SECOND DERIVATIVES FOR OPTIMIZING EIGENVALUES OF SYMMETRIC MATRICES

MICHAEL L. OVERTON* AND ROBERT S. WOMERSLEY†

Abstract. Let A denote an $n \times n$ real symmetric matrix-valued function depending on a vector of real parameters, $x \in \Re^m$. Assume that A is a twice continuously differentiable function of x, with the second derivative satisfying a Lipschitz condition. Consider the following optimization problem: minimize the largest eigenvalue of $A(x)$. Let x^* denote a minimum. Typically, the maximum eigenvalue of $A(x^*)$ is multiple, so the objective function is not differentiable at x^*, and straightforward application of Newton's method is not possible. Nonetheless, the formulation of a method with local quadratic convergence is possible. The main idea is to minimize the maximum eigenvalue subject to a constraint that this eigenvalue has a certain multiplicity. The manifold Ω of matrices with such multiple eigenvalues is parameterized using a matrix exponential representation, leading to the definition of an appropriate Lagrangian function. Consideration of the Hessian of this Lagrangian function leads to the second derivative matrix used by Newton's method. The convergence proof is nonstandard because the parameterization of Ω is explicitly known only in the limit. In the special case of multiplicity one, the maximum eigenvalue is a smooth function and the method reduces to a standard Newton iteration.

Key words. nonsmooth optimization, multiple eigenvalues

AMS subject classifications. 15A18, 65F15, 65K10, 90C25

1. Introduction. Let A denote an $n \times n$ real symmetric matrix-valued function depending on a vector of real parameters, $x \in \Re^m$. Assume that A depends smoothly on x, specifically that it is at least twice continuously differentiable, with the second derivative satisfying a Lipschitz condition in x. Denote the eigenvalues of $A(x)$ by

$$\lambda_1(x) \geq \cdots \geq \lambda_n(x).$$

The eigenvalues λ_i are Lipschitz continuous functions of x [7] and, in any region where they are distinct from one another, it is well known that they are (Fréchet) differentiable; in fact, they inherit the C^2 smoothness of the function $A(x)$ [7, p.134]. Let $\widehat{x}$ be given, with

$$A(\widehat{x}) = \widehat{Q}\widehat{\Lambda}\widehat{Q}^T, \quad , \widehat{Q}^T\widehat{Q} = I, \tag{1.1}$$

where

$$\widehat{\Lambda} = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n), \quad \widehat{Q} = [\widehat{q}_1, \ldots, \widehat{q}_n]. \tag{1.2}$$

Thus, $\{\widehat{\lambda}_i\}$ and $\{\widehat{q}_i\}$ are respectively the eigenvalues and an orthonormal set of eigenvectors of $A(\widehat{x})$. Assume that $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$, so that $\widehat{\lambda}_i = \lambda_i(\widehat{x})$. Then formulas for the first and second partial derivatives of the eigenvalues λ_i at $x = \widehat{x}$, assuming that the $\widehat{\lambda}_i$ are distinct, are

$$\frac{\partial \lambda_i(\widehat{x})}{\partial x_k} = \widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_k} \widehat{q}_i \tag{1.3}$$

* Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York. The work of this author was supported in part by National Science Foundation Grant CCR-9101649. This author would also like to acknowledge the kind hospitality of the Centre for Process Systems Engineering, Imperial College, London, where part of this work was conducted with support from the U.K. Science and Engineering Research Council.

† School of Mathematics, University of New South Wales, Australia

and

$$\frac{\partial^2}{\partial x_k \partial x_j}\lambda_i(\widehat{x}) = \widehat{q}_i^T \frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_j}\widehat{q}_i + 2\sum_{s\neq i} \frac{\widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{q}_s \widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_j}\widehat{q}_s}{\widehat{\lambda}_i - \widehat{\lambda}_s}. \tag{1.4}$$

The first of these formulas is well known, and the second may be found in a variety of sources; see [8], [9], as well as (in a somewhat less accessible form) [7, p.95]. Both will follow as special cases of the results given in this paper.

