\rangle] \\ \sum_{j=1}^{m} \mu^j_1 - 1 \\ \mu^{r+1}_1[\bar{f}^{r,r+1}(x_2) + \langle \nabla \bar{f}^{r,r+1}(x_2), (x_1 - x_2)\rangle + \frac{1}{2}\langle (x_1 - x_2), \bar{f}^{r,r+1}_{xx}(x_2)(x_1 - x_2)\rangle] \\ \vdots \\ \mu^m_1[\bar{f}^{r,m}(x_2) + \langle \nabla \bar{f}^{r,m}(x_2), (x_1 - x_2)\rangle + \frac{1}{2}\langle (x_1 - x_2), \bar{f}^{r,m}_{xx}(x_2)(x_1 - x_2)\rangle] \end{bmatrix} \tag{3.26}$$

where $\bar{f}^{ij} \triangleq f^i - f^j$ and $\nabla \bar{f}^{ij}$ and $\bar{f}^{ij}_{xx}$ are defined similarly. The following result summarizes the useful characteristics of the function $g(\cdot,\cdot)$.

Proposition 3.3: Let $\hat{z}$ be a Danskin point for P, and suppose that Assumptions 2.1-2.3 and 3.1-3.2

are satisfied. Let $g(\cdot,\cdot)$ be defined as in (3.26). Then the following are true:

(i) $g(\hat z,\hat z) = 0$;

(ii) $g_{z_1}(\cdot,\cdot)$ exists and is continuous in some neighborhood of $(\hat z,\hat z)$ and is invertible at $(\hat z,\hat z)$;

(iii) There exist neighborhoods M and N of $\hat z$ and a function $Z(\cdot)$ satisfying (3.16) - (3.19);

(iv) $Z((x,\mu)) = Z((x,\bar\mu)) \quad \forall\ \mu, \bar\mu \in \mathbb{R}^m$ such that $(x,\mu), (x,\bar\mu) \in M$;

(v) There exists a neighborhood $U \subset M$ containing $\hat z$ such that $\forall\ z \in U$, $Z(z)$ is the (unique) solution to the search direction problem (2.8);

(vi) There exists a neighborhood G of $\hat x$ such that $\forall\ x \in G$, the search direction map $H(x)$ is a singleton. Furthermore, $H(\hat x) = \{ 0 \}$ and $H(\cdot)$ is continuous at $\hat x$;

(vii) The multiplier vector μ_+ defined by $(x_+,\mu_+) = Z((x,\mu))$ are continuous in x at $\hat x$.

(viii) The vector $(x_+,\mu_+) = Z((x,\mu))$ satisfies (3.20) - (3.23).

Proof: That (i) is true follows immediately from the definition of a Danskin point and the fact that $A(\hat x) = \{ 1,\dots,r \}$. It is straightforward to verify that $g_{z_1}(\cdot,\cdot)$ exists and is continuous in some neighborhood of $(\hat z,\hat z)$. To show that $g_{z_1}(\hat z,\hat z)$ is invertible, observe that

$$g_{z_1}(\hat z,\hat z) = \begin{bmatrix} l_{xx}(\hat z) & \nabla f^1(\hat x) & \cdots & \nabla f^{r-1}(\hat x) & \nabla f^r(\hat x) & \cdots & \nabla f^m(\hat x) \\ \hat\mu^1 \nabla \bar f^{r,1}(\hat x)^T & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \hat\mu^{r-1} \nabla \bar f^{r,r-1}(\hat x)^T & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 0 & \bar f^{r,r+1}(\hat x) & 0 \\ \vdots & & & & & & \\ 0 & 0 & \cdots & 0 & 0 & \cdots & \bar f^{r,m}(\hat x) \end{bmatrix} \tag{3.27}$$

Suppose $g_{z_1}(\hat z,\hat z)(h,v)^T = 0$. Then since $\bar f^{r,i}(\hat x) > 0$ for all $i \notin A(\hat x)$ it is clear that $v_{r+1} = v_{r+2} = \dots = v_m = 0$. Next $\hat\mu^i \langle \nabla \bar f^{r,i}(\hat x), h \rangle = 0$ for all $i \in \underline{r-1}$ and, since the strict complementarity Assumption 3.1 implies that $\hat\mu^i > 0$ for all $i \in \underline{r}$, it follows that $\langle \nabla \bar f^{r,i}(\hat x), h \rangle = 0$ for all $i \in \underline{r-1}$. In addition, since $\hat z$ is a Danskin point, $\sum_{i=1}^{r} \hat\mu^i \nabla f^i(\hat x) = 0$. Combining these facts we have

$$\langle \sum_{i=1}^{r} \hat{\mu}^i \rangle \langle \nabla f^r(\hat{x}), h \rangle = 0, \tag{3.28}$$

which implies that $\langle \nabla f^i(\hat{x}), h \rangle = 0, \ \forall \ i \in \underline{r}$. Since $v_{r+1}, \ldots, v_m$ are all zero, the top n components of $g_{z_1}(\hat{z}, \hat{z})(h, v)^T$ are given by

$$\sum_{i=1}^{r} \hat{\mu}^i f^i_{xx}(\hat{x}) h + \sum_{i=1}^{r} \hat{v}^i \nabla f^i(\hat{x}) = 0. \tag{3.29}$$

Hence $\langle h, \sum_{i=1}^{r} \hat{\mu}^i f^i_{xx}(\hat{x}) h \rangle = 0$. It now follows from Assumption 3.2 that $h = 0$. It remains to show that $v_1 = v_2 = \ldots = v_r = 0$. Since $h = 0$, it follows from (3.29)

$$\sum_{i=1}^{r} \hat{v}^i \nabla f^i(\hat{x}) = 0. \tag{3.30}$$

Next, since $g_{z_1}(\hat{z}, \hat{z})(h, v)^T = 0$, it follows that

$$\sum_{i=1}^{r} \hat{v}^i = 0. \tag{3.31}$$

Since by Assumption 3.1, $\{ \nabla f^i(\hat{x}) \}_{i \in A(\hat{x})}$ are affinely independent, (3.30), (3.31) imply that $v_1 = v_2 = \ldots = v_r = 0$. Hence we conclude that $g_{z_1}(\hat{z}, \hat{z})$ is invertible, and so (ii) is proved. Consequently it is clear that $g(\cdot, \cdot)$ satisfies the assumptions of Theorem 3.2, thus proving (iii).

To prove (iv), we deduce from the definition (3.26) that $g(\bar{z}, (x, \mu)) = g(\bar{z}, (x, \underline{\mu})) \ \forall \ \bar{z}, x, \mu, \underline{\mu}$. Suppose that $(x, \mu), (x, \underline{\mu}) \in M$. Then by definition of $Z(\cdot)$ we have $g(Z((x, \mu)), (x, \mu)) = 0$, and by the above remark, we see that $g(Z((x, \mu)), (x, \underline{\mu})) = 0$. (3.19) in Theorem 3.2 allows us to conclude that (iv) is true. More generally, it should be clear that this result implies that $Z(\cdot)$ depends only on it's first argument (x).

Because of (3.18) in Theorem 3.2, and of the fact that $\overline{f}^{\,i}(\hat{x}) > 0 \ \forall \ i \notin A(\hat{x})$, we may chose a neighborhood $U \subset M$ of $\hat{z}$, such that for all $z \in U$ and $(x_+, \mu_+) = Z(z)$, $\mu_+^1, \mu_+^2, \ldots, \mu_+^r > 0$, and $\overline{f}^{\,i}(x) + \langle \nabla \overline{f}^{\,i}(x), x_+ - x \rangle + \frac{1}{2} \langle (x_+ - x) \overline{f}^{\,i}_{xx}(x)(x_+ - x) \rangle > 0$ for all $i \notin A(\hat{x})$. Since $g(Z(z), z) = 0$, we see that z_+ satisfies 3.20-3.23 and therefore is a Danskin point for the search direction problem (2.8). In addition, it is the only such point in the neighborhood N of part (iii). By Assumption 2.3, the search direction problem is convex. Hence we conclude that $x_+ - x$ is the unique solution to the search direction problem. This proves (v).

Since U in (v) is an open set containing $\hat{z}$, it should be clear that for some $\varepsilon > 0$, the set $B(\hat{x},\varepsilon) \times \{ \hat{\mu} \}$ is contained in U. Let $G \triangleq B(\hat{x},\varepsilon)$. Suppose $x \in G$. Then clearly $(x,\hat{\mu}) \in U$ and therefore, by (v), $Z((x,\hat{\mu}))$ is the unique solution to the search direction problem (2.8). Hence $H(x)$ is a singleton. Since by Lemma 2.1 $H(\cdot)$ is u.s.c., it follows that it is continuous on G. Since $Z(\hat{z}) = \hat{z}$, it follows that $H(\hat{x}) = \{ 0 \}$, and so (vi) is true.

