INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

- 1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.**
- 2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.**
- 3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.**
- 4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.**
- 5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.**

Xerox University Microfilms

300 North Zeeb Road
Ann Arbor, Michigan 48106

76-21,667

BYRD, Richard Henry, 1949-
LOCAL CONVERGENCE OF THE DIAGONALIZED
METHOD OF MULTIPLIERS.

Rice University, Ph.D., 1976
Mathematics

Xerox University Microfilms, Ann Arbor, Michigan 48106

RICE UNIVERSITY

LOCAL CONVERGENCE OF THE
DIAGONALIZED METHOD OF MULTIPLIERS

by

Richard Henry Byrd

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

A handwritten signature in cursive script, appearing to read "Richard Taped", written over a horizontal line.

Thesis Director's Signature

Houston, Texas
April, 1976

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor Richard A. Tapia for his invaluable guidance and encouragement. I would also like to thank Jorge Nocedal for his help in experimental programming and Velva Power for her great patience in typing this thesis.

TABLE OF CONTENTS

I. Introduction	Page 1
II. The Method of Multipliers with Bounded Penalty Constants	
1. The Standard Method	Page 6
2. The Diagonalized Method	Page 8
III. Convergence with Unbounded Penalty Constants	
1. Large Penalty Constants and Updates	Page 16
2. Diagonalization with Large Penalty Constants	Page 22
3. Conclusion	Page 28

1. Introduction

In this study we will consider algorithms for the solution of the following problem.

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g(x) = 0 \end{aligned} \tag{1.1}$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ have three continuous derivatives.

We will denote the Fréchet derivative of a function φ at x by $\varphi'(x)$. If φ is considered as a function of two vector variables x and y we will denote the partial derivative with respect to x by $\partial_x \varphi(x, y)$, and $[\partial_x \varphi(x, y)]^T$ by $\nabla_x \varphi(x, y)$. Likewise, second derivatives will be denoted by $H\varphi(x, y)$ and $H_x \varphi(x, y)$.

A large class of algorithms for solving problem (1.1) involve the augmented Lagrangian.

$$L(x, \lambda, c) = f(x) + g(x)^T \lambda + \frac{1}{2} c g(x)^T g(x)$$

where $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$ and $c \in \mathbb{R}^+$. If $x^* \in \mathbb{R}^n$ is such that $\nabla g(x^*)$ is of rank m , then a necessary condition for x^* to be a solution of problem (1.1) is that there exist λ^* such that

$$\nabla_x L(x^*, \lambda^*, c) = 0 \tag{1.2}$$

$$g(x^*) = 0$$

Such a λ^* must be unique. Throughout this study we will assume that $\nabla g(x^*)$ is of rank m and that $H_x L(x^*, \lambda^*, c)$ is nonsingular. It may be noted that the constant c does not affect condition (1.2).

Hestenes in [1] and Powell in [2] independently suggest approximating (x^*, λ^*) by the sequence $\{(x^k, \lambda^k)\}$ where λ^0 and c^0 are given, x^k is a minimizer of $L(x, \lambda^k, c^k)$, and $\lambda^{k+1} = \lambda^k + c^k g(x^k)$. It can be shown, Hestenes [1], that in a neighborhood of λ^* it is possible to choose c^k sufficiently large that the minimizer x^k exists. This method is commonly referred to as the method of multipliers.

Other ways of picking λ^k have been proposed, all of the form $\lambda^{k+1} = U(x^k, \lambda^k, c^k)$. The function U is referred to as an update function. The selection of λ in the method of multipliers may be looked at in two ways, both based on the following argument.

Since $\nabla L(x^*, \lambda^*, c) = 0$ and $H_L(x^*, \lambda^*, c)$ is nonsingular, then by the implicit function theorem there is a function $x: \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that $x(\lambda) = x^*$, $\nabla L(x(\lambda), \lambda, c) = 0$ for all λ in some neighborhood of λ^* , and for each λ , $x(\lambda)$ is the unique zero. Then we may regard the Lagrange multiplier update as an iterative formula for solving $g(x(\lambda)) = 0$. Buys [3] has shown that the point (x^*, λ^*) may be regarded as a saddle point, i.e., that λ^* is a local maximum of $L(x(\lambda), \lambda, c)$ for c sufficiently large. Then since $L'(x(\lambda), \lambda, c) = g(x(\lambda))^T$ the update proposed by Hestenes and Powell

$$U(x, \lambda, c) = \lambda + cg(x) \quad (1.3)$$

may be thought of as a version of the gradient method. Buys also suggests the use of a Newton step on the equation $L'(x(\lambda), \lambda, c) = g(x(\lambda))^T = 0$. Since, by the implicit function theorem, $x'(\lambda) = -H_x L(x(\lambda), \lambda, c)^{-1} \nabla_x g(x(\lambda))$ this approach yields the Buys update

$$U(x, \lambda, c) = \lambda + [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} g(x) \quad (1.4)$$

Clearly, requiring x^k to be an exact or even an approximate minimizer of $L(x, \lambda^k, c^k)$ is in general either theoretically impossible or in the latter

case a large amount of work. Buys and also Bertsekas [4] analyze the convergence of $\{(x^k, \lambda^k)\}$ when x^k is required to be within some distance of the exact minimizer. Still it seems inefficient to require an accurate calculation of x^k when λ^k is determined by a single iteration. Miele et al [5] and Tapia [6] argue that it is more efficient to update the multiplier after just one step of the minimization algorithm for x^k ; Tapia refers to this modification as a diagonalization of the method of multipliers.

When an exact minimization is not used to find x^k it is no longer true that $\nabla_x L(x^k, \lambda^k, c^k) = 0$. In this case one may propose a number of updates which reduce to the Hestenes-Powell update or the Buys update in the case of exact minimization.

In particular, if in (1.3) we substitute $-\nabla g(x)_L^{-1}[\nabla f(x) + c\nabla g(x)g(x)]$, for λ , which is valid if $\nabla_x L(x, \lambda, c) = 0$, we get the update

$$U(x, \lambda, c) = -\nabla g(x)_L^{-1} \nabla f(x) . \quad (1.5)$$

Here $\nabla g(x)_L^{-1}$ refers to the left inverse of the matrix. If the particular left inverse $\nabla g(x)_L^{-1} = [\nabla g(x)^T \nabla g(x)]^{-1} \nabla g(x)^T$ is used we get the projection update

$$U(x, \lambda, c) = -[\nabla g(x)^T \nabla g(x)]^{-1} \nabla g(x)^T \nabla f(x) \quad (1.6)$$

which has been considered by Buys and implemented by Miele and Levy in [5] and Fletcher in [7].

If we use the same substitution in the Buys update (1.4) we obtain

$$U(x, \lambda, c) = -\nabla g(x)_L^{-1} [\nabla f(x) + c\nabla g(x)g(x)] + [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} g(x) . \quad (1.7)$$

In the special case $\nabla g(x)_L^{-1} = [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} \nabla g(x)^T H_x L(x, \lambda, c)^{-1}$ we have the update proposed by Tapia.

$$U(x, \lambda, c) = [\nabla g(x)^T H_L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} [g(x) - \nabla g(x)^T H_x L(x, \lambda, c)^{-1} (\nabla f(x) + c \nabla g(x)^T g(x))]. \quad (1.8)$$

In formulas (1.5) - (1.8) λ only enters through $\nabla_x L$ and $H_x L$ and it can be shown that $\partial_\lambda U(x^*, \lambda^*, c) = 0$. Such update formulas are referred to by Tapia as local approximation formulas.

Under the assumption that the penalty constants c^k remain bounded for all k , the convergence properties of the method of multipliers using Newton's method and updates (1.3), (1.4), (1.6), and (1.8) were analyzed by Tapia. For the multiplier method with exact minimization formulas (1.4) and (1.8) are equivalent and only they give quadratic convergence. With the diagonalized multiplier method this equivalence no longer holds and only (1.8) gives quadratic convergence.