However, if $A(x)$ has multiple eigenvalues at a point $x = \widehat{x}$, its eigenvalues, while still Lipschitz continuous, may not generally be written as differentiable functions of several variables at $x = \widehat{x}$. For example, consider

$$A(x) = \begin{bmatrix} 1+x_1 & x_2 \\ x_2 & 1-x_1 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_{1,2} = 1 \pm \sqrt{x_1^2 + x_2^2}.$$

Thus λ_1, the largest eigenvalue of $A(x)$, is generally not a smooth function of x; furthermore, it cannot even be written as the maximum of n smooth functions of x, if x has two or more components. Also, the eigenvectors of $A(x)$ cannot generally be written as continuous functions of x; this is a consequence of the fact that eigenvectors corresponding to simple eigenvalues are unique (up to sign and normalization) while those corresponding to multiple eigenvalues are not.

Generally speaking, applications involving eigenvalues of matrices depending on free parameters fall into one of two categories. In the first, it is specified that some or all of the eigenvalues $\lambda_i(x)$ achieve some given values λ_i^*; this is known as an *inverse eigenvalue problem*. If these given values are distinct, the inverse eigenvalue problem may be formulated as a differentiable system of nonlinear equations, and, if the number of free parameters and the number of equations is the same, the application of Newton's method is straightforward, using (1.3). In [4] it was shown how, even in the multiple eigenvalue case, the inverse eigenvalue problem may be formulated as a differentiable system of nonlinear equations, so that Newton methods, with generic quadratic convergence, are applicable.

In the second class of applications, the eigenvalues are not required to have particular values, but rather it is desired to solve some *optimization problem involving the eigenvalues*. A particularly common case is the "min-max" problem

$$\min_{x \in \Re^m} \phi(x) \tag{1.5}$$

where $\phi(x) = \lambda_1(x)$, the largest eigenvalue of $A(x)$. Let x^* be a locally unique minimizer of ϕ. If x^* has the property that the eigenvalue $\lambda_1(x^*)$ is simple, i.e. has multiplicity one, then the function to be minimized, λ_1, is twice continuously differentiable in a neighborhood of x^*, and Newton's method for unconstrained minimization may be applied, using the Hessian matrix defined by (1.4). However, it is more often the case that $A(x^*)$ has multiple eigenvalues; this is a consequence of the optimization objective, which in driving all the eigenvalues down as much as possible usually forces the coalescence of some of them. In such a case λ_1 is generally not differentiable at $x = x^*$.

This paper is concerned with the formulation of a method to solve optimization problems involving eigenvalues in exactly this case, where multiple eigenvalues occur at the solution. We shall show that the correct problem formulation leads to a method with generic quadratic convergence. This method was first given by [10], inspired in part by [3,4]. Quadratic convergence was demonstrated by numerical examples. The purpose of the present paper is primarily to prove the quadratic convergence property for the method presented in [10], justifying the Hessian matrix formulas given there, which were originally derived only formally and stated without any derivation or proof. The ideas of this paper can be applied to other classes of eigenvalue and singular value optimization problems, e.g. those discussed in [1,6,11,12,14,18,19], as well as many other references which can be found in these papers. However, we concentrate on the model problem (1.5). We consider only the issue of local convergence. For details of how to use the method and related methods in practice, see [11].

2. Tensor Notation. We shall have frequent need to refer to the first and second derivatives, with respect to several variables, of matrix-valued functions. Such objects are, respectively, tensors in three and four dimensions, a matrix being a tensor in two dimensions. We shall use subscripts to denote differentiation: thus A_x and A_{xx} refer to the first and second derivatives of the matrix-valued function A, with respect to the variable $x \in \Re^m$. Rather than attempt to describe the elements of a tensor, however, we shall describe its action as a linear operator, the result having the same dimension as the undifferentiated quantity, whether a matrix, a vector, or a scalar. For example, we write $[A_x \Delta x]$ to mean

$$\sum_{k=1}^{m} \{\Delta x\}_k \frac{\partial}{\partial x_k} A$$

and $[A_{xx} \Delta x \Delta x]$ to mean

$$\sum_{k=1}^{m} \sum_{j=1}^{m} \{\Delta x\}_k \{\Delta x\}_j \frac{\partial^2}{\partial x_k \partial x_j} A.$$