It is obvious from (3.18) in Theorem 3.2 that $Z(\cdot)$ is continuous at $\hat{z}$. To prove that (vii) is true, it suffices to show that $Z(\cdot)$ depends only on x in some neighborhood of $\hat{x}$. Let $x \in G$, then as above we have $(x,\hat{\mu}) \in U$, and so $Z((x,\hat{\mu}))$ is well defined. Because of (iv) and the fact that $g(\bar{z},(x,\mu)) = g(\bar{z},(x,\Pi))$ for all $\bar{z},x,\mu,\Pi$, we obtain that $g(Z((x,\hat{\mu})),(x,\mu)) = 0$ for all $x \in G$, $\forall\ \mu$. Part (viii) follows immediately from the fact that $g(Z(z),z) = 0$. This concludes the proof. ■

In light of Proposition 3.3, Proposition 3.1 can be strengthened as follows.

Proposition 3.4: Let $\hat{z}$ be a Danskin point for problem **P** at which Assumptions 2.1, 3.1 and 3.2 hold, then $\hat{x}$ is a local minimizer for problem **P**, and there exists a neighborhood W such that $\hat{z}$ is the unique Danskin point of **P** in W.

Proof: Define $G(z) \triangleq g(z,z)$, where $g(\cdot,\cdot)$ is defined by (3.26). Notice that $G_z(\hat{z}) = g_{z_1}(\hat{z},\hat{z})$, and so it is invertible. Thus by the inverse function Theorem, there exists some neighborhood O of $\hat{z}$, such that $\hat{z}$ is the unique zero of $G(\cdot)$ in O. Furthermore, by choosing $W \subset O$ sufficiently small, we may ensure that $\bar{f}^j(x) > 0\ \forall\ i \notin A(\hat{x})$, and $\mu^1,\mu^2,\ldots,\mu^r > 0$, $\quad \forall\ (x,\mu) \in O$. Hence if $z \in O$ is a Danskin point of **P**, then it should be clear that $g(z) = 0$ and hence $z = \hat{z}$. This proves the Proposition. ■

Note that Proposition 3.4 implies that $\hat{x}$ is an isolated local minimizer for **P**. To provide quantitative bounds on the variation of $Z(\cdot)$, we use a version of Robinson's Theorem 2.2 [16]. Our proof is somewhat simpler than that in [16].

Proposition 3.5: ([16, Theorem 2.2]) Suppose that (a) $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ is a neighborhood of the point $(\hat{x},\hat{y})$, (b) the function $g:\Omega \to \mathbb{R}^n$ is continuous at $(\hat{x},\hat{y})$, and (c) the following hold:

$$g(\hat{x},\hat{y}) = 0; \tag{3.32}$$

$$g_x(\cdot,\cdot) \text{ exists on } \Omega \text{ and is continuous at } (\hat{x},\hat{y}); \tag{3.33}$$

$$g_x(\hat{x},\hat{y}) \text{ is invertible.} \tag{3.34}$$

Then there exist neighborhoods $N \subset \mathbb{R}^n$ and $M \subset \mathbb{R}^m$ of $\hat{x}$ and $\hat{y}$ respectively, and a function $Z: M \to N$ such that (3.16)-(3.19) are satisfied. Furthermore, for every $\delta \in (0,1)$, there exist neighborhoods $\overline{N} \subset N$ and $\overline{M} \subset M$ of $\hat{x}$ and $\hat{y}$ respectively, such that

$$\|x - Z(y)\| \le \frac{1}{1-\delta} \|g_x(\hat{x},\hat{y})^{-1}\| \|g(x,y)\| \quad \forall\ (x,y) \in \overline{N} \times \overline{M}. \tag{3.35}$$

Proof: The existence of neighborhoods N, M and of the function $Z(\cdot)$ follows directly from Theorem 3.2. Now, let $\delta \in (0,1)$, $\Gamma \triangleq g_x(\hat{x},\hat{y})^{-1}$, and select $\overline{N} \subset N$ and $\overline{M} \subset M$ to be open neighborhoods of $\hat{x}$ and $\hat{y}$ respectively, such that

$$\|g_x(\hat{x},\hat{y}) - g_x(x,y)\| \le \frac{\delta}{\|\Gamma\|} \quad \forall\ (x,y) \in \overline{N} \times \overline{M}. \tag{3.36}$$

Without loss of generality, $\overline{M}$ and $\overline{N}$ may be taken to be convex, and $\overline{M}$ may be taken to be small enough to ensure that $Z(\overline{M}) \subset \overline{N}$. Following Robinson [16], we define $\Phi^y(x) \triangleq x - \Gamma g(x,y)$. For $(x,y) \in \overline{N} \times \overline{M}$, $\Phi^y(\cdot)$ is differentiable, and:

$$\Phi^y_x(x) = I - \Gamma g_x(x,y) \tag{3.37}$$

$$= \Gamma (g_x(\hat{x},\hat{y}) - g_x(x,y)). \tag{3.38}$$

It follows from (3.38) that $\|\Phi^y_x(x)\| \le \delta$, and since $\Phi^y(x_1) - \Phi^y(x_2) = \int_0^1 \Phi^y_x(x_1 + s(x_2 - x_1))ds(x_1 - x_1)$,

$$\|\Phi^y(x_1) - \Phi^y(x_2)\| \le \delta \|x_1 - x_2\| \quad \forall\ x_1,x_2 \in \overline{N} \quad \forall\ y \in \overline{M}, \tag{3.39}$$

where the norm on $\Phi^y_x(\cdot)$ is the induced norm. Substituting for $\Phi^y(\cdot)$ gives

$$\|x_1 - \Gamma g(x_1,y) - (x_2 - \Gamma g(x_2,y))\| \le \delta \|x_1 - x_2\|. \tag{3.40}$$

Using the inequality $|a| - |b| \le |a - b|$, we obtain from (3.40)

$$|x_1 - x_2| - \|\Gamma(g(x_1,y) - g(x_2,y))\| \le \delta |x_1 - x_2| , \tag{3.41}$$

from which it follows that

$$|x_1 - x_2| \frac{(1-\delta)}{\|\Gamma\|} \le |g(x_1,y) - g(x_2,y)| . \tag{3.42}$$

By letting $x_1 = x$ and $x_2 = Z(y)$, and noting that $g(x_2,y) = 0$, we obtain the required result. ■

To complete the derivation of the bound on the variation of $Z(\cdot)$, a growth condition on $g(\cdot,\cdot)$ must be established.

Lemma 3.6: Suppose Assumptions 2.1-2.3 and 3.1-3.2 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Let $g(\cdot,\cdot)$ be defined as in (3.26). Then there exists a neighborhood $K \subset \mathbb{R}^{n+m}$ of $\hat{z}$ and a $L > 0$ such that

$$|g(z_1,z_1) - g(z_1,z_2)| \le L |z_1 - z_2|^2 \quad \forall\, z_1,z_2 \in K . \tag{3.43}$$

Proof: The proof is obtained by expanding (3.26), and noting that by Assumption 2.1, the matrix valued functions $f^i_{xx}(\cdot)$, $i \in \underline{m}$, are locally Lipschitz continuous. ■

We are now ready to establish the first main result of this section.

Theorem 3.7: Suppose that Assumptions 2.1-2.3 and 3.1-3.2 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Let $g(\cdot,\cdot)$ be defined as in (3.26), and let $M \times N \subset \mathbb{R}^{n+m} \times \mathbb{R}^{n+m}$, be a neighborhood of $\hat{z}$ $Z: M \to N$ be such that (i) $g(Z(z),z) = 0 \quad \forall\, z \in M$, (ii) $Z(\hat{z}) = \hat{z}$, (iii) $Z(\cdot)$ is continuous at $\hat{z}$, and (iv) for all $z \in M$, $Z(z)$ is the unique zero of $g(\cdot,z)$ in N.