If the penalty constants c^k increase to infinity better convergence has been observed. Bertsekas [4] proves that this strategy, used with exact minimization yields superlinear convergence, and it can be seen that for large C the Hestenes-Powell update approaches the Buys update. In computational experiments Miele et al [5] observe better than R-linear convergence in using (1.6) in a diagonalized method.

It is natural to ask what class of updates yields quadratic or super-linear convergence, and how one may preserve this convergence under diagonalization. One may generalize the concept of a diagonalization to that of a j -step diagonalization, one where one takes j steps of a particular minimization algorithm after each Lagrange multiplier update; this allows one to achieve a balance between the iterations in x and in λ . In this study we consider the convergence properties of j -step diagonalizations of the method of multipliers with various updates in conjunction with Newton's method or Newton-like methods.

In section II we consider the case of a bounded sequence of penalty constants and we obtain a characterization of those updates which give quadratic convergence if we minimize exactly. We then show that this class of updates, which includes the Buys update, will give quadratic convergence for a two-step diagonalization with respect to Newton's method. Conditions are also given for quadratic convergence of a one-step diagonalization, which conditions are essentially only satisfied by the Tapia update. A general family of updates which includes all proposed updates and essentially all likely updates is derived and analyzed.

In section III the effect of unbounded increase of the penalty constant is considered. It is shown that a wide class of updates yields quadratic convergence for a two-step diagonalization under these conditions and that such updates will give convergence rates between one and two depending on the rate of increase of the penalty constant. Although these updates are in general simpler, it is shown that a large penalty constant will decrease the region of guaranteed convergence for Newton's method.

II. The Method of Multipliers with Bounded Penalty Constants

1. The standard method. Consider the method of multipliers implemented with the update $\lambda^{k+1} = U(x^k, \lambda^k, c^k)$ such that for some $\mu > 0$, $c^k \leq \mu$ for all k . We will consider first the standard method of multipliers (exact minimization), and determine what conditions on the update characterize quadratic convergence.

As stated in the introduction the requirement of exact minimization defines by the implicit function theorem a function $x: \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that $\nabla_x L(x(\lambda), \lambda, c) = 0$ for all λ in a neighborhood of the solution λ^* and such that $x(\lambda^*) = x^*$. We may regard the update as an iterative method for solving the equation $g(x(\lambda)) = 0$.

An iterative process which generates a sequence $\{y^k\}$ in $\mathbb{R}^n$ for any starting value is said to converge locally quadratically to y^* if there exists constants $\gamma, r > 0$ such that $\|y^0 - y^*\| < r$ implies $\|y^{k+1} - y^*\| \leq \|y_k - y^*\|^2$ for all k .

This is a slightly stronger statement than saying y^* is a point of attraction at the iterative process and the iteration is Q-quadratically convergent to y^* . However, all updates in the literature which are Q-quadratically convergent are also locally quadratically convergent. It is known that local quadratic convergence is characterized as follows.

Theorem 2.1. If S maps $\mathbb{R}^n$ into itself, satisfies $S(x^*) = x^*$, and has a continuous Fréchet derivative which is Lipschitz at x^* , then the iterative process $y^{k+1} = S(y^k)$ is locally quadratically convergent to y^* if and only if $S'(y^*) = 0$.

The sufficiency part of this theorem is proved in Ortega and Rheinboldt [9] and the necessity is easily shown. We will now give a characterization of

updates formulas yielding quadratic convergence.

Lemma 2.1. Consider the iteration $\lambda^{k+1} = U(x^k, \lambda^k, c^k)$ where x^k is the solution of $\nabla L(x, \lambda^k, c^k) = 0$ and c^k is chosen by $c^k = \tilde{c}(x^{k-1})$; where $\tilde{c}$ is continuous at x^* and U is a C^2 function from R^{n+m+1} to R^n .

This iterative process is locally quadratically convergent if and only if

$$\partial_x U(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c^*) = \frac{\partial}{\partial \lambda} U(x^*, \lambda^*, c^*) \quad (2.1)$$

Proof: The iterative procedure $\lambda^{k+1} = U(x(\lambda^k), \lambda^k, \tilde{c}(x(\lambda^k)))$ is a standard functional iteration which is locally quadratically convergent if and only if $U'(x(\lambda^*), \lambda^*, c(\lambda^*)) = 0$. Note that $\partial_c U(x^*, \lambda^*, c) = 0$ for all c . Now by the implicit function theorem

$$x'(\lambda^*) = -\partial_x \nabla_x L(x^*, \lambda^*, c^*)^{-1} \partial_\lambda \nabla L(x^*, \lambda^*, c^*)$$

$$-H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*) .$$

Therefore,

$$\begin{aligned} U'(x(\lambda^*), \lambda^*, c(\lambda^*)) &= \partial_x U(x^*, \lambda^*, c^*) x'(\lambda^*) + \partial_\lambda U(x^*, \lambda^*, c^*) \\ &= -\partial_x U(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*) \\ &\quad + \partial_\lambda U(x^*, \lambda^*, c^*) . \end{aligned}$$

Since U is C^2 and $H_x L$ is differentiable, the conclusion follows.

Note that since $H_x L(x^*, \lambda^*)$ is nonsingular, quadratic convergence in λ is equivalent to quadratic convergence in x . Also note that the requirement that the function $\tilde{c}$ be continuous at x^* rules out updates in which c goes to infinity. Actually it is often the case that (2.1) is satisfied for all c in which case quadratic convergence for any bounded

sequence $\{c_k\}$ is assured.

2. The diagonalized method. By a j -step diagonalization of the method of multipliers we mean the modification of replacing exact minimization by j steps of Newton's method. Since in this case x is no longer a function of λ we must perform the analysis in (x, λ) space, i.e., R^{n+m} . As the penalty constant c does not affect the analysis we will suppress it from our notation. We define two iteration functions Λ and N , both mapping R^{n+m} into itself by

$$\Lambda(x, \lambda) = (x, U(x, \lambda))$$

$$N(x, \lambda) = (x - H_x L(x, \lambda)^{-1} \nabla_x L(x, \lambda), \lambda)$$

Intermediate iterates may be defined by

$$(x^{k+1,0}, \lambda^{k+1}) = \Lambda(x^k, \lambda^k)$$

$$(x^{k+1,\ell}, \lambda^{k+1}) = N(x^{k+1,\ell-1}, \lambda^{k+1}) .$$

For a j -step diagonalization

$$(x^{k+1}, \lambda^{k+1}) = (x^{k+1,j}, \lambda^{k+1}) = S(x^k, \lambda^k)$$

Note that (x^*, λ^*) is a fixed point for N , Λ , and S . Note also that $S'(x^*, \lambda^*) = N'(x^*, \lambda^*)^j \Lambda'(x^*, \lambda^*)$.

A straightforward calculation shows that

$$\Lambda'(x^*, \lambda^*) = \begin{bmatrix} I & 0 \\ \partial_x U(x^*, \lambda^*) & \partial_\lambda U(x^*, \lambda^*) \end{bmatrix}$$

and

$$N'(x^*, \lambda^*) = \begin{bmatrix} 0 & -H_x L(x^*, \lambda^*)^{-1} \nabla g(x^*) \\ 0 & I \end{bmatrix}$$

Relations between these two quantities allow us to characterize quadratic

convergence.

Theorem 2.2. Given a C^2 update formula $U(x, \lambda)$ the following are equivalent:

- i) $\partial_x U(x^*, \lambda^*) H_x L(x^*, \lambda^*)^{-1} \nabla g(x^*) = \partial_\lambda U(x^*, \lambda^*)$
- ii) With exact minimization $x^k \rightarrow x^*$ locally quadratically
- iii) With a j -step diagonalization $j \geq 2$, $1 \leq \ell < j$
 $(x^{k, \ell}, \lambda^k) \rightarrow (x^*, \lambda^*)$ locally quadratically.