We shall reserve square brackets [,] for this purpose, and we shall use parentheses (,) primarily to mean "evaluated at". We shall use braces {, } to indicate expression precedence. For example, the first and second derivatives of $\phi(x) \equiv \lambda_1(x)$ at $x = \widehat{x}$, when $\lambda_1(\widehat{x})$ is simple, given by (1.3)–(1.4), are written in tensor notation as

$$[\phi_x(\widehat{x}) \Delta x] = \widehat{q}_1^T [A_x(\widehat{x}) \Delta x] \widehat{q}_1$$

and

$$[\phi_{xx}(\widehat{x}) \Delta x \Delta x] = \widehat{q}_1^T [A_{xx} \Delta x \Delta x] \widehat{q}_1 + 2 \sum_{s \neq 1} \frac{\{\widehat{q}_1^T [A_x \Delta x] \widehat{q}_s\}^2}{\widehat{\lambda}_1 - \widehat{\lambda}_s}.$$

Because the second derivative of a twice continuously differentiable function is symmetric with respect to its two arguments of differentiation, there is no ambiguity in this notation. There should be no confusion between those subscripts indicating differentiation and those indicating components.

We shall use $\| \cdot \|$ to denote the Euclidean vector norm. The expression $A \bullet B$, where A and B are symmetric matrices of the same dimension, means the matrix

inner product

$$A \bullet B = \text{tr } AB.$$

The operator "vec" maps the set of symmetric matrices of dimension t into the corresponding vector space $\Re^{t(t+1)/2}$, multiplying the off-diagonal components by the factor $\sqrt{2}$ so that

$$(\text{vec } A)^T(\text{vec } B) = A \bullet B.$$

Consequently

$$\|\text{vec } A\| = \|A\|_F,$$

the Frobenius norm of A.

3. The Matrix Exponential Formulation. Let x^* be a locally unique minimizer of $\phi \equiv \lambda_1$, and let $\lambda_i^* = \lambda_i(x^*)$, $i = 1, \ldots, n$. Suppose that

$$\lambda_1^* = \cdots = \lambda_t^* \ > \ \lambda_{t+1}^* > \cdots > \lambda_n^* \tag{3.1}$$

i.e. the maximum eigenvalue of $A(x^*)$ has multiplicity t, but all other eigenvalues are simple. The latter assumption usually holds in practice; it could be relaxed, at the cost of more complex notation. Let

$$\Lambda_1^* = \lambda_1^* I, \quad \Lambda_2^* = \text{Diag}(\lambda_{t+1}^*, \ldots, \lambda_n^*), \tag{3.2}$$

the identity block having order t, and let $Q^* = [q_1^*, \ldots, q_n^*]$ be a corresponding orthogonal basis of eigenvectors, with

$$Q_1^* = [q_1^* \ldots q_t^*], \quad Q_2^* = [q_{t+1}^* \cdots q_n^*]. \tag{3.3}$$

The matrix Q_2^* is unique, up to the choice of signs for its columns, but the matrix Q_1^* is not, since any particular choice of basis may be rotated by postmultiplying by a $t \times t$ orthogonal matrix.

It was shown in [11] that a necessary condition for x^* to minimize $\phi(x)$ is that there exist a t by t symmetric matrix V^*, with V^* positive semi-definite, such that

$$\text{tr } V^* = 1, \quad V^* \bullet \{Q_1^*\}^T[A_x(x^*)\Delta x]Q_1^* = 0, \tag{3.4}$$

for all Δx. In the case $t = 1$, when Q_1^* consists of a single column q_1^*, this reduces to the statement that $\{q_1^*\}^T[A_x(x^*)\Delta x]q_1^* = 0$, equivalently $[\phi_x(x^*)\Delta x] = 0$ for all Δx, i.e. the gradient of $\phi(x^*)$ is zero. If $A(x)$ is an affine function, the necessary condition is also sufficient for optimality.

We wish to consider the correct local formulation of a Newton-based method so that quadratic convergence to x^* is obtained generically. We assume that the optimal multiplicity t is known. This is not the case in practice, and must be determined during the course of the computation, as explained in [10,11]. If t is set incorrectly, the method to be described would converge locally to a minimizer of ϕ subject to the wrong multiplicity constraint, which might not be a minimizer of ϕ. This can be avoided, by computing an approximation to V^* and verifying that the necessary conditions for optimality, including the positive semi-definite condition on V^*, are satisfied. See [11] for discussion of the case where all optimality conditions except the positive semi-definite condition are satisfied.