Then there exists a neighborhood V of $\hat{z}$ such that if $z_0 \in V$, then the sequence $\{ z_i \}_{i=0}^{\infty}$ defined by $z_{i+1} \triangleq Z(z_i)$ is well defined, and converges to $\hat{z}$ with R-order at least 2 (i.e. quadratically). Specifically, there exist constants $\eta \in (0,1)$, $\bar{K} > 0$ such that

$$|\hat{z} - z_i| \le \bar{K} \eta^{2^i} , \ \forall\, i \ge 0 . \tag{3.44}$$

Proof: Let $\varepsilon \in (0,1)$ be such that $\bar{B}(\hat{z},\varepsilon) \subset K \cap M \cap W$, where K, and W are the open neighborhoods of $\hat{z}$ specified in Lemma 3.6 and Proposition 3.4 respectively. Let $\gamma \in (0,1)$, and L be as (3.43). Suppose, in addition, that ε is small enough to ensure that $Z^2(\cdot)$ (i.e. the composition of Z with itself) is well defined on $\bar{B}(\hat{z},\varepsilon)$, and that for all $z \in \bar{B}(\hat{z},\varepsilon)$ both z and $Z(z)$ are contained in $\bar{N} \cap \bar{M} \cap K$, where $\bar{N}$ and $\bar{M}$ are the neighborhoods defined in Proposition 3.5 for $\delta = 1/2$. Let $\bar{L} \triangleq 2\|g_{z_1}(\hat{z},\hat{z})^{-1}\|L$ and let $\rho > 0$ such that for all $z \in B(\hat{z},\rho)$,

$$\|\hat{z} - z\| \le \varepsilon/2\,, \tag{3.45}$$

$$\|Z(z) - z\| \le \gamma\min\{\frac{1}{\bar{L}},\frac{\varepsilon/2}{1+2\bar{L}}\}\,. \tag{3.46}$$

Let $V \triangleq B(\hat{z},\rho)$. To continue, we must establish that the sequence $\{ z_i \}_{i=0}^{\infty}$ is well defined. Note that ε has been selected so that for all $z \in \bar{B}(\hat{z},\varepsilon)$,

$$\begin{aligned}\|Z^2(z) - Z(z)\| &\le \frac{1}{1-1/2}\|g_{z_1}(\hat{z},\hat{z})^{-1}\|\|g(Z(z),Z(z))\| \\ &\le 2\|g_{z_1}(\hat{z},\hat{z})^{-1}\|\|g(Z(z),Z(z)) - g(Z(z),z)\| \\ &\le 2\|g_{z_1}(\hat{z},\hat{z})^{-1}\|L\|Z(z) - z\|^2\,.\end{aligned} \tag{3.47}$$

Suppose $z_0 \in \bar{B}(\hat{z},\rho)$. Then $z_1 = Z(z_0)$ is well defined. Let $\alpha \triangleq \bar{L}\|z_1 - z_0\|$, then by construction, $\alpha < 1$, and hence for all $n \ge 0$

$$\sum_{i=0}^{n}\alpha^i \le \frac{1}{(1-\alpha)}\,. \tag{3.48}$$

Then

$$\|z_1 - z_0\| \le \gamma\varepsilon/2 \le \varepsilon/2\,, \tag{3.49}$$

$$\|z_1 - \hat{z}\| \le \|z_1 - z_0\| + \|z_0 - \hat{z}\| \le \varepsilon\,. \tag{3.50}$$

Therefore $z_1 \in B(\hat{z},\varepsilon)$, and so z_2 is well defined. We now proceed by induction. We have

$$\begin{aligned}\|z_2 - z_1\| &\le \bar{L}\|z_1 - z_0\|^2 \\ &\le \alpha\,\|z_1 - z_0\|\,,\end{aligned} \tag{3.51}$$

and

$$|z_2 - \hat{z}| \le |z_2 - z_1| + |z_1 - z_0| + |z_0 - \hat{z}| \tag{3.52}$$

$$\le (\alpha+1)|z_1 - z_0| + \varepsilon/2. \tag{3.53}$$

By using (3.48) we obtain

$$|z_2 - \hat{z}| \le \frac{|z_1 - z_0|}{1-\bar{L}|z_1 - z_0|} + \varepsilon/2$$

$$\le \frac{(\varepsilon/2)/(1+2\bar{L})}{1-\bar{L}(\varepsilon/2)/(1+2\bar{L})} + \varepsilon/2 \le \varepsilon. \tag{3.54}$$

Now suppose that for some $k \ge 2$, (i) $z_i \in \bar{B}(\hat{z},\varepsilon) \ \forall\ i \le k$, and (ii) $|z_{i+1} - z_i| \le \alpha |z_i - z_{i-1}|$ for all $i \in \underline{k-1}$. Then clearly z_{k+1} is well defined, and as above

$$|z_{k+1} - \hat{z}| \le \sum_{i=1}^{k+1} |z_i - z_{i-1}| + |z_0 - \hat{z}|$$

$$\le \sum_{i=0}^{k} \alpha^i\, |z_1 - z_0| + \varepsilon/2$$

$$\le \frac{|z_1 - z_0|}{1-\alpha} + \varepsilon/2 \le \varepsilon. \tag{3.55}$$

Hence $z_{k+1} \in \bar{B}(\hat{z},\varepsilon)$, and in addition

$$|z_{k+1} - z_k| \le \bar{L}|z_k - z_{k-1}|^2 \le \bar{L}|z_1 - z_0|\,|z_k - z_{k-1}| \le \alpha |z_k - z_{k-1}|. \tag{3.56}$$

It follows from this that the sequence $\{ z_i \}_{i=0}^{\infty}$ is well defined. To show that the sequence converges, suppose that $k,l \ge 0$. Then

$$|z_{k+l} - z_k| \le \sum_{i=k}^{k+l-1} |z_{i+1} - z_i| \le \sum_{i=k}^{k+l-1} \alpha^i |z_1 - z_0| \le \frac{\alpha^k}{1-\alpha} |z_1 - z_0|. \tag{3.57}$$

Hence the sequence is Cauchy, and so converges to some $\bar{z} \in \bar{B}(\hat{z},\varepsilon)$. We may conclude from this that $g(\bar{z},\bar{z}) = 0$, and by Proposition 3.4 that $\bar{z} = \hat{z}$.

To prove quadratic convergence, note that

$$\|z_2 - z_1\| \le \bar{L}\|z_1 - z_0\|^2 \le \frac{1}{\bar{L}}\alpha^2. \tag{3.58}$$

Suppose that for $k \ge 1$, $\|z_k - z_{k-1}\| \le \alpha^{2^{k-1}}/\bar{L}$. Then

$$\|z_{k+1} - z_k\| \le \bar{L}\|z_k - z_{k-1}\|^2 \le \bar{L}(\frac{1}{\bar{L}}\alpha^{2^{k-1}})^2 \le \frac{1}{\bar{L}}\alpha^{2^k}. \tag{3.59}$$

Furthermore since $\hat{z} - z_k = \sum_{i=k}^{\infty}(z_{i+1} - z_i)$, we have the following estimate

$$\|\hat{z} - z_k\| \le \sum_{i=k}^{\infty}\|z_{i+1} - z_i\| \le \frac{1}{\bar{L}}\sum_{i=k}^{\infty}\alpha^{2^i} \le \frac{\alpha^{2^k}}{\bar{L}}\sum_{i=0}^{\infty}\alpha^{2^k(2^i - 1)}, \tag{3.60}$$

and since $2^k(2^i - 1) \ge i$, for all $i,k \ge 0$, we have

$$\|z_{k+1} - z_k\| \le \frac{1}{\bar{L}(1 - \alpha)}\alpha^{2^k}. \tag{3.61}$$

Letting $\bar{K} \triangleq \dfrac{1}{\bar{L}(1 - \alpha)}$ and $\eta \triangleq \alpha$, we obtain the required result. ■

To show that the step size eventually becomes 1, we must establish a bound on $\theta(\cdot)$ near a solution.

Lemma 3.8: Suppose Assumptions 2.1-2.3 and 3.1-3.2 are satisfied and that $\hat{z} = (\hat{x},\hat{\mu})$ is a Danskin point for P. Then there exists $K > 0$ and a neighborhood O of $\hat{x}$ such that

$$\theta(x) \le -K\|h\|^2 \quad \forall\, x \in O, \quad \forall\, h \in H(x). \tag{3.62}$$

Proof: By Proposition 3.3 property (vi), we may assume that $H(\cdot)$ is a singleton. Hence we denote the search direction by $h(x)$. The search direction is continuous as a function of x and furthermore $h(\hat{x}) = 0$. By properties (vii) and (viii) of Proposition 3.3, the multipliers μ_+ defined by (3.20) - (3.23), for $x_+ = x + h(x)$ are continuous at $\hat{x}$. Suppose that the Lemma is false. Then there exists a sequence $x_k \to \hat{x}$, such that the following holds

$$\theta(x_k) > -\frac{1}{k}\|h_k\|^2 \quad \forall\, k \in \mathbb{N}, \tag{3.63}$$

where $h_k \triangleq h(x_k)$. Since by Proposition 3.4 $\hat{x}$ is an isolated local minimizer, we must have that

$h_k \neq 0$ for all k sufficiently large. Without loss of generality we may assume that this is so for all $k \in \mathbb{N}$. Let $u_k \triangleq h_k / \| h_k \|$, and let μ_k be the multiplier vector, defined by (3.20) - (3.23), corresponding to the solution of the search direction problem at x_k. It should be clear that the multipliers μ_k converge to the multiplier vector $\hat{\mu}$ at the solution $\hat{x}$. Clearly, we must have that $u_k \xrightarrow{K} \hat{u}$, for some infinite subsequence $K \subset \mathbb{N}$.