Secondly, i), ii), or iii) implies that with a j -step diagonalization $x^{k, \ell} \rightarrow x^*$ and $\lambda^{k, \ell} \rightarrow \lambda^*$ locally quadratically.

Proof: From Lemma 2.1 we know i) and ii) are equivalent.

Clearly $[N'(x^*, \lambda^*)]^2 = N'(x^*, \lambda^*)$. Also denoting $N'(x^*, \lambda^*)$ and $\Lambda'(x^*, \lambda^*)$ by N' and Λ' we have

$$N' \Lambda' = \begin{bmatrix} -H_x L^{-1} \nabla g \partial_x U & -H_x L^{-1} \nabla g \partial_x U \\ \partial_x \bar{U} & \partial_x \bar{U} \end{bmatrix} \quad (2.2)$$

$$\Lambda' N' = \begin{bmatrix} 0 & -H_x L^{-1} \nabla g \\ 0 & -\partial_x U H_x L^{-1} \nabla_\lambda U \end{bmatrix} \quad (2.3)$$

$$N' \Lambda' N' = \begin{bmatrix} 0 & H_x L^{-1} \nabla g [\partial_x U H_x L^{-1} - \partial_\lambda U] \\ 0 & -\partial_x U H_x L^{-1} \nabla g + \partial_x \bar{U} \end{bmatrix} \quad (2.4)$$

If we define the phase shifted iteration function $\tilde{S}_\ell$ such that $(x^{k+1, \ell}, \lambda^{k+1}) = \tilde{S}_\ell(x^{k, \ell}, \lambda^k)$, it follows that

$$\begin{aligned}\tilde{S}'_j(x^*, \lambda^*) &= N'(x^*, \lambda^*)^{\ell} \Lambda'(x^*, \lambda^*) N'(x^*, \lambda^*)^{j-\ell} \\ &= N'(x^*, \lambda^*) \Lambda' x^*, \lambda^*) N'(x^*, \lambda^*)\end{aligned}$$

since $N'(x^*, \lambda^*)$ is idempotent. Thus, since $\tilde{S}'_j$ is continuously differentiable the iteration $\{(x^{k,\ell}, \lambda^k)\}$ is locally quadratically convergent if and only if $\tilde{S}'_j(x^*, \lambda^*) = 0$ which is true if and only if i) is true.

For the last part one need only note that the quadratic convergence of $\{(x^{k,\ell}, \lambda^k)\}$ implies R-quadratic convergence of $\{x^{k,\ell}\}$ and $\{\lambda^k\}$ and thus R-quadratic convergence of $\{x^k\}$.

Note that $S'(x^*, \lambda^*) = N'N'\Lambda'(x^*, \lambda^*) = N'\Lambda'(x^*, \lambda^*) \neq 0$ so that we do not expect $\{(x^k, \lambda^k)\}$ to converge locally quadratically, and indeed the convergence of x^k and λ^k is generally out of phase.

Since $N'(x^*, \lambda^*)$ is idempotent extra Newton steps have no effect on convergence rates and therefore it would seem we lose nothing by using only a 2-step diagonalization, at least locally.

In the case of a one-step diagonalization two kinds of good behavior are possible, one is an immediate corollary.

Corollary 2.1 . A one-step diagonalization of the method of multipliers is 2-step locally quadratically convergent if and only if equation (2.1) is satisfied by the update. In this case it is also true that the R-order for a single iteration will be at least $\sqrt{2}$.

Proof: Two iterations of a one-step diagonalization are given by $S_2(x, \lambda) = N \circ \Lambda \circ N \circ \Lambda(x, \lambda)$. Therefore if we let $\Gamma = \partial_x U(x^*, \lambda^*) H_x L^{-1}(x^*, \lambda^*) - \partial_x U(x^*, \lambda^*)$ it follows that

$$S_2'(x^*, \lambda^*) = (N' \Lambda')^2 = \begin{bmatrix} H_x L^{-1} \nabla g \Gamma \partial_x U & H_x L^{-1} \nabla g \Gamma \partial_\lambda U \\ -\Gamma \partial_x U & \Gamma \partial_\lambda U \end{bmatrix}$$

which is zero if and only if (2.1) is satisfied.

Better single iteration results may be obtained if stronger conditions are put on the update formula.

Theorem 2.2. The one-step diagonalization of the method of multipliers is locally quadratically convergent in (x, λ) if and only if $\partial_x U(x^*, \lambda^*) = \partial_\lambda U(x^*, \lambda^*) = 0$.

Proof. From equation (2.2) it is clear that $N' \Lambda' = 0$ if and only if $\partial_x U(x^*, \lambda^*) = \partial_\lambda U(x^*, \lambda^*) = 0$. As before, quadratic convergence in R^{n+m} implies R-quadratic convergence of $\{x^k\}$ and $\{\lambda^k\}$.

The fact that from (2.3) $\Lambda' N'$ is never zero only precludes Q-quadratic convergence in R^{n+m} . The R-quadratic convergence still exists since we are looking at the same iterates out of phase with each other.

3. Application to Particular Update Functions. Based on the conditions that we have found for quadratic convergence of one-step and two-step diagonalizations, we would like to try to compare the known updates with respect to these conditions and to determine what other updates might have high convergence rates. All Lagrange multiplier update formulas appearing in the literature have one of the following two forms:

$$U(x, \lambda, c) = \lambda + B(x, \lambda, c) g(x) \quad (2.5)$$

$$U(x, \lambda, c) = -\nabla g(x)_L^{-1} \nabla f(x) + B(x, \lambda, c) g(x), \quad (2.6)$$

where $\nabla g(x)_L^{-1}$ is a left inverse of $\nabla g(x)$ which may also depend on λ

and c , and the function B maps R^{n+m+1} into $L(R^m)$. The Hestenes-Powell update and the Buys update are of the form (2.5), and the projection formula and the update proposed by Tapia are of form (2.6).

We wish to consider the most general form that we can in order to determine a large class of updates having good convergence properties. We know that we must have $U(x^*, \lambda^*, c) = \lambda^*$, that is, $U(x, \lambda, c) = \lambda + \psi(x, \lambda, c)$ where $\psi(x^*, \lambda^*, c) = 0$. The only functions dependent on f and g with this property are ∇L and g . Therefore a reasonable general form is

$$U(x, \lambda, c) = \lambda + A(x, \lambda, c) \nabla_x L(x, \lambda, c) + B(x, \lambda, c) g(x) \quad (2.7)$$

where A maps R^{n+m+1} into $L(R^n, R^m)$ and B maps R^{n+m+1} into $L(R^m)$. If we use exact minimization $\nabla L(x^*, \lambda^*, c) = 0$ and all three forms are equivalent. With diagonalization this is no longer true, but it is clear that (2.7) includes all updates of the form (2.5) and, if we let $A(x, \lambda, c) = \nabla g(x)_L^{-1}$, those of the form (2.6). Also (2.7) includes Tapia's general class of updates [6]. We now characterize local quadratic convergence for updates of this form.

Theorem 2.3. Let U be a C^2 update function of the form (2.7). If U is used with the method of multipliers with either exact minimization or a two-step diagonalization, convergence is local quadratic if and only if $B(x^*, \lambda^*, c^*) = [\nabla g(x^*)^T H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*)]^{-1}$.

Proof. By Theorem 2.1 we need A and B such that

$$\partial_x U(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*) = \partial_\lambda U(x^*, \lambda^*, c^*).$$

This has the form

$$[A(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c^*)^{-1} + B(x^*, \lambda^*, c^*) \nabla g(x^*)^T H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*)] = I + A(x^*, \lambda^*, c^*) \nabla g(x^*) .$$

This proves the theorem.