Assuming, then, that the optimal value of t is known, the local minimizer x^* of ϕ clearly also locally solves the constrained problem

$$\min_{x,\omega} \quad \omega \tag{3.5}$$

$$\text{subject to} A(x) \in \Omega(t,\omega) \tag{3.6}$$

where $x \in \Re^m$, ω is a real parameter, and $\Omega(t,\omega)$ is the *set of matrices whose greatest eigenvalue has multiplicity t and value ω*. The set $\Omega(t,\omega)$ is an analytic manifold contained in the space of n by n symmetric matrices. The structure of this manifold is well known. It was observed as early as 1929 [17] that the number of conditions imposed on the space of symmetric matrices by the restriction that a matrix lie on this manifold is $\frac{t(t+1)}{2}$. In other words, the codimension of the manifold $\Omega(t,\omega)$ is $\frac{t(t+1)}{2}$. Formulas for the tangent space to the manifold $\Omega(t,\omega)$ at any point can be computed using standard techniques in differential geometry [13,15]. Much less obvious, however, is how to parameterize a description of the manifold which is suitable for the application of Newton methods. This is really the main point of the paper.

The key idea, following [4], is to parameterize the orthogonal matrix of eigenvectors using a *matrix exponential*. Any orthogonal matrix P with $\det P = 1$ can be represented by

$$P = e^Y = I + Y + \tfrac{1}{2}Y^2 + \cdots,$$

where Y is skew-symmetric, i.e. $Y = -Y^T$. Since eigenvector signs are arbitrary, the assumption that $\det P = 1$ is not a restriction. A proof that this representation is always possible and locally unique is given in the Appendix.

Let $\widehat{x}$ be a given point, with the eigenvalues and eigenvectors of $A(\widehat{x})$ given by (1.1)–(1.2). Let

$$\widehat{\Lambda}_1 = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_t), \quad \widehat{\Lambda}_2 = \mathrm{Diag}(\widehat{\lambda}_{t+1}, \ldots, \widehat{\lambda}_n), \tag{3.7}$$

and let

$$\widehat{Q}_1 = [\widehat{q}_1 \ldots \widehat{q}_t], \quad \widehat{Q}_2 = [\widehat{q}_{t+1} \ldots \widehat{q}_n]. \tag{3.8}$$

Define the twice continuously differentiable $n \times n$ symmetric matrix-valued function

$$\widehat{F}(x, Y, \omega, \Theta) = \begin{bmatrix} \omega I & 0 \\ 0 & \Theta \end{bmatrix} - e^{-Y}\widehat{Q}^T A(x)\widehat{Q}e^Y, \tag{3.9}$$

where $x \in \Re^m$, ω is a real scalar, $\Theta = \mathrm{Diag}(\theta_1, \ldots, \theta_{n-t})$ is a real diagonal matrix of order $n - t$, and Y is a real $n \times n$ skew-symmetric matrix. From the context, it is clear that I is used to mean the identity matrix of order t. Subsequent block matrices will have dimensions conforming with those of $\widehat{F}$. We shall find it useful to write

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ -{Y_{12}}^T & Y_{22} \end{bmatrix}, \tag{3.10}$$

where Y_{11} and Y_{22} are skew-symmetric but Y_{12} is not. Note that the definition of $\widehat{F}$ depends on $\widehat{x}$ through $\widehat{Q}$. Of course, $\widehat{Q}$ could be removed from $\widehat{F}$ by absorbing it into e^Y. The reason for the explicit inclusion of $\widehat{Q}$ in the definition of $\widehat{F}$ is so that the function e^Y can always be expanded about $Y = 0$.

Now consider the nonlinear program

$$\min_{x,Y,\omega,\Theta} \quad \omega \tag{3.11}$$

$$\text{subject to} \widehat{F}(x,Y,\omega,\Theta) = 0. \tag{3.12}$$

It is clear that if $\{x,Y,\omega,\Theta\}$ solves (3.12), with $\omega > \theta_i$, $i = 1,\ldots,n-t$, then $\{x,\omega\}$ satisfies the constraint (3.6), with $A(x)$ having eigenvalues $\omega,\ldots,\omega,\theta_1,\ldots\theta_{n-t}$, and eigenvectors given by the columns of $\widehat{Q}e^Y$. Conversely, if x,ω satisfy (3.6), then, regardless of $\widehat{Q}$, (3.12) has a solution $\{x,Y,\omega,\Theta\}$, with $\theta_i = \lambda_{t+i}(x)$ and $e^Y = \widehat{Q}^T Q$, where Q is an orthogonal matrix of eigenvectors for $A(x)$.