(i) Suppose that $\hat{u}$ satisfies $\langle \nabla f^j(\hat{x}), \hat{u} \rangle = 0$, for all $j \in A(\hat{x})$. Then because of Assumption 3.2,

$$\langle \hat{u}, l_{xx}(\hat{x}, \hat{\mu}) \hat{u} \rangle \geq m_0 > 0 . \tag{3.64}$$

It follows that there exists a k_0 such that for all $k \geq k_0$, $k \in K$,

$$\langle h_k, l_{xx}(x_k, \mu_k)\, h_k \rangle \geq \frac{m_0}{2} \| h_k \|^2 . \tag{3.65}$$

From (3.24) we obtain that

$$\theta(x_k) = \sum_{j=1}^{m} \mu_k^j (f^j(x_k) - \psi(x_k)) + \langle \sum_{j=1}^{m} \mu_k^j \nabla f^j(x_k), h_k \rangle + \frac{1}{2} \langle h_k, l_{xx}(x_k, \mu_k)\, h_k \rangle . \tag{3.66}$$

Making use of (3.20), (3.66) and the fact that $f^j(x) - \psi(x) \leq 0$, (3.66) simplifies out to

$$\theta(x_k) \leq -\frac{1}{2} \langle h_k, l_{xx}(x_k, \mu_k)\, h_k \rangle . \tag{3.67}$$

Hence for k sufficiently large we obtain a contradiction.

(ii) Suppose that $\langle \nabla f^l(\hat{x}), \hat{u} \rangle \neq 0$ for some $l \in A(\hat{x})$. Let L be the subspace spanned by the active gradients at the solution, i.e. $L \triangleq \mathrm{sp}\{ \nabla f^j(\hat{x}) \}_{j \in A(\hat{x})}$ and let $L^\perp$ denote the orthogonal complement of L. Note that $\hat{u} \notin L^\perp$. Let the projections of a vector x onto L and $L^\perp$ be denoted by the superscripts 1,2, respectively, i.e., $x = x^1 + x^2$ with $x^1 \in L$, $x^2 \in L^\perp$. From the continuity of the projection operators, it should be clear that there exists a k_1 such that $\forall\ k \geq k_1$, $k \in K$,

$$\| u_k^2 \| / \| u_k^1 \| \leq 2 \| \hat{u}^2 \| / \| \hat{u}^1 \| . \tag{3.68}$$

Combining this with the fact that $\| h \|^2 = \| h^1 \|^2 + \| h^2 \|^2$ implies that $\| h_k \| \leq L \| h_k^1 \|$ for some $L < \infty$ and for $k \in K$ sufficiently large.

Let $g(\cdot,\cdot)$ be as in (3.26), then for k sufficiently large $g(z_{k+1},z_k) = 0$ and $\mu^j_{k+1} > 0$ for all $j \in A(\hat{x})$. Hence, from (3.25) we conclude that

$$\bar{f}^{ij}(x_k) + \langle \nabla\bar{f}^{ij}(x_k),h_k\rangle + \frac{1}{2}\langle h_k,\bar{f}^{ij}_{xx}(x_k)h_k\rangle \geq 0 \quad \forall\ i \in A(\hat{x}),\ \ \forall\ j \in \mathbf{m}, \tag{3.69}$$

which may be written as

$$f^i(x_k) - f^j(x_k) - \langle \nabla f^j(x_k),h_k\rangle - \frac{1}{2}\langle h_k,f^j_{xx}(x_k)\,h_k\rangle \geq -\langle \nabla f^i(x_k),h_k\rangle - \frac{1}{2}\langle h_k,f^i_{xx}(x_k)\,h_k\rangle. \tag{3.70}$$

By maximizing first the left and then the right hand side of (3.70), with respect to $i \in A(\hat{x})$, and noting that $A(x_k) \subset A(\hat{x})$ for k sufficiently large, we obtain

$$\psi(x_k) - f^j(x_k) - \langle \nabla f^j(x_k),h_k\rangle - \frac{1}{2}\langle h_k,f^j_{xx}(x_k)\,h_k\rangle \geq \max_{i \in A(\hat{x})} -\langle \nabla f^i(x_k),h_k\rangle - \frac{1}{2}\langle h_k,f^i_{xx}(x_k)\,h_k\rangle \tag{3.71}$$

$$\geq \max_{i \in A(\hat{x})} -\langle \nabla f^i(x_k),h_k\rangle - L_2\|h_k\|^2, \tag{3.72}$$

for some $L_2 > 0$. To continue, let $C \triangleq \mathrm{co}\{\ \nabla f^j(\hat{x})\ \}_{j \in A(\hat{x})}$, then strict complementarity (Assumption 3.1) implies that $0 \in \mathrm{ri}\,C$ (where ri denotes the relative interior of a set). Hence there exists an $\eta > 0$ such that $\bar{B}(0,\eta) \cap L \subset C$. Hence, for any $h \in L$, we must have

$$\max_{i \in A(\hat{x})} \langle \nabla f^i(\hat{x}),h\rangle = \max_{v \in C}\langle v,h\rangle \geq \max_{v \in \bar{B}(0,\eta) \cap L} \langle v,h\rangle = \eta\,\|h\|., \tag{3.73}$$

Since the $\nabla f^j(\cdot)$ are continuous, it follows that there exists a neighborhood O_1 of $\hat{x}$ such that for all $x \in O_1$

$$\max_{i \in A(\hat{x})} \langle \nabla f^i(x),h\rangle \geq \frac{\eta}{2}\|h\|. \tag{3.74}$$

Substituting this into (3.72) yields

$$\psi(x_k) - f^j(x_k) - \langle \nabla f^j(x_k),h_k\rangle - \frac{1}{2}\langle h_k,f^j_{xx}(x_k)\,h_k\rangle \geq \frac{\eta}{2}\|h^1_k\| - L_2\|h_k\|^2$$

$$\geq \frac{\eta}{2L}\|h_k\| - L_2\|h_k\|^2 \quad \forall\ j \in \mathbf{m}. \tag{3.75}$$

Therefore we obtain

$$\theta(x_k) \le -\frac{\eta}{2L}\|h_k\| + L_2\|h_k\|^2, \tag{3.76}$$

which results in a contradiction for k sufficiently large. ■

The second main result of this section shows that in the vicinity of a Danskin point, the step size calculation returns a step length of unity.

Theorem 3.9: Suppose Assumptions 2.1-2.3 and 3.1-3.2 are satisfied and that $\hat{x}$ is a Danskin point for P. Then there exists a neighborhood O of $\hat{x}$ such that

$$\psi(x+h) - \psi(x) - \alpha\theta(x) \le 0 \quad \forall\, x \in O,\ \forall\, h \in H(x). \tag{3.77}$$

Proof: As in the previous Lemma, we note that by Proposition 3.3 property (vi), there is a $\rho > 0$ such that $H(x)$ is a singleton for all $x \in B(\hat{x},\rho)$. For $x \in B(\hat{x},\rho)$, we denote the search direction by $h(x)$, and observe that $h(\cdot)$ is continuous and $h(\hat{x}) = 0$. From Taylor's Theorem, we have that

$$\begin{aligned}
&\psi(x+h(x)) - \psi(x) - \alpha\theta(x) \\
&= \max_{\mu\in\Sigma}\left\{\sum_{j=1}^{m}\mu^j[f^j(x)-\psi(x)] + \langle\sum_{j=1}^{m}\mu^j\nabla f^j(x),h(x)\rangle + \int_0^1(1-s)\langle h(x), l_{xx}(x+s\,h(x),\mu)h(x)\rangle ds\right\} - \alpha\theta(x) \\
&\le \max_{\mu\in\Sigma}\left\{\sum_{j=1}^{m}\mu^j[f^j(x) - \psi(x)] + \langle\sum_{j=1}^{m}\mu^j\nabla f^j(x),h(x)\rangle + \frac{1}{2}\langle h(x), l_{xx}(x,\mu)\,h(x)\rangle\right\} \\
&\qquad + \max_{\mu\in\Sigma}\int_0^1(1-s)\langle h(x),(l_{xx}(x+s\,h(x),\mu) - l_{xx}(x,\mu))\,h(x)\rangle ds - \alpha\theta(x) \\
&\le (1-\alpha)\theta(x) + L\|h(x)\|^3 \\
&\le -(1-\alpha)K\|h(x)\|^2 + L\|h(x)\|^3.
\end{aligned} \tag{3.78}$$

Hence for x sufficiently close to $\hat{x}$, the step size will be 1. ■

4. THE ALGORITHM (Non-Convex Problems)

If any of the $f^j(\cdot)$ in (1.1) are non-convex, it is possible that the search direction problem (2.8) becomes non-convex. The effect of this is that the relations described in Lemma 2.2 are no longer true, and so we can no longer guarantee descent at a non-stationary point. In this section a simple modification is described which augments the Hessians $f^j_{xx}(\cdot)$ of non-convex functions to ensure a convex search direction problem.