It is interesting to note that the matrix function A makes no difference at all. The theorem implies that essentially the only update of the form (2.6) yielding local quadratic convergence is the Buys update. $U(x, \lambda, c) = \lambda + [\nabla g^T H_x L^{-1} \nabla g]^{-1} g$, or something approaching it asymptotically. An update of the form (2.6) which is quadratically convergent must look like

$$U(x, \lambda, c) = -\nabla g(x)^T L^{-1} \nabla f(x) - c g(x) + [\nabla g(x)^T H_x L^{-1}(x, \lambda, c) \nabla g(x)]^{-1} g(x) .$$

Tapia has considered update formulas of this form in the context of one-step diagonalization and exact minimization with left inverses of the form $\nabla g_L^{-1} = [\nabla g^T B \nabla g]^{-1} \nabla g^T B$ where B is a matrix valued function. Note that neither the projection update nor the Hestenes-Powell update satisfy this relation.

We now consider the same question for one-step diagonalization.

Theorem 2.4. The only update of the form (2.7) which yields local quadratic convergence in (x, λ) for the one-step diagonalization is Tapia's update, that is

$$U(x, \lambda, c) = -[\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla f(x) - g(x)] - c g(x) .$$

Proof. By Theorem 2.2 we must have $\partial_x U(x^*, \lambda^*, c^*) = 0$. Therefore

$A(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c) + B(x^*, \lambda^*, c^*) \nabla g(x^*) = 0$ but we know what B must be from the previous theorem so we obtain

$$A(x^*, \lambda^*, c^*) H_x L(x^*, \lambda^*, c^*) + [\nabla g(x^*)^T H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*)]^{-1} \nabla g(x^*) = 0$$

which implies

$$A(x^*, \lambda^*, c^*) = [\nabla g(x^*)^T H_x L(x^*, \lambda^*, c^*)^{-1} \nabla g(x^*)]^{-1} \nabla g(x^*) H_x L(x^*, \lambda^*, c^*)^{-1} .$$

Therefore

$$\begin{aligned}
U(x, \lambda, c) &= \lambda + [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} [\nabla g(x)^T A_x L(x, \lambda, c)^{-1} \\
&\quad \nabla_x L(x, \lambda, c) + g(x)] \\
&= [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} [-\nabla g(x)^T H_x L(x, \lambda, c) \nabla f(x) + g(x)] - c g(x).
\end{aligned}$$

This uniqueness should not be surprising since Tapia [6] showed not only quadratic convergence of the one-step diagonalization, but that use of equation (2.8) is equivalent to using Newton's method on problem (1.2) in $\mathbb{R}^{n+m}$.

Since it is a widely held opinion that in practice superlinear convergence is nearly as good as quadratic convergence, it is reasonable to ask whether any updates exhibit superlinear convergence and how this is affected by diagonalization. In the context which we are considering here, with a single step update formula in conjunction with Newton's method, and with bounded penalty constants the situation is fairly simple.

Theorem 2.5. If the method of multipliers is used with a C^2 update function U the following are equivalent :

- i) R -superlinear convergence with exact minimization.
- ii) R -superlinear convergence with a one-step diagonalization.
- iii) For some integer p , with exact minimization, local quadratic convergence of every p -th iterate.
- iv) For the same integer p as in iii), using a two-step diagonalization on every p -th iterate and a one-step diagonalization on all other iterates, local quadratic convergence of the iterates $\{x^{kp,1}, \lambda^{kp}\}$.

Proof. As shown in Lemma 2.1 the derivative of the iteration function for exact minimization at the solution is $\partial_\lambda U - \partial_x U H_x L^{-1} \nabla g$ which we will denote

by Γ . It is well known that R-superlinear convergence occurs if and only if the spectral radius $\rho(r) = 0$. For the one-step diagonalization the iteration function is $N^0 \Lambda$ whose derivative at the solution is given by (2.2). It is easy to show that

$$(N' \Lambda')^P = \begin{bmatrix} H_x L^{-1} \nabla g(-\Gamma)^{P-1} \partial_x U & H_x L^{-1} \nabla g(-\Gamma)^{P-1} \partial_\lambda U \\ \Gamma^{P-1} \partial_x U & \Gamma^{P-1} \partial_\lambda U \end{bmatrix}$$

Therefore $\rho(N' \Lambda') = 0$ if and only if $\rho(\Gamma) = 0$ which implies i) and ii) are equivalent. Since p -step quadratic convergence is equivalent to p -th order nilpotence of the derivative of the iteration function i) and iii) are equivalent. Noting that

$$(N' \Lambda')^{PN} = \begin{bmatrix} 0 & -H_x L^{-1} \nabla g(-\Gamma)^P \\ 0 & \Gamma^P \end{bmatrix}$$

it is clear that iii) is equivalent to iv), which proves the theorem.

This result, however, is of little practical interest since there is no proposed update formula with $\rho(r) = 0$ unless $\Gamma = 0$. Indeed, if we look at updates of the form (2.7) we see that in order to have $\rho(\Gamma) = 0$, in light of theorem 3.1 we must have

$$\rho(B(\nabla g^T H_x L^{-1} \nabla g) - I) = 0$$

that is, B must be almost an inverse of $\nabla g^T H_x L^{-1} \nabla g$, a condition which is unlikely to be achieved by an explicit function B unless it is the exact inverse.

III. Convergence with Unbounded Penalty Constants.

1. Large penalty constants and updates. A number of updates exhibit accelerated convergence if the sequence of penalty constants $\{c^k\}$ increases to infinity. Indeed, Bertsekas proves superlinear convergence for the Hestenes-Powell update with exact minimization if the penalty constants satisfy this condition. Miele et al [5] implement a diagonalized scheme with unbounded penalty constants, but find it necessary to use a special scheme for increasing c^k in order to get good results.

In this section we consider the effect of an unbounded sequence of penalty constants on the properties of various update formulas, and then take into account the destabilizing effect of this on Newton's method in order to determine efficient schemes for increasing the penalty constant.

To facilitate our discussion we define the following functional which is actually the classical Lagrangian.

$$F(x, \lambda) = L(x, \lambda, c) = f(x) + \lambda^T g(x)$$

so that

$$L(x, \lambda, c) = F(x, \lambda) + \frac{1}{2} c g(x)^T g(x) .$$

The following lemma is similar to a result proved by Bartsekas [4] and is the basis for many of the results in the rest of this section.

Lemma 3.1. For F and L as defined above and for $H_x F(x, \lambda)$ nonsingular and positive definite on the null space of $g'(x)$ we have

$$H_x L(x, \lambda, c)^{-1} \nabla g(x) = \frac{1}{c} H_x F^{-1}(x, \tilde{\lambda}) \nabla g(x) [\nabla g(x)^T H_x F(x, \tilde{\lambda})^{-1} \nabla g(x) + \frac{1}{c} I]^{-1} \quad (3.1)$$

$$\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x) \leq \frac{1}{c} I - \frac{1}{c^2} [\nabla g(x)^T H_x F(x, \tilde{\lambda})^{-1} \nabla g(x) + \frac{1}{c} I]^{-1} \quad (3.2)$$

where $\tilde{\lambda} = \lambda + cg$. Additionally given $\alpha < 1$ $H_x F(x, \tilde{\lambda})$ and

$\nabla g(x)^T H_x F(x, \tilde{\lambda})^{-1} \nabla g(x)$ are invertible for all c such that

$$c \leq \alpha \|Hg(x)\|^{-1} \|g(x)\|^{-1} \min \left(\|H_x F(x, \lambda)^{-1}\|, \inf_{g'(x)\eta=0} \frac{\eta^T H_x F(x, \lambda) \eta}{\|\eta\|^2} \right)$$