The number of equations in (3.12) is $\frac{n(n+1)}{2}$. Formulation (3.11)–(3.12) introduces additional variables Y,Θ which are not present in (3.5)–(3.6), with corresponding space dimension $\frac{n(n-1)}{2} + n - t = \frac{n(n+1)}{2} - t$. The difference between the number of equations and number of extra variables is t, which is *not the codimension of* $\Omega(t,\omega)$. This shows that there is a difficulty with regularity in the parameterization of $\Omega(t,\omega)$ given by (3.12).

This difficulty is clarified by a key observation. Consider

$$F^*(x,Y,\omega,\Theta) = \begin{bmatrix} \omega I & 0 \\ 0 & \Theta \end{bmatrix} - e^{-Y}(Q^*)^T A(x) Q^* e^Y$$

and the associated nonlinear program

$$\min_{x,Y,\omega,\Theta} \quad \omega \tag{3.13}$$

$$\text{subject to} F^*(x,Y,\omega,\Theta) = 0. \tag{3.14}$$

The functions $\widehat{F}$ and F^* coincide if $\widehat{x} = x^*$ and the same basis $\widehat{Q} = Q^*$ is used in both definitions. We have $A(x^*) \in \Omega(t,\lambda_1^*)$ and

$$(Q^*)^T A(x^*) Q^* = \begin{bmatrix} \lambda_1^* I & 0 \\ 0 & \Lambda_2^* \end{bmatrix}$$

so Y satisfying (3.14) is not unique if $t > 1$. Specifically, any Y of the form

$$Y = \begin{bmatrix} Y_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

solves (3.14) with $x = x^*$, $\omega = \lambda_1^*$, $\Theta = \Lambda_2^*$. Consequently, to obtain regularity in (3.13)–(3.14), the additional condition

$$Y_{11} = 0 \tag{3.15}$$

should be imposed in (3.10). The number of equations in (3.14) reduced by the dimension of the space of variables Y, Θ is then

$$\frac{n(n+1)}{2} - \left(\frac{n(n-1)}{2} - \frac{t(t-1)}{2}\right) - (n - t) = \frac{t(t+1)}{2},$$

which *is* the codimension of $\Omega(t,\omega)$. Ideally then, we would like to parameterize (3.5)–(3.6), not by (3.11)–(3.12), but by (3.13)–(3.14) together with (3.10), (3.15). However, this is not possible in practice, because Q^* is *known only in the limit*. The

best we can do is to use (3.11)–(3.12), where $\widehat{Q}$ is the matrix of eigenvectors for $\widehat{x}$, the current best approximation to the solution x^*. Thus, we shall work with *a different function $\widehat{F}$ at each step of the iteration.*

But now a second key point must be emphasized. Although the Y_{11} variables are redundant in (3.14), they are *not* redundant in (3.12) if $\widehat{x} \neq x^*$, or more specifically if $A(\widehat{x}) \notin \Omega(t, \widehat{\lambda}_1)$. On the contrary, the *freedom in Y_{11} is necessary* to ensure that a feasible solution to (3.12) exists in general. Clearly, the closer $\widehat{x}$ is to x^*, i.e. the closer $A(\widehat{x})$ is to $\Omega(t, \widehat{\lambda}_1)$, the closer the Y_{11} variables come to being redundant. This observation is quantified by the following theorem, which follows directly from [4], Corollary 3.1 and subsequent remarks. It will be convenient to denote the variables $\{x, Y, \omega, \Theta\}$ collectively by a single variable Z, which lies in a space of dimension $\frac{n(n+1)}{2} + m + 1 - t$.

THEOREM 1. *There exist $\epsilon > 0$, $C < \infty$ such that, if $\|\widehat{x} - x^*\| \leq \epsilon$, then $\widehat{F}(Z) = 0$ has a solution $\widehat{Z}^* = \{x^*, \widehat{Y}^*, \lambda_1^*, \Lambda_2^*\}$ with*

$$\|\widehat{Y}^*\| \leq C\|\widehat{x} - x^*\|$$

and with the leading t by t block of $\widehat{Y}^*$ satisfying

$$\|\widehat{Y}_{11}^*\| \leq C\|\widehat{x} - x^*\|^2.$$

Here $\widehat{Y}^*$ and $\widehat{Z}^*$ are so denoted because, unlike x^*, they depend on the choice of function $\widehat{F}$.