Let $m_0 > 0$ and let $\gamma(\cdot)$ be defined by

$$\gamma(x) \triangleq \max\{\, 0, \frac{m_0}{2} - \lambda_{\min}(f^j_{xx}(x))\,\}, \tag{4.1}$$

where $j \in \underline{m}$ and $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of its argument. We define the new optimality function $\theta(\cdot)$ and the search direction map $H(\cdot)$ as follows

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, \sum_{j=1}^{m} \mu^j (f^j_{xx}(x) + \gamma(x) I)\, h \rangle \right\}, \tag{4.2}$$

$$H(x) \triangleq \operatorname*{argmin}_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^{m} \mu^j \tilde{f}^j(x) + \langle \sum_{j=1}^{m} \mu^j \nabla f^j(x), h \rangle + \frac{1}{2} \langle h, \sum_{j=1}^{m} \mu^j (f^j_{xx}(x) + \gamma(x) I)\, h \rangle \right\}. \tag{4.3}$$

The algorithm for solving the non-convex case of **P** can now be stated.

Algorithm 4.1.

Data: $x_0 \in \mathbb{R}^n, m_0 > 0,\ \alpha, \beta \in (0,1),\ S \triangleq \{\, 1, \beta, \beta^2, \ldots \,\}$.

Step 0: Set $i = 0$.

Step 1: Compute $\theta(x_i)$, and $h_i \in H(x_i)$.

Step 2: Compute the step length

$$\lambda_i \triangleq \max\{\, \lambda \in S \mid \psi(x_i + \lambda h_i) - \psi(x_i) \le \lambda \alpha \theta(x_i) \,\}. \tag{4.4}$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$.

Step 4: Replace i by $i+1$ and go to Step 1. ■

Global convergence of the modified algorithm follows almost immediately from the convex case. In the present case, only Assumption 2.1 is required, and the proofs of the following results are essentially those of results in Section 2 and are omitted.

Lemma 4.1: Suppose Assumption 2.1 is satisfied, then the function $\theta : \mathbb{R}^n \to \mathbb{R}$ is continuous, and the search direction map $H : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is u.s.c. ■

Lemma 4.2: Suppose Assumption 2.1 is satisfied, then

$$\theta(x) \geq d\psi(x;h) \quad \forall\ h \in H(x); \tag{4.5}$$

$$\theta(x) = 0 \ \text{ if and only if } \ 0 \in \partial\psi(x). \tag{4.6}$$

Theorem 4.3: Suppose Assumption 2.1 is satisfied. Then any accumulation point $\hat{x}$ of an infinite sequence $\{ x_i \}_{i=0}^{\infty}$, generated by Algorithm 4.1 satisfies $0 \in \partial\psi(\hat{x})$. ■

5. QUADRATIC CONVERGENCE (Non-Convex Problems)

To establish superlinear convergence, we require an additional assumption which ensures that $\psi(\cdot)$ is convex near a solution.

Assumption 5.1: For any Danskin point $(\hat{x},\hat{\mu})$ for problem **P**, (1.1), there exists a $m_0 > 0$ such that

$$\langle h, f^j_{xx}(\hat{x})h \rangle \geq m_0 \|h\|^2 , \ \forall\ h \in \mathbb{R}^n, \forall\ j \in A(\hat{x}), \tag{5.1}$$

where $A(\cdot)$ is the active function set defined by (2.16). ■

The results of this section are almost identical to those of Section 3, and in most cases the proofs are omitted.

Proposition 5.1: Let $(\hat{x},\hat{\mu})$ be a Danskin point for problem **P** at which Assumptions 2.1, 3.1 and 5.1 hold. Then $\hat{x}$ is a local minimizer for problem **P**.

■

To continue, we note that a necessary and sufficient condition for a vector h to be optimal for the search direction problem (4.2) is that there exists a $z_+ = (x_+,\mu_+)$ satisfying $x_+ - x = h$ and

$$\sum_{j=1}^{m} \mu^j_+ \{ \nabla f^j(x) + (f^j_{xx}(x) + \gamma^j(x)I)(x_+ - x) \} = 0, \tag{5.2}$$

$$\mu_+ \geq 0, \tag{5.3}$$

$$\sum_{j=1}^{m} \mu^j_+ = 1, \tag{5.4}$$

$$\mu^j_+ \{ \theta(x) - [f^j(x) - \psi(x) + \langle \nabla f^j(x), x_+ - x \rangle + \frac{1}{2} \langle (x_+ - x), (f^j_{xx}(x) + \gamma^j(x)I)(x_+ - x) \rangle] \} = 0, \tag{5.5}$$

for all $j \in \underline{m}$. Condition (5.5) is equivalent to

$$\mu^j_+\{\ [f^k(x) - \psi(x) + \langle \nabla f^k(x), x_+ - x\rangle + \frac{1}{2}\langle (x_+ - x), (f^j_{xx}(x) + \gamma^j(x) I)(x_+ - x)\rangle] -$$

$$[f^j(x) - \psi(x) + \langle \nabla f^j(x), x_+ - x\rangle + \frac{1}{2}\langle (x_+ - x), (f^j_{xx}(x) + \gamma^j(x) I)(x_+ - x)\rangle]\ \} \le 0\ ,\ \forall\ j,k \in \underline{m}\,. \tag{5.6}$$

Suppose that $\hat{z}$ is a Danskin point for problem **P** at which Assumption 2.1 holds. Without loss of generality, we may assume that $A(\hat{x}) = \{\ 1,\ldots,r\ \}$. Let the function $g : \mathbb{R}^{n+m}\times\mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ be defined as in (3.26) except that each occurrence of $f^j_{xx}(x)$ is replaced by $f^j_{xx}(x) + \gamma^j(x)I$. The following result is essentially the same as Proposition 3.3, with the addition of property (viii) whose proof is obvious under Assumption 5.1.

Proposition 5.2: Suppose Assumptions 2.1, 3.1 and 5.1 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Let $g(\cdot,\cdot)$ be as above. Then

(i) $g(\hat{z},\hat{z}) = 0$;

(ii) $g_{z_1}(\cdot,\cdot)$ exists and is continuous in some neighborhood of $(\hat{z},\hat{z})$ and is invertible at $(\hat{z},\hat{z})$;

(iii) There exist neighborhoods M and N of $\hat{z}$ and a function $Z(\cdot)$ satisfying (3.16) - (3.19);

(iv) $Z((x,\mu)) = Z((x,\overline{\mu}))$ for all $\mu,\overline{\mu} \in \mathbb{R}^m$ such that $(x,\mu),(x,\overline{\mu}) \in M$;

(v) There exists a neighborhood $U \subset M$ containing $\hat{z}$ such that for all $z \in U$, $Z(z)$ is the unique solution to the search direction problem (4.2);

(vi) There exists a neighborhood G of $\hat{x}$ such that $\forall\ x \in G$, the search direction map $H(x)$ is a singleton. In addition $H(\hat{x}) = \{\ 0\ \}$ and $H(\cdot)$ is continuous at $\hat{x}$;

(vii) The multiplier μ_+, defined by $(x_+,\mu_+) = Z((x,\mu))$, is continuous in x at $\hat{x}$;

(viii) There exists neighborhood O of $\hat{x}$ such that for all $x \in O$, $\gamma^j(x) = 0$ for all $j \in A(\hat{x})$. ■

As in Section 3, Proposition 5.1 can be strengthened to give the following result.

Proposition 5.3: Suppose that Assumptions 2.1, 3.1 and 5.1 hold and that $\hat{z} = (\hat{x},\hat{\mu})$ is a Danskin point for problem **P**. Then $\hat{x}$ is a local minimizer for problem **P**, and there exists a neighborhood W of $\hat{z}$ such that $\hat{z}$ is the unique Danskin point of **P** in W.

Proof: Since property (viii) of Proposition 5.2 implies that

$$\sum_{j=1}^{m} \hat{\mu}^j (f^j_{xx}(\hat{x}) + \gamma^j(\hat{x}) I) = \sum_{j=1}^{m} \hat{\mu}^j f^j_{xx}(\hat{x}), \tag{5.7}$$

the proof of this Proposition is the same as the one for Proposition 3.4. ∎

A quantitative bound on the variation of $Z(\cdot)$ is provided, as before, by Proposition 3.5. Lemma 3.6 has to be modified slightly to account for the extra terms in $g(\cdot,\cdot)$.