Also the quantity $\beta = \sup \|\nabla g(x)^T H_x F(x, \tilde{\lambda})^{-1} \nabla g(x)\|$ over all such c exists and

$$\|H_x L(x, \lambda, c)^{-1} \nabla g(x)\| \leq \frac{1}{c} \frac{1}{1-\alpha} \|H_x F(x, \lambda)^{-1} \nabla g(x)\| \frac{1}{\beta^{-1} - c^{-1}} \quad (3.3)$$

and

$$\|\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x) - \frac{1}{c} I\| \leq \frac{1}{c^2} \frac{1}{\beta^{-1} - c^{-1}} \quad (3.4)$$

Proof. $H_x L(x, \lambda, c) = H_x F(x, \tilde{\lambda}) + c \nabla g(x) \nabla g(x)^T$

By the Sherman-Morrison formula letting $\tilde{F} = F(x, \tilde{\lambda})$

$$\begin{aligned} H_x L^{-1} &= H_x \tilde{F}^{-1} \nabla g - H_x \tilde{F}^{-1} c \nabla g [I + c \nabla g^T H_x \tilde{F}^{-1} \nabla g]^{-1} \nabla g^T H_x \tilde{F}^{-1} \nabla g \\ &= \frac{1}{c} H_x \tilde{F}^{-1} \nabla g [\nabla g^T H_x \tilde{F}^{-1} \nabla g + \frac{1}{c} I]^{-1}. \end{aligned}$$

Equation (3.2) follows easily. Additionally

$$\begin{aligned} \|H_x F(x, \tilde{\lambda})^{-1}\| &\leq \|H_x F(x, \lambda)^{-1}\| [1 - \|H_x F(x, \lambda)^{-1}\| \|cHg(x)g(x)\|]^{-1} \\ &\leq \frac{1}{1-\alpha} \|H_x F(x, \lambda)^{-1}\|. \end{aligned}$$

To show $\nabla g^T H_x \tilde{F}^{-1} \nabla g$ invertible, note that, given the restrictions on c , for any η such that $g'(x)\eta = 0$

$$\eta^T H_x \tilde{F} \eta = \eta^T H_x F \eta + \eta^T c H g g^T \eta \geq \left[\inf_{g'(x)\eta=0} \frac{\eta^T H_x F \eta}{\|\eta\|^2} - c \|Hg\| \|g\| \right] \|\eta\|^2 > 0$$

where the infimum exists since $H_x F$ is positive definite on the null space

of $g'(x)$. Thus $H_x \tilde{F}$ is positive definite on the same subspace. But then if $\nabla g^T H_x \tilde{F}^{-1} \nabla g \eta = 0$ for some η , we would have

$$(H_x \tilde{F}^{-1} \nabla g \eta)^T H_x \tilde{F} H_x \tilde{F}^{-1} \nabla g \eta > 0$$

$$\eta^T \nabla g^T H_x \tilde{F}^{-1} \nabla g \eta > 0$$

a contradiction.

Since $\nabla g^T H_x \tilde{F}^{-1} \nabla g$ is invertible for all c in a closed interval the supremum β must exist. Therefore

$$\begin{aligned} \|[\nabla g^T H_x \tilde{F}^{-1} \nabla g + \frac{1}{c} I]^{-1}\| &\leq \|[\nabla g^T H_x \tilde{F}^{-1} \nabla g]^{-1}\| [1 - \|\nabla g^T H_x \tilde{F}^{-1} \nabla g^{-1}\| \frac{1}{c}]^{-1} \\ &\leq \beta [1 - \frac{\beta}{c}]^{-1} \end{aligned}$$

and the last two inequalities follow.

The restrictions placed on c in the statement of the lemma are probably stronger than necessary but they are satisfied by almost any reasonable scheme for picking c . The positive definiteness of $H_x F(x, \lambda)$ on the null space of $g'(x)$ is of course satisfied in a neighborhood of the solution to a constrained minimization problem if $H_x F(x^*, \lambda^*)$ is non singular.

If we apply the second result of Lemma 3.1 to the Buys update, $U(x, \lambda, c) = \lambda + [\nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)]^{-1} g(x)$ we see that as c goes to infinity $[\nabla g^T H_x L^{-1} \nabla g]^{-1}$ goes to cI and this update tends toward the Hestenes-Powell update as noted by Bertsekas. In this section we will show that the tendency towards fast updates for large c occurs for a much larger class of updates.

In Section II it was shown that local convergence for either exact minimization or for diagonalization is dependent on the matrix

$$\Gamma(x, \lambda, c) = \partial_{\lambda} U(x, \lambda, c) - \partial_x U(x, \lambda, c) H_x L(x, \lambda, c)^{-1} \nabla g(x).$$

Updates for which Γ goes to zero for large c should yield fast convergence and therefore we will give them a special designation.

Definition 3.1. An update function is called pseudo-quadratic if there exist constants r, γ_1 and γ_2 such that $\|x - x^*\|, \|\lambda - \lambda^*\| < r$ implies $\Gamma(x, \lambda, c) \leq \frac{\gamma_1}{c} + \gamma_2(\|x - x^*\| + \|\lambda - \lambda^*\|)$ for c sufficiently large.

We would like to show that this class of updates is large and to do that we note the following definition due to Tapia.

Definition 3.2. An update function which is not dependent on λ is called an approximation formula. If it satisfies $\partial_{\lambda} U(x^*, \lambda^*, c) = 0$ for all c it is called a local approximation formula.

The projection update (1.6) is an example of an approximation formula and it can be shown that any u-date of form (2.6) is a local approximation formula. One useful property of such updates is given in the following lemma.

Lemma 3.2. The Hestenes-Powell update, and all local approximation formulas whose second derivative is bounded in a neighborhood of the solution uniformly in c are pseudo-quadratic updates.

Proof. For the Hestenes-Powell update

$$\begin{aligned} \|\Gamma(x, \lambda, c)\| &= \|I - c \nabla g(x)^T H_x L(x, \lambda, c)^{-1} \nabla g(x)\| \\ &\leq \frac{1}{c} \frac{1}{\beta(x, \lambda)^{-1} - c^{-1}} \end{aligned}$$

where $\beta(x, \lambda)$ is the β of Lemma 3.1 and is bounded in a neighborhood of (x^*, λ^*) .

For a local approximation formula

$$\begin{aligned} \|\Gamma(x, \lambda, c)\| &= \|\partial_\lambda U(x, \lambda, c)\| + \|\partial_x U(x, \lambda, c)\| \|H_x^{-1}(x, \lambda, c) \nabla g(x)\| \\ &\leq \sup_{\|x - x^*\|, \|\lambda - \lambda^*\| \leq r} \|\partial_\lambda U'(x, \lambda, c)\| [\|\lambda - \lambda^*\| + \|x - x^*\|] + \\ &\quad \frac{1}{c} \sup_{\|x - x^*\|, \|\lambda - \lambda^*\| \leq r} [\partial_x U(x, \lambda, c)] \beta(x, \lambda) \left(1 - \frac{\beta(x, \lambda)}{c}\right)^{-1} \\ &\leq \gamma_2 (\|\lambda - \lambda^*\| + \|x - x^*\|) + \gamma_1 \frac{1}{c} . \end{aligned}$$

Obvious examples of approximation formulas whose derivatives are bounded are the projection formula and any C^2 approximation formula not dependent on c .

Given a pseudo-quadratic update it is easy to prove fast convergence for exact minimization which we will do here. It should be noted that by local superlinear and local p -order convergence we mean the same strengthening of the usual notions as with local quadratic convergence.

Theorem 3.1. If a pseudo-quadratic update is used with exact minimization and with a sequence of penalty constants such that $c_k \rightarrow \infty$ then $\lambda_k \rightarrow \lambda^*$ locally superlinearly for c^0 sufficiently large. If in addition we choose $c_{k+1} = \frac{\tilde{c}}{g(x_k)^\alpha}$, $0 \leq \alpha \leq 1$, $\lambda_k \rightarrow \lambda^*$ locally with order $1 + \alpha$.