Roughly speaking, the Y variables describe the rotation of the eigenvectors $\widehat{q}_i$ needed to transform them to eigenvectors of $A(x^*)$, while Y_{11} describes the rotation of the first t of these eigenvectors within the t-dimensional space they span. The rotation of the latter kind becomes relatively unimportant, as $\widehat{x} \to x^*$, because of the nonuniqueness of the eigenvectors of $A(x^*)$.

Straightforward application of Newton's method to solve (3.11)–(3.12) is not satisfactory, since inclusion of the Y_{11} variables, which are redundant in the limit, prevents rapid convergence. On the other hand, setting $Y_{11} = 0$ in (3.11)–(3.12) makes (3.12) infeasible in general. We shall see that the solution to these difficulties is to remove Y_{11} from each linearization step, but include Y_{11} in the convergence analysis of this procedure. Thus, our convergence analysis is nonstandard.

Let us calculate the derivatives of $\widehat{F}$. The appearance of the matrix exponential function in the definition makes this an easy task. We obtain

$$[\widehat{F}_x \Delta x] = -e^{-Y}\widehat{Q}^T[A_x(x)\Delta x]\widehat{Q}e^Y; \tag{3.16}$$

$$[\widehat{F}_Y \Delta Y] = -B - B^T, \quad \text{where} \tag{3.17}$$

$$B = \{-\Delta Y + \tfrac{1}{2}\{\Delta Y\}Y + \tfrac{1}{2}Y\{\Delta Y\} + O(Y^2)\}\widehat{Q}^T A(x)\widehat{Q}\{I + Y + O(Y^2)\};$$

$$[\widehat{F}_\omega \Delta\omega] = \begin{bmatrix} \Delta\omega\ I & 0 \\ 0 & 0 \end{bmatrix}; \tag{3.18}$$

$$[\widehat{F}_\Theta \Delta\Theta] = \begin{bmatrix} 0 & 0 \\ 0 & \Delta\Theta \end{bmatrix}. \tag{3.19}$$

Here $\Delta x, \Delta Y, \Delta\omega, \Delta\Theta$ are variables with the same dimensions as x, Y, ω, Θ, respectively; for example ΔY, like Y, is an n by n skew-symmetric matrix, with

$$\Delta Y = \begin{bmatrix} \{\Delta Y\}_{11} & \{\Delta Y\}_{12} \\ -\{\Delta Y\}_{12}^T & \{\Delta Y\}_{22} \end{bmatrix}, \tag{3.20}$$

where ΔY_{11} and ΔY_{22} are skew-symmetric (but ΔY_{12} is not). We shall use ΔZ to denote $\{\Delta x, \Delta Y, \Delta\omega, \Delta\Theta\}$.

Now let us evaluate $\widehat{F}$ and its derivatives $\widehat{F}_x$, $\widehat{F}_Y$ at the point

$$\widehat{Z} = \{\widehat{x}, \widehat{Y}, \widehat{\lambda}_1, \widehat{\Lambda}_2\}, \tag{3.21}$$

where

$$\widehat{Y} = 0,$$

this equation being essential to keep the formulas simple. The derivatives $\widehat{F}_\omega$ and $\widehat{F}_\Theta$ are constant. We have

$$\widehat{F}(\widehat{Z}) = \begin{bmatrix} \widehat{\lambda}_1 I - \widehat{\Lambda}_1 & 0 \\ 0 & 0 \end{bmatrix}; \tag{3.22}$$

$$[\widehat{F}_x(\widehat{Z})\Delta x] = -\widehat{Q}^T [A_x(\widehat{x})\Delta x]\widehat{Q}; \tag{3.23}$$

$$[\widehat{F}_Y(\widehat{Z})\Delta Y] = -\widehat{\Lambda}\{\Delta Y\} + \{\Delta Y\}\widehat{\Lambda} = \tag{3.24}$$

$$\begin{bmatrix} -\widehat{\Lambda}_1\{\Delta Y\}_{11} + \{\Delta Y\}_{11}\widehat{\Lambda}_1 & -\widehat{\Lambda}_1\{\Delta Y\}_{12} + \{\Delta Y\}_{12}\widehat{\Lambda}_2 \\ \widehat{\Lambda}_2\{\Delta Y\}_{12}^T - \{\Delta Y\}_{12}^T\widehat{\Lambda}_1 & -\widehat{\Lambda}_2\{\Delta Y\}_{22} + \{\Delta Y\}_{22}\widehat{\Lambda}_2 \end{bmatrix}.$$