Lemma 5.4: Suppose that Assumptions 2.1, 3.1 and 5.1 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Let $g(\cdot,\cdot)$ be as above, and let $Z(\cdot)$ be as in Proposition 5.2. Then there exists a neighborhood $K \subset \mathbb{R}^{n+m}$ of $\hat{z}$ and a $L > 0$ such that

$$|g(Z(z),Z(z)) - g(Z(z),z)| \le L\|Z(z) - z\|^2, \ \forall\, z \in K. \tag{5.8}$$

Proof: The proof is essentially the same as that of Lemma 3.6, provided (5.7) is taken into account. ∎

The proof of local quadratic convergence is almost identical to that of the convex case.

Theorem 5.5: Suppose that Assumptions 2.1, 3.1 and 5.1 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Let $g(\cdot,\cdot)$ be as above, and let $Z(\cdot)$ be as in Proposition 5.2. Then there exists a neighborhood V of $\hat{z}$ such that if $z_0 \in V$, then the sequence $\{ z_i \}_{i=0}^{\infty}$ defined by $z_{i+1} \triangleq Z(z_i)$ is well defined and converges to $\hat{z}$ with R-order at least 2 (i.e. quadratically). Specifically, there exist constants $\eta \in (0,1)$, $\bar{K} > 0$ such that

$$\|\hat{z} - z_i\| \le \bar{K}\eta^{2^i}, \ \forall\, i \ge 0. \tag{5.9}$$

∎

To show that the step size eventually becomes 1, we must establish a rate bound on $\theta(\cdot)$ near a local solution.

Lemma 5.6: Suppose that Assumptions 2.1, 3.1 and 5.1 are satisfied and that $\hat{z}$ is a Danskin point for **P**. Then there exists $K > 0$ and a neighborhood O of $\hat{x}$ such that

$$\theta(x) \le -K\|h\|^2 \quad \forall\, x \in O, \quad \forall\, h \in H(x). \tag{5.10}$$

Proof: The proof is similar to that of case (i) in Lemma 3.8, and is based on the fact that by construction

$$\langle h_k, (f^j_{xx}(x_k) + \gamma^j(x_k)I)h_k \rangle \geq \frac{m_0}{2}\|h_k\|^2, \tag{5.11}$$

where x_k, h_k are defined as in the proof of Lemma 3.8. ■

The second main result of this section shows that in the vicinity of a Danskin point, the step size calculation returns a step length of unity.

Theorem 5.7: Suppose that Assumptions 2.1, 3.1 and 5.1 are satisfied and that $\hat{z}$ is a Danskin point for P. Then there exists a neighborhood O of $\hat{x}$ such that

$$\psi(x+h) - \psi(x) - \alpha\theta(x) \leq 0 \quad \forall\, x \in O,\ \forall\, h \in H(x). \tag{5.12}$$

Proof: As in Lemma 5.6, we note that by Proposition 5.2 property (vi), $H(x)$ is a singleton near $\hat{x}$. In this case, we denote the search direction by $h(x)$ and observe that $h(\cdot)$ is continuous and $h(\hat{x}) = 0$. In addition, in a neighborhood of $\hat{x}$ $\psi(\cdot)$ may be written as

$$\psi(x) = \max_{j \in A(\hat{x})} f^j(x)$$

In view of Taylor's Theorem and Proposition 5.2 (viii), because $h(\hat{x}) = 0$, for x sufficiently close to $\hat{x}$, we may write

$$\begin{aligned}
&\psi(x + h(x)) - \psi(x) - \alpha\theta(x) \\
&= \max_{j \in A(\hat{x})} \left\{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \int_0^1 (1-s)\langle h(x), f^j_{xx}(x + s\,h(x))h(x) \rangle ds \right\} - \alpha\theta(x) \\
&\leq \max_{j \in A(\hat{x})} \left\{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \frac{1}{2}\langle h(x), f^j_{xx}(x)\, h(x) \rangle \right\} \\
&\qquad\qquad + \max_{j \in A(\hat{x})} \int_0^1 (1-s)\langle h(x), (f^j_{xx}(x+s\,h(x)) - f^j_{xx}(x))\, h(x) \rangle ds - \alpha\theta(x) \\
&\leq (1-\alpha)\theta(x) + L\|h(x)\|^3 \\
&\leq -(1-\alpha)K\|h(x)\|^2 + L\|h(x)\|^3.
\end{aligned} \tag{5.13}$$

Hence for x sufficiently close to $\hat{x}$, the step size will be 1. ■

6. NUMERICAL EXPERIENCE (Convex problems)

Three examples which illustrate the performance of Algorithm 2.1 are considered. For the purpose of comparison, the Pshenichnyi linearization method is also applied to these problems. In each case the cost function value is indicated at each iteration. All of the problems considered satisfy Assumptions 2.1-2.3, 3.1 and 3.2. The Armijo step size parameters are taken to be $\alpha = 0.1$ and $\beta = 0.5$ in all cases. In figures 1-3, the solid line represents Algorithm 2.1, and the dashed line represents the Pshenichnyi linearization method.

Test Problem 6.1: (figure 6.1)

$$\psi(x) \triangleq \max\{ e^{\frac{x_1^2}{1000}+(x_2-1)^2}, e^{\frac{x_1^2}{1000}+(x_2+1)^2} \}. \tag{6.1}$$

The initial point is $x_0 \triangleq (50.0,0.05)^T$.

Test Problem 6.2: (figure 6.2) Define $f : \mathbb{R}^{10} \to \mathbb{R}$ as

$$f(x) \triangleq e^{(0.0001\, x_1)^2+x_2^2+x_3^2+(2x_4)^2+x_5^2+\ldots+x_{10}^2}, \tag{6.2}$$

and let $e_1 \triangleq (1,0,\ldots,0)^T$. Define the cost function as

$$\psi(x) \triangleq \max\{ f(x+2e_1), f(x-2e_1) \}. \tag{6.3}$$

The initial point is $x_0 \triangleq (100.0,0.1,0.1,\ldots,0.1)^T$.

Test Problem 6.3: (figure 6.3) Define $f^j : \mathbb{R}^{50} \to \mathbb{R}$ for all $j \in \underline{50}$ as

$$f^j(x) \triangleq \sum_{i=0}^{50} \alpha_i^j \, e^{(x_i-t_i^j)^2}, \tag{6.4}$$

where $\alpha_i^j \in [0,150]$ and $t_i^j \in [0,1.5]$. The cost function is defined as

$$\psi(x) \triangleq \max_{j \in \underline{50}} f^j(x) \tag{6.5}$$

The initial point is $x_0 \triangleq (1.0,1.0,\ldots,1.0)^T$.

APPENDIX

In the previous algorithms, quadratic convergence was obtained by assuming that $\psi(\cdot)$ was strongly convex in a neighborhood of a local solution. By relaxing this assumption slightly, superlinear convergence can still be obtained, at the expense of a more complex algorithm. This algorithm requires that the reduced active Hessians be sufficiently positive definite at local solutions.

Assumption A.1: At any Danskin point $(\hat{x},\hat{\mu})$ of problem **P**, there exists a $m_0 > 0$ such that

$$\langle h, f^j_{xx}(\hat{x})h\rangle \geq m_0\|h\|^2, \tag{A.1}$$

for all $j \in A(\hat{x})$ and for all h satisfying

$$\langle \nabla f^j(\hat{x}), h\rangle = 0\,, \qquad \forall\, j \in A(\hat{x}), \tag{A.2}$$

where $A(\cdot)$ is the active function set defined by (2.16). ■

Let $x \in \mathbb{R}^n$, $\delta \in \mathbb{R}^m$, $\gamma \in \mathbb{R}$. For $i \in \mathbb{N}$, we shall denote by I_i subsets of $\underline{m}$ and by G_i matrices whose range is $\mathbb{R}^n$. We define the quantities $\theta(\cdot,\cdot,\cdot,\cdot)$, $H(\cdot,\cdot,\cdot,\cdot)$, $I(\cdot,\cdot,\cdot,\cdot,\cdot)$ and $G(\cdot,\cdot)$ as follows:

$$\theta(x,\delta,\gamma,G) \triangleq \min_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^m \mu^j \bar{f}^j(x) + \langle \sum_{j=1}^m \mu^j \nabla f^j(x), h\rangle + \frac{1}{2}\langle h, \sum_{j=1}^m \mu^j (f^j_{xx}(x) + \delta^j I + \gamma G G^T)h\rangle \right\}, \tag{A.3}$$

$$H(x,\delta,\gamma,G) \triangleq \operatorname*{argmin}_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^m \mu^j \bar{f}^j(x) + \langle \sum_{j=1}^m \mu^j \nabla f^j(x), h\rangle + \frac{1}{2}\langle h, \sum_{j=1}^m \mu^j (f^j_{xx}(x) + \delta^j I + \gamma G G^T)h\rangle \right\} \tag{A.4}$$

$$I(x,\delta,\gamma,G,h) \triangleq \Bigg\{ j \in \underline{m} \mid \bar{f}^j(x) + \langle \nabla f^j(x), h\rangle + \frac{1}{2}\langle h, (f^j_{xx}(x) + \delta^j I + \gamma G G^T)h\rangle$$

$$= \max_{\mu \in \Sigma} \left\{ \sum_{j=1}^m \mu^j \bar{f}^j(x) + \langle \sum_{j=1}^m \mu^j \nabla f^j(x), h\rangle + \frac{1}{2}\langle h, \sum_{j=1}^m \mu^j (f^j_{xx}(x) + \delta^j I + \gamma G G^T)h\rangle \right\} \Bigg\}. \tag{A.5}$$