Proof. We are dealing with the iteration $\lambda^{k+1} = S_{c^k}(\lambda^k)$ where $S_c(\lambda) = U(x(\lambda), \lambda, c)$. As in Theorem 2.1,

$$S'_c(\lambda) = \partial_\lambda U(x, \lambda, c) - \partial_x(x, \lambda, c) H_x L(x, \lambda, c)^{-1} \nabla g(x)$$

$$= \Gamma(x, \lambda, c)$$

$$\leq \frac{1}{c} \gamma_1 + \gamma_2 [\|x - x^*\| + \|\lambda - \lambda^*\|]$$

$$\leq \frac{1}{c} \gamma_1 + \gamma_3 \|\lambda - \lambda^*\|$$

$$\text{since } \|x - x^*\| \leq \sup_{\mu} \|\mu - \lambda^*\| \leq r \|\mathbf{x}'(\mu)\| \|\lambda - \lambda^*\|$$

where r is the radius of the ball where U is pseudo-quadratic.

Now,

$$\|\lambda^{k+1} - \lambda^*\| \leq \|S_{c^k}(\lambda^k) - S_{c^k}(\lambda^*)\|$$

$$\leq \|S'_{c^k}(\lambda^* + \theta(\lambda^k - \lambda^*))\| \|\lambda^k - \lambda^*\| \quad 0 \leq \theta \leq 1$$

$$\leq \left[\frac{1}{c^k} \gamma_1 + \gamma_3 \|\lambda^k - \lambda^*\|\right] \|\lambda^k - \lambda^*\|$$

which implies local superlinear convergence if $c^k \rightarrow \infty$. If

$$c^k = \frac{\tilde{c}}{g(x^k)^\alpha} \leq \tilde{c} \gamma \|\lambda^k - \lambda^*\|^{-\alpha} \quad \text{then we have}$$

$$\|\lambda^{k+1} - \lambda^*\| \leq \gamma_1 \gamma \tilde{c} \|\lambda^k - \lambda^*\|^{1+\alpha} + \gamma_3 \|\lambda^k - \lambda^*\|^2.$$

which implies local $1+\alpha$ order convergence.

The most interesting feature of this result is that it allows us to obtain quadratic convergence with updates such as the projection formula or the Hestenes-Powell update which require no information about the second derivative or inversion of the Hessian, whereas when c is bounded all update formulas yielding superlinear convergence contain the Hessian

of the Lagrangian. Some heuristic justification for this may be seen by noting that as c increases we are accomplishing more of the effort of satisfying the constraint in the exact minimization stage and the Lagrange multiplier update need only keep up with the x convergence linearly, which is certainly true of any approximation formula. If we believe this it is only natural to ask whether we can retain this quadratic convergence under diagonalization.

2. Diagonalization with large penalty constants. In Section II we saw that once the basic condition $\Gamma(x^*, \lambda^*, c) = 0$ is satisfied good convergence is exhibited for diagonalization as well as for exact minimization. However, computational experience has shown that with a one-step diagonalization, with the penalty constant determined by $c^{k+1} = \frac{\tilde{c}}{\|g(x^k)\|}$, and the projection formula used as an update, convergence is extremely slow. Even if the Tapia update, which gives quadratic convergence for bounded penalty constants, is used convergence is at best linear if the penalty constant is increased as above. However, Miele et al [5] have obtained fast convergence from an implementation where $c^{k+1} = \frac{\tilde{c}}{g(x^k)}$ if $g(x^k)^T g(x^k) \geq \nabla L(x^k, \lambda^k, c^k)^T \nabla L(x^k, \lambda^k, c^k)$ and $c^{k+1} = c^k$ otherwise.

The results for diagonalization in Section II went through easily because the iteration function for Newton's method has a zero derivative at the solution and its second derivative was bounded. However, when the penalty constant changes we are looking at a sequence of iteration functions indexed on c^k , and their second derivatives are not uniformly bounded if the sequence $\{c^k\}$ is not bounded. Therefore one cannot adapt any of the standard proofs of quadratic convergence. It seems then that what we gain in

the update as c gets large we lose in the Newton step. However, we can show that two Newton steps in succession will give fast convergence if the ratio between the constants is chosen correctly.

We denote by $\bar{N}_c$ the x coordinate of a Newton step with penalty constant c ; i.e., $\bar{N}_c(x, \lambda) = x - H_x L(x, \lambda, c)^{-1} \nabla_x L(x, \lambda, c)$. Similarly to the notation introduced in Section II $N_c(x, \lambda) = (\bar{N}_c(x, \lambda), \lambda)$. We now prove a lemma clarifying the behavior of Newton's method for large values of c .

Lemma 3.3. If f and g are C^3 functions and if we define the Newton step as above, there exist constants $\tilde{c}$, r , and γ such that for any (x, λ) with $\|x - x^*\| \leq r$, $\|\lambda - \lambda^*\| \leq r$, and any $c \leq \frac{\tilde{c}}{\|g(x)\|}$ we have

$$\|\bar{N}_c(x, \lambda) - x^*\| \leq \gamma[\|x - x^*\| + \|\lambda - \lambda^*\| + c\|g(x)\|]\|x - x^*\| + \frac{\gamma}{c}\|\lambda - \lambda^*\| \quad (3.5)$$

and

$$\|g(\bar{N}_c(x, \lambda))\| \leq \gamma(\|x - x^*\|^2 + \|x - x^*\| \|\lambda - \lambda^*\|) + \frac{\gamma}{c}\|\lambda - \lambda^*\| \quad (3.6)$$

Proof. $N_c(x, \lambda) - x^* = N_c(x, \lambda) - N_c(x^*, \lambda) + N_c(x^*, \lambda) - N_c(x^*, \lambda^*)$.

Considering first the second two terms,

$$\begin{aligned} \bar{N}_c(x^*, \lambda) - \bar{N}_c(x^*, \lambda^*) &= x^* - H_x L(x^*, \lambda, c)^{-1} \nabla_x L(x^*, \lambda, c) - x^* \\ \bar{N}_c(x^*, \lambda) - \bar{N}_c(x^*, \lambda^*) &= H_x L(x^*, \lambda, c)^{-1} \nabla g(x^*) (\lambda - \lambda^*)^T \\ \|\bar{N}_c(x^*, \lambda) - \bar{N}_c(x^*, \lambda^*)\| &\leq \frac{1}{c} \gamma \|\lambda - \lambda^*\| \end{aligned} \quad (3.7)$$

by Lemma 3.1,

Using a form of the mean value theorem, we get from the second term

$$\bar{N}_c(x, \lambda) - \bar{N}_c(x^*, \lambda) = \sum_{i=1}^n \alpha_i \partial_x \bar{N}_c(z_i, \lambda, c) (x - x^*)$$

where $z_i = x^* + \theta_i (x - x^*)$, $0 \leq \theta_i \leq 1$, $i = 1, n$, and $\sum_{i=1}^n \alpha_i = 1$, $\alpha_i \geq 0$.

$$\begin{aligned}
\partial_z \bar{N}_c(z, \lambda, c) &= I - H_z L(z, \lambda, c)^{-1} H_z L(z, \lambda, c) - \partial_z \left(H_z L(z, \lambda, c)^{-1} \right) \nabla_z L(z, \lambda, c) \\
&= H_z L(z, \lambda, c)^{-1} \partial_z H_z L(z, \lambda, c) H_z L(z, \lambda, c)^{-1} \nabla_z L(z, \lambda, c)
\end{aligned}$$

by differentiating the inverse .