Notice that the leading t by t block of this matrix is zero if and only if $\widehat{\Lambda}_1$ is a multiple of the identity matrix, i.e. $A(\widehat{x}) \in \Omega(t, \widehat{\lambda}_1)$.

An immediate consequence of Theorem 1 which we shall need later is

$$\|\widehat{Z} - \widehat{Z}^*\| = O(\|\widehat{x} - x^*\|), \tag{3.25}$$

using (3.21) and the Lipschitz continuity of the eigenvalues.

The rest of the paper is organized as follows. In the next section, we analyze the special case $\frac{t(t+1)}{2} = m + 1$, when the dimension of the variable space matches the number of conditions imposed by the multiple eigenvalue, and hence quadratic convergence to a local solution of (1.5) can be achieved by a method which only uses first derivative information. In the subsequent section, we consider the general case, where second derivative information is necessary.

4. A Special Case. In this section we assume $\frac{t(t+1)}{2} = m+1$, where t, as before, is the multiplicity of λ_1^*. This is the case when the number of variables equals the number of conditions imposed by the multiple eigenvalue, and hence x^* is a locally unique solution of (3.14), given a nonsingularity condition to be defined shortly. Consider the following iteration.

ITERATION 1. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by (1.1)–(1.2), and* $\widehat{F}$ *by (3.9). Let* $\widehat{Z} = \{\widehat{x}, 0, \widehat{\lambda}_1, \widehat{\Lambda}_2\}$.
2. *Solve the* n *by* n *symmetric matrix equation*

$$[\widehat{F}_Z(\widehat{Z})\Delta Z] = -\widehat{F}(\widehat{Z}) \tag{4.1}$$

for ΔZ*, imposing also the condition*

$$\{\Delta Y\}_{11} = 0. \tag{4.2}$$

Set $\overline{Z} = \widehat{Z} + \Delta Z$.

3. *Replace* $\widehat{x}$ *by* $\overline{x}$*, the* x *component of* $\overline{Z}$*. Go to Step 1.*

Iteration 1 consists of a *Newton iteration applied to a varying function*, since the function which is differentiated, $\widehat{F}$, changes at each step. Such a situation is not unusual; see [5,16]. The linear system (4.1) is equivalent to

$$[\widehat{F}_x(\widehat{Z})\Delta x] + [\widehat{F}_Y(\widehat{Z})\Delta Y] + [\widehat{F}_\omega(\widehat{Z})\Delta\omega] + [\widehat{F}_\Theta(\widehat{Z})\Delta\Theta] = -\widehat{F}(\widehat{Z}). \tag{4.3}$$

Because of the assumption that $\frac{t(t+1)}{2} = m+1$, together with the fact that Y_{11} is constrained to be zero, this is a system of $\frac{n(n+1)}{2}$ equations in the same number of variables. Examining (3.18)–(3.24), we see that it separates very conveniently. Imposing the condition $\{\Delta Y\}_{11} = 0$, the 1,1 block of (4.3) reduces to the t by t symmetric matrix equation

$$\Delta\omega\ I - \widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 = \widehat{\Lambda}_1 - \widehat{\lambda}_1 I. \tag{4.4}$$

Let us denote this system of linear equations by

$$\widehat{K}\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix} = \widehat{b} \tag{4.5}$$

where

$$\widehat{K} = [\text{vec } I, \ -\text{vec } (\widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_1}\widehat{Q}_1), \ldots, -\text{vec } (\widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_m}\widehat{Q}_1)] \tag{4.6}$$

and

$$\widehat{b} = \text{vec } (\widehat{\Lambda}_1 - \widehat{\lambda}_1 I). \tag{4.7}$$

Note that $\widehat{K}$ has dimension $t(t+1)/2$ by $m+1$, i.e. it is square under the assumptions of this section. (The operator "vec" was defined at the end of Section 2.)