Next we turn to the matrix of $G(x,I)$. Let the cardinality of I be r and let $\{\, i_j \,\}_{j=1}^r = I$ be an ordering of the indices in I such that $i_j < i_{j+1}$. Then

$$G(x,I) \triangleq \left[\nabla f^{i_1}(x) \cdots \nabla f^{i_{r-1}}(x)\right]. \tag{A.6}$$

The algorithm for solving **P** may now be stated.

Algorithm A.1.

Data: $x_0 \in \mathbf{R}^n$, $m_0 > 0$, $\alpha, \beta \in (0,1)$, $S \triangleq \{ 1, \beta, \beta^2, \ldots \}$, $I_0 \subset \underline{m}$.

Step 0: Set i = 0.

Step 1: Compute $G_i \triangleq G(x_i, I_i)$ and N_i, an orthonormal matrix whose columns span the orthogonal complement of the range of G_i.

Step 2: Compute $\delta_i \in \mathbf{R}^m$ according to

$$\delta_i^j \triangleq \max\{ 0, \frac{m_0}{2} - \lambda_{\min}(N_i^T f_{xx}^j(x_i) N_i) \} , \; j \in \underline{m}. \tag{A.7}$$

Step 3: Compute γ_i as follows. Let

$$\hat{\gamma}_i \triangleq \min\{ \gamma \mid \lambda_{\min}(f_{xx}^j(x_i) + \delta^j I + \gamma G_i G_i^T) \geq \frac{m_0}{4} \quad \forall j \in \underline{m} \}. \tag{A.8}$$

Then

$$\gamma_i \triangleq \begin{cases} \gamma_{i-1} & \text{if } \gamma_{i-1} \geq \hat{\gamma}_i \\ \hat{\gamma}_i + 1 & \text{if } \gamma_{i-1} < \hat{\gamma}_i \end{cases} \tag{A.9}$$

Step 4: Compute $\theta(x_i, \delta_i, \gamma_i, G_i)$ and $h_i \in H(x_i, \delta_i, \gamma_i, G_i)$.

Step 5: Compute $I_{i+1} \triangleq I(x_i, \delta_i, \gamma_i, G_i, h_i)$.

Step 6: Compute the step size

$$\lambda_i \triangleq \max\{ \lambda \in S \mid \psi(x_i + \lambda h_i) - \psi(x_i) \leq \lambda\alpha\theta(x_i) \}. \tag{A.10}$$

Step 7: Set $x_{i+1} = x_i + \lambda_i h_i$.

Step 8: Replace i by i+1 and go to Step 1. ■

Theorem A.1: Suppose Assumptions 2.1, 3.1 and A.1 are satisfied and that $\{ x_i \}_{i=0}^{\infty}$ and $\{ \gamma_i \}_{i=0}^{\infty}$ are bounded sequences generated by algorithm A.1. Then $\{ x_i \}_{i=0}^{\infty}$ converges to a Danskin point $\hat{x}$ with R-order at least $\sqrt{2}$. ■

The proof of the above theorem is similar in nature to that given for the previous algorithms, but considerably more laborious and full of technical details. Since Algorithm A.1 does not represent a very large advance over Algorithm 4.1, the proof is omitted to save space.

REFERENCES

[1] Charalambous, C. and Conn, A. R., "An efficient method to solve the minimax problem directly," *SIAM J. Numer. Anal.* 15(1)(February 1978).

[2] Daugavet, V. A. and Malozemov, V. N., "Quadratic rate of convergence of a linearization method for solving discrete minimax problems," *U.S.S.R. Comput. Maths. Math. Phys.* 21(4) pp. 19-28 (1981).

[3] Fletcher, R., "A model algorithm for composite nondifferentiable optimization problems," *Mathematical Programming Study* 17 pp. 67-76 (1982).

[4] Hald, J. and Madsen, K., "Combined LP and Quasi-Newton methods for minimax optimization," *Mathematical Programming* 20 pp. 49-62 (1981).

[5] Kantorovich, L. V. and Akilov, G. P., *Functional Analysis, (second edition)*, Pergamon Press, Oxford, England (1982).

[6] Madsen, K. and Schjaer-Jacobsen, H., "Linearly constrained minimax optimization," *Mathematical Programming* 14 pp. 208-223 (1978).

[7] Maratos, N., *Exact penalty function algorithms for finite dimensional and control optimisation problems*, Ph.D. Thesis, University of London, London (1978).

[8] Mayne, D. Q., Polak, E., and Sangiovanni-Vincentelli, A., "Computer-Aided Design via Optimization: A Review," *Automatica* 18(2) pp. 147-154 (1982).

[9] Mayne, D. Q. and Polak, E., "A Superlinearly Convergent Algorithm for Constrained Optimization Problems," *Mathematical Programming Study* 16 pp. 45-61 (1982).

[10] Pironneau, O. and Polak, E., "On the Rate of Convergence of Certain Methods of Centers," *Mathematical Programming* 2(2) pp. 230-258 (1972).

[11] Polak, E., Trahan, R., and Mayne, D. Q., "Combined phase I - phase II methods of feasible directions," *Mathematical Programming* 17(1) pp. 32-61 (1979).

[12] Polak, E., Mayne, D. Q., and Stimler, D. M., "Control System Design via Semi-Infinite Optimization: A Review," *Proceedings of the IEEE* 72(12) pp. 1777-1794 (1984).

[13] Polak, E., "On the Mathematical Foundations of Nondifferentiable Optimization in Engineering Design," *SIAM Review* 29(1) pp. 21-89 (March 1987).

[14] Polyak, B. T., *Introduction to Optimization, (in Russian)*, , Nauka, Moscow (1983).

[15] Pshenichnyi, B. N. and Danilin, Yu. M., *Numerical methods in extremal problems (Chislennye metody v ekstremal'nykh zadachakh)*, , Nauka, Moscow (1975).

[16] Robinson, S. M., "Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms," *Mathematical Programming* 7 pp. 1-16 (1974).

[17] Salcudean, S. E., *Algorithms for Optimal Design of Feedback Compensators*, Ph.D. Thesis, Department of EECS, University of California, Berkeley, California (1986).

[18] Womersley, R. S. and Fletcher, R., "An Algorithm for Composite Nonsmooth Optimization Problems," *Journal of optimization theory and applications* 48(3)(March 1986).

[19] Wuu, T. L. S., *DELIGHT.MIMO: An Interactive System for Optimization-Based Multivariable Control System Design.*, Ph.D. Thesis, Department of EECS, University of California, Berkeley, California (1986).