$$\begin{aligned}
\partial_z \bar{N}_c(z, \lambda, c) &= H_z L(z, \lambda, c)^{-1} \left[\partial_z H_z F(z, \tilde{\lambda}) + c Hg(z) \nabla g(z)^T + c \nabla g(z) Hg(z) \right] \\
&\quad H_z L(z, \lambda, c)^{-1} \left[\nabla_z F(z, \lambda) + c \nabla g(z) g(z) \right]
\end{aligned}$$

where $\tilde{\lambda} = \lambda + cg(z)$. Note that by our hypothesis $\|\tilde{\lambda}\| \leq \|\lambda\| + \tilde{c}$ for all z and c . In applying Lemma 3.1 to the above expression we note that in every term but one the expression $H_z L(z, \lambda, c)^{-1} \nabla g(z)$ occurs as many times as the multiplier c . Since the operators $H_z L(z, \lambda, c)^{-1}$, $\partial_z H_z F(z, \tilde{\lambda})$ and Hg are all bounded uniformly in c in a ball of radius r about x^* , we get that

$$\begin{aligned}
\partial_x \bar{N}_c(x, \lambda, c) &= M_1 \nabla_z F(z, \lambda) + M_2 g(z)^T + c^2 H_z L(z, \lambda, c)^{-1} Hg \nabla g(z)^T \\
&\quad H_z L(z, \lambda, c)^{-1} \nabla g(z) g(z)
\end{aligned} \tag{3.8}$$

where M_1 and M_2 are operators bounded uniformly in c . Noting that $\nabla_z F(z, \lambda) \leq \gamma(\|z - x^*\| + \|\lambda - \lambda^*\|) \leq \gamma(\|x - x^*\| + \|\lambda - \lambda^*\|)$ and

$g(z) \leq \gamma\|z - x^*\| \leq \gamma\|z - x^*\|$ for some constant γ we get from (3.8)

$$\begin{aligned}
\|\bar{N}_c(x, \lambda) - \bar{N}_c(x^*, \lambda)\| &\leq \sum_{i=1}^n \alpha_i \gamma(\|x - x^*\| + \|\lambda - \lambda^*\|) \|x - x^*\| \\
&\leq \gamma(\|x - x^*\| + \|\lambda - \lambda^*\|) \|x - x^*\|.
\end{aligned}$$

Combined with equation (3.5) this gives us the result (3.3).

For the second part we note that

$$\|g(N(z, \lambda)) - g(x^*)\| \leq \|\nabla g(z)\|^T (\bar{N}_c(z, \lambda) - x^*)\| + \gamma \|\bar{N}_c(z, \lambda) - x^*\|^2$$

for some constant γ . But from equation (3.8) we see

$$\begin{aligned} \|\nabla g(z) (\bar{N}_c(z, \lambda) - x^*)\| &\leq \gamma [\|x - x^*\| + \|\lambda - \lambda^*\|] \|x - x^*\| + \frac{\gamma}{c} \|\lambda - \lambda^*\| \\ &\quad + c^2 \gamma \|\nabla g(z) H_z L(z, \lambda, c)\|^2 \|g(x)\| \\ &\leq \gamma [\|x - x^*\| + \|\lambda - \lambda^*\|] \|x - x^*\| + \frac{\gamma}{c} \|\lambda - \lambda^*\|. \end{aligned}$$

From this the results, equation (3.6) follows immediately.

The essence of this result is that for large c one Newton step does not necessarily act in a local superlinear fashion or even like a contraction. However, the quantity $g(x)$ is reduced superlinearly by a Newton step. This fact means that there is much better behavior for two Newton steps in succession as we show here.

Lemma 3.4. Under the conditions of Lemma 3.3 there is a constant γ such that for $c \leq \frac{\tilde{c}}{\|g(x)\|}$ and $c' \leq \frac{\tilde{c}}{\|g(\bar{N}(x, \lambda))\|}$

$$\begin{aligned} \|\bar{N}_c, (N_c(x, \lambda), \lambda') - x^*\| &\leq \gamma [\|x - x^*\| + \|\lambda - \lambda^*\|] [1 + c' \|x - x^*\| + \frac{c'}{c}] \|x - x^*\| \\ &\quad + \frac{\gamma}{c} \|\lambda' - \lambda^*\| \end{aligned} \quad (3.9)$$

Additionally if U is a local approximation formula such that

$$\begin{aligned} \|U(x, \lambda, c) - \lambda^*\| &\leq \gamma' \|x - x^*\| \quad \text{for some } \gamma' \quad \text{and if } \lambda \quad \text{is chosen so that} \\ \|\lambda - \lambda^*\| &\leq \gamma' \|x - x^*\| \quad \text{and } c' \geq c \geq 1 \quad \text{then} \end{aligned}$$

$$\|\bar{N}_c, (\bar{N}_c(x, \lambda), \lambda) - x^*\| \leq \gamma [c' (\|x - x^*\| + \frac{1}{c}) (c \|x - x^*\| + \frac{1}{c}) \|x - x^*\| + \frac{1}{c} \|x - x^*\|] \quad (3.10)$$

and

$$\begin{aligned} \|\bar{N}_c, (\bar{N}_c(x, \lambda), U(x, \lambda, c)) - x^*\| &\leq \gamma [c' (\|x - x^*\| + \frac{1}{c}) \|x - x^*\| + \frac{1}{c}] \\ &\quad (c \|x - x^*\| + \frac{1}{c}) \|x - x^*\| \end{aligned} \quad (3.11)$$

Proof. Substituting equation (3.6) with c into equation (3.5) with c' immediately yields (3.9). To get (3.10) let $\epsilon = \|x - x^*\|$ and $\epsilon' = \|\bar{N}_c(x, \lambda) - x^*\|$. Then our assumptions with (3.5) and (3.6) give $\epsilon' \leq \gamma c \epsilon^2 + \frac{\gamma}{c} \epsilon$ and $\|g(N(x, \lambda))\| \leq \gamma \epsilon^2 + \frac{\gamma}{c} \epsilon$ where γ is used to denote any constant independent of x, λ and c . Clearly, then

$$\begin{aligned} \|\bar{N}_c, (\bar{N}_c(x, \lambda), \lambda) - x^*\| &\leq \gamma(\epsilon' + \epsilon + c'g)\epsilon' + \frac{\gamma}{c'}\epsilon \\ &\leq \gamma(\epsilon + c'\epsilon^2 + \frac{c'}{c}\epsilon)(c\epsilon^2 + \frac{1}{c}\epsilon) + \frac{\gamma}{c'}\epsilon \\ &\leq \gamma c'(\epsilon + \frac{1}{c})(c\epsilon + \frac{1}{c})\epsilon + \frac{\gamma}{c'}\epsilon \end{aligned}$$

In the case of (3.11)

$$\begin{aligned} \|\bar{N}_c, (\bar{N}_c(x, \lambda), U(x, \lambda, c) - x^*)\| &\leq \gamma(\epsilon' + c'g)\epsilon' + \frac{\gamma}{c'}\epsilon' \\ &\leq \gamma(c\epsilon^2 + \frac{1}{c}\epsilon + c'(\epsilon^2 + \frac{1}{c}\epsilon) + \frac{1}{c'}) (c\epsilon^2 + \frac{1}{c}\epsilon) \\ &\leq \gamma(c'\epsilon^2 + \frac{c'}{c}\epsilon + \frac{1}{c'}) (c\epsilon + \frac{1}{c})\epsilon \end{aligned}$$

and (3.11) easily follows.

With this lemma the dependence of convergence on updating schemes becomes clear. Equation (3.11) indicates that for any initial x and λ two Newton steps will give local quadratic convergence if the next λ is linearly dependent on $\bar{N}_c(x, \lambda)$, and if $c = c' = \frac{\tilde{c}}{\|x - x^*\|}$ for some fixed $\tilde{c}$. Note, however, that if we choose $c = \frac{\tilde{c}}{\|x - x^*\|}$ and $c' = \tilde{c} \|g(\bar{N}_c(x, \lambda))\|^{-1} \geq \gamma \tilde{c} \frac{1}{\|x - x^*\|^2}$ the term $\frac{c'}{c}$ prevents us from getting quadratic convergence. Lemma 3.3 gives an indication of why the implementation of Miele mentioned earlier works well. In that scheme c is held constant if $\|g\|$ is smaller than $\|\nabla L\|$, but from (3.5) we see that this would make the term $c\|g(x)\|$

negligible, and equation (3.6) shows that whenever we do not get a quadratic step $\|g\|$ should be smaller than $\|\nabla L\|$.