The 1,2 block of (4.3) is the t by $n-t$ matrix equation

$$-\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 - \widehat{\Lambda}_1\{\Delta Y\}_{12} + \{\Delta Y\}_{12}\widehat{\Lambda}_2 = 0 \tag{4.8}$$

which can be solved for $\{\Delta Y\}_{12}$ in terms of Δx by

$$\Delta y_{ij} = \frac{\widehat{q}_i^T[A_x(\widehat{x})\Delta x]\widehat{q}_j}{\widehat{\lambda}_j - \widehat{\lambda}_i}, \tag{4.9}$$

for $1 \leq i \leq t$, $t < j \leq n$; the denominator is bounded away from zero for $\widehat{x}$ in a small enough neighborhood of x^*. The 2,1 block of (4.3) contains the same information as the 1,2 block. The 2,2 block of (4.3) is

$$\Delta\Theta - \widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 - \widehat{\Lambda}_2\{\Delta Y\}_{22} + \{\Delta Y\}_{22}\widehat{\Lambda}_2 = 0. \tag{4.10}$$

The off-diagonal equations of this symmetric system can be solved for $\{\Delta Y\}_{22}$ in a manner similar to equation (4.9), while the diagonal equations, which vanish in the last two terms, can be solved for $\Delta\Theta$.

In fact, though, we see that each step of Iteration 1 actually requires solving *only* one linear system for $\Delta\omega$ and Δx, namely (4.5), a system of $\frac{t(t+1)}{2}$ linear equations in $m+1$ variables and therefore square by assumption. The variables ΔY and $\Delta\Theta$ are *not required* to continue with the next iteration; their only purpose is their use in the problem formulation and convergence analysis. Iteration 1 is therefore equivalent to:

ITERATION 2. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by (1.1)–(1.2).*
2. *Solve the linear system* $\widehat{K}\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix} = \widehat{b}$*, defined in (4.6)–(4.7), for* $\Delta\omega$, Δx.

Set $\overline{x} = \widehat{x} + \Delta x$.

3. *Replace* $\widehat{x}$ *by* $\overline{x}$*, and go to Step 1.*

Let us now analyze the rate of convergence of Iteration 1, equivalently Iteration 2. We first need:

THEOREM 2. *Define*

$$K^* = [\text{vec } I, \quad -\text{vec } (Q_1^{*T}\frac{\partial A(x^*)}{\partial x_1}Q_1^*), \ldots, -\text{vec } (Q_1^{*T}\frac{\partial A(x^*)}{\partial x_m}Q_1^*)]. \tag{4.11}$$

Then the smallest singular value of K^* *is independent of the choice of basis* Q_1^*.

Proof. The freedom in Q_1^* is that it may be postmultiplied by any t by t orthogonal matrix. The smallest singular value of K^* is, by definition,

$$\min_{\Delta\omega^2 + \|\Delta x\|^2 = 1} \|K^*\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix}\|. \tag{4.12}$$

The vector norm being minimized is in fact

$$\|\Delta\omega\ I - \{Q_1^*\}^T[A_x(x^*)\Delta x]Q_1^*\|_F$$

(see the discussion at the end of Section 2). This quantity is not changed if Q_1^* is postmultiplied by an orthogonal matrix. □

Using this result, we can speak unambiguously about whether or not K^* is singular. The convergence result may now be stated.

THEOREM 3. *Suppose* K^* *is nonsingular. Then there exist constants* ϵ *and* C *such that, if* $\|\widehat{x} - x^*\| \leq \epsilon$*, then*

$$\|\overline{x} - x^*\| \leq C\|\widehat{x} - x^*\|^2.$$

Consequently, Iteration 1, equivalently Iteration 2, generates points $\widehat{x}$ *which converge quadratically to the solution* x^*.

Proof. That Iterations 1 and 2 generate the same point $\widehat{x}$ follows from the equivalence of (4.1)–(4.2) with (4.5), (4.8),(4.10). Expanding $\widehat{F}$ in a Taylor series about $\widehat{Z}$, using the point $\widehat{Z}^*$ whose existence is guaranteed by Theorem 1, gives

$$0 = \widehat{F}(\widehat{Z}^*) = \widehat{F}(\widehat{Z}) + [\widehat{F}_Z(\widehat{Z})\{\widehat{