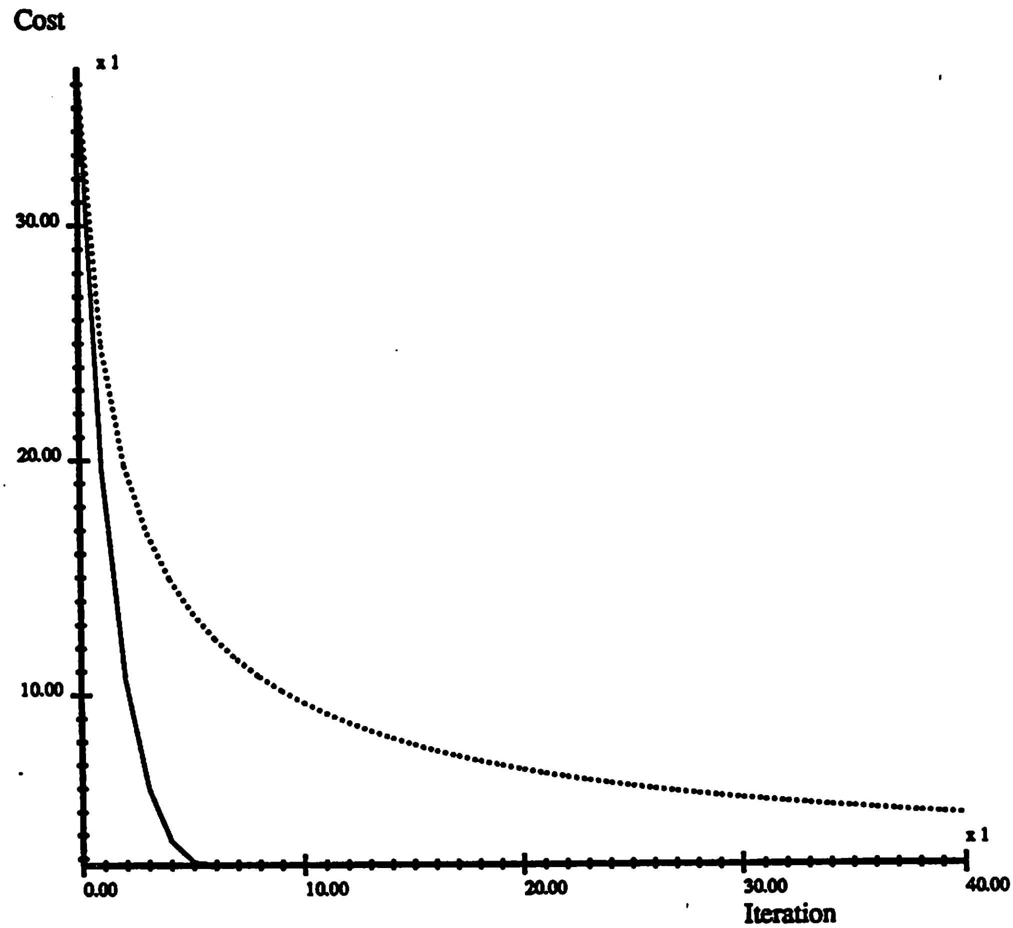

Figure 6.1

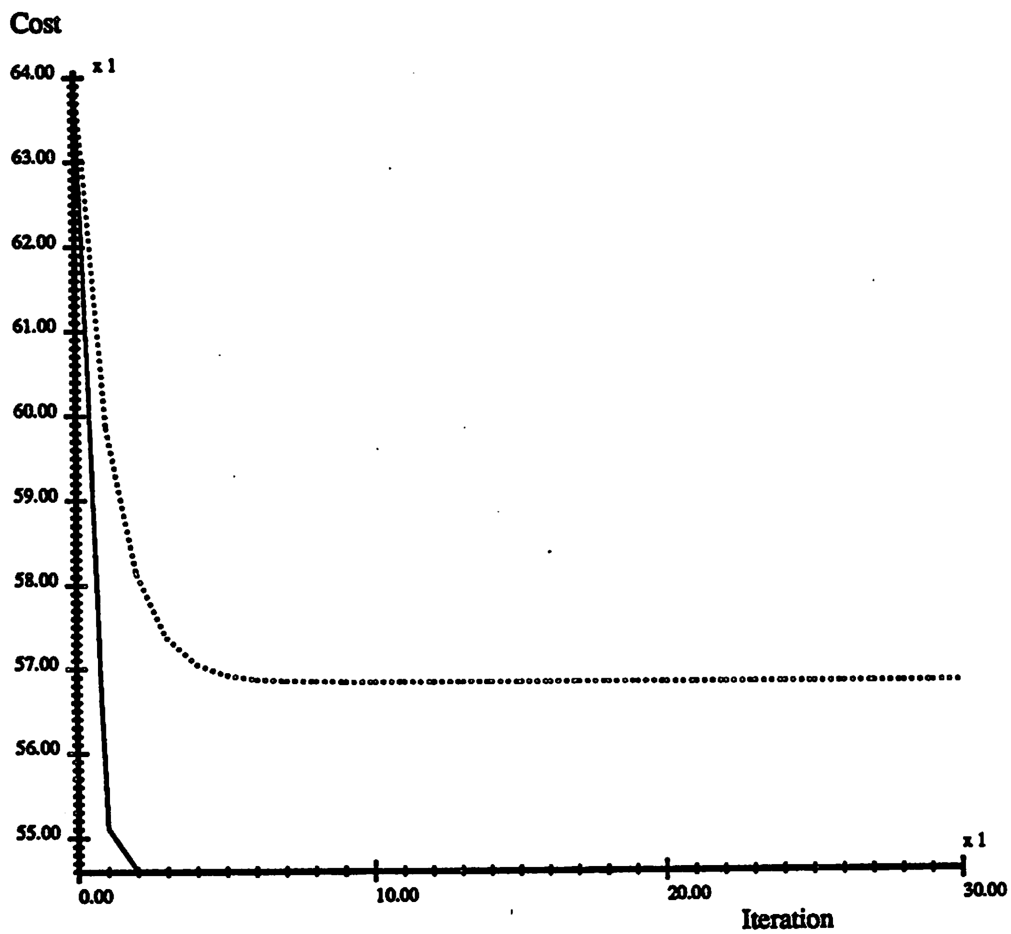
Cost
x 1
64.00
63.00
62.00
61.00
60.00
59.00
58.00
57.00
56.00
55.00
x 1
0.00
10.00
20.00
30.00
Iteration

Figure 6.2

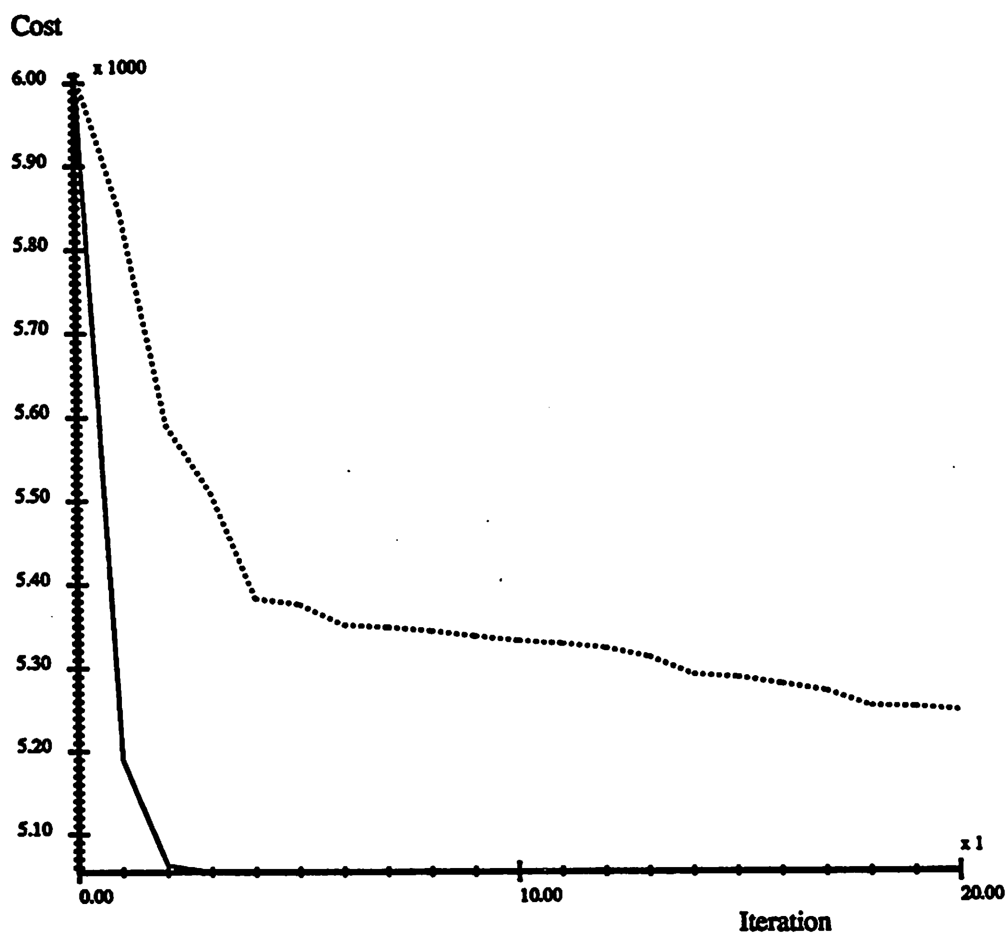
Cost
x 1000
6.00
5.90
5.80
5.70
5.60
5.50
5.40
5.30
5.20
5.10
x 1
0.00
10.00
20.00
Iteration

Figure 6.3