We now wish to determine what the best convergence attainable is for one- and two-step diagonalizations. Equation (3.10) represents the effect of two Newton steps. The error in the Lagrange multiplier, if we make an update after the second Newton step, is simply the error in x times some fixed constant. If we wish better than quadratic convergence, the last term implies that we must have $c' \geq \tilde{c}\|x - x^*\|^{-q}$ for some $q > 1$. However, there is a term $c'\|x - x^*\|^3$ which would then be greater than $\|x - x^*\|^{3-q}$ thus losing quadratic convergence. Thus quadratic convergence is the best we can hope for as we state here.

Theorem 3.2. The two-step diagonalization $x^{k+1} + \bar{N}_c, (\bar{N}_c(x^k, \lambda^k), \lambda^k)$, $\lambda^{k+1} = U(x^{k+1}, \lambda^k, c')$ where $c' = c = \tilde{c}\|\nabla_x F(x, \lambda)\|^{-1}$ is locally quadratically convergent in x for $\tilde{c}$ sufficiently small.

Proof. Near (x^*, λ^*) , $\|\nabla_x F(x, \lambda)\|$ is of the same order as $\|x - x^*\|$ since $H_x F(x^*, \lambda^*)$ is non-singular, thus $c_1 \leq c\|x - x^*\| \leq c_2$ in a neighborhood of (x^*, λ^*) for some c_1, c_2 . Substituting into equation (3.10) gives the result.

In the case of a one-step diagonalization one can get two-step behavior which is a little better than quadratic. To investigate this, let $\epsilon = \|x - x^*\|$ and assume $c = \tilde{c}\epsilon^{-p}$, $c' = \tilde{c}\epsilon^{-q}$. Then the error after two complete iterations is, by (3.9), proportional to $(\epsilon^{-q}(\epsilon + \epsilon^p)\epsilon + \epsilon^q)(\epsilon^{1-p} + \epsilon^p)\epsilon$ or, since p must be less than 1, $(\epsilon^{1+p-q} + \epsilon^q)(\epsilon^{1-p} + \epsilon^p)\epsilon$. Clearly the the best we can do here is to let $q = \frac{p+1}{2}$ which gives an error of $(\epsilon^{1-p/2} + \epsilon^{3p/2})\epsilon^{3/2}$. The optimal choice of p is $\frac{1}{2}$ implying $q = \frac{3}{4}$.

and indicating a convergence rate of $2\frac{1}{4}$, although we could get any rate between 1 and $2\frac{1}{4}$. Indeed, $p=q=1$ gives quadratic convergence as in the two-step diagonalization. The scheme $p=\frac{1}{2}$, $q=\frac{3}{4}$ is actually identical to the simple scheme described in the following theorem.

Theorem 3.3. The one-step diagonalization $c = \|\nabla F(x^k, \lambda^k)\|^{-\frac{1}{2}}$, $x^{k+1} = \bar{N}_c(x^k, \lambda^k)$, $\lambda^{k+1} = U(x^{k+1}, \lambda^k, c)$ is $1\frac{1}{2}$ order locally convergent.

Proof. As in Theorem 3.2, the quantity $c\|x - x^*\|^{\frac{1}{2}}$ is near constant and the result is immediate from equation (3.5).

Note that the rate for two iterations is $(1\frac{1}{2})^2 = 2\frac{1}{4}$, and for the following step $c = \|\nabla F(x^{k+1}, \lambda^{k+1})\|^{-\frac{1}{2}} = \|x^k - x^*\|^{3/4}$ which is the $\frac{1}{2}$, $\frac{3}{4}$ scheme mentioned above.

The schemes presented in Theorems 3.2 and 3.3 seem to majorize the others in this analysis; one offers a convergence rate of $\frac{1}{4}$ more, the other saves one update step. However, given the instability of letting c tend toward infinity, these benefits are of questionable significance.

3. Conclusion. Our results clearly indicate that as a local strategy it is wasteful to do exact minimization or even to take more than two Newton steps toward a solution. In the case of bounded penalty constants, the work involved in using Tapia's update is only slightly more than in the Buys update or other second order updates. There are no extra evaluations and the only algebraic saving is in not having to solve the $n \times n$ system for $HL^{-1}\nabla f$ which is only a difference of order n^2 out of $\frac{1}{3}n^3$. Against this must be balanced the work of the extra functional evaluations to two Newton steps. It would seem then that when one is using Newton's method and has the Hessian of the Lagrangian available that the Tapia update with a one-step

diagonalization is the most efficient.

In the case of unbounded penalty constants the simpler updates such as the projection become competitive. Both the two-step diagonalization and the one-step diagonalization described in Theorem 3.3 give convergence rates of two or more for every two Newton steps. The projection formula involves extra computation of the order of m^2n where m is the number of constraints. Tapia's update involves m back substitutions which is mn^2 operations beyond what is done in the Newton step. Thus the projection formula might be advantageous if there are many constraints and the functions involved are simple to evaluate. However, one must then also reckon with instability introduced in letting the penalty constant go to infinity.

BIBLIOGRAPHY

1. Hestenes, M.R., "Multiplier and gradient methods," Journal of Optimization Theory and Applications, Vol. 4, pp. 303-320, 1969.
2. Powell, M.J.D., "A method for nonlinear constraints in minimization problems," in Optimization, Edited by R. Fletcher, Academic Press, London, England, 1969.
3. Buys, J.D., "Dual algorithms for constrained optimization," Ph.D. thesis, Rijksuniversiteit, de Leiden, the Netherlands, 1972.
4. Bertsekas, D.P., "Combined primal dual and penalty methods for constrained minimization," SIAM Journal on Control, Vol. 13, pp. 521-543, 1975.
5. Miele, A., Moseley, P.E., Levy, A.V., and Coggins, G.M., "On the method of multipliers for mathematical programming problems," Journal of Optimization Theory and Applications, Vol. 10, pp. 1-33, 1972.
6. Tapia, R.A., "Quasi-Newton methods for constrained optimization," Department of Mathematical Sciences Technical Report, Rice University, Houston, Texas, June 1975.
7. Fletcher, R., "An ideal penalty function for constrained optimization," Journal of the Institute of Mathematics and its Applications, Vol. 15, pp. 319-342, 1975.
8. Tapia, R.A., "Newton's method for problems with equality constraints," SIAM Journal on Numerical Analysis, Vol. 11, pp. 174-196, 1974.
9. Ortega, J.M., and Rheinboldt, W.C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, New York, 1970.
10. Tapia, R.A., "Newton's method for optimization problems with equality constraints," SIAM Journal on Numerical Analysis, Vol. 11, pp. 874-886, 1974.
11. Luenberger, D.G., Introduction to Linear and Nonlinear Programming, Addison Wesley, Reading, Massachusetts, 1973.
12. Haarhof, P.D., and Buys, J.D., "A new method for the optimization of a nonlinear function subject to nonlinear constraints," Computing Journal, Vol. 13, pp. 178-184, 1970.
13. Rupp, R.D., "On the combination of the multiplier method of Hestenes and Powell with Newton's method," Journal of Optimization Theory and Applications, Vol. 15, pp. 167-187, 1975.

14. Powell, M.J.D., "A new algorithm for constrained optimization," in Nonlinear Programming, Edited by J. B. Rosen, O. L. Mangasarian and K. Ritter, Academic Press, New York, 1970.
15. Bard, Y., and Greenstadt, J., "A modified Newton method for optimization with equality constraints," in Optimization, Edited by R. Fletcher, Academic Press, London, England, 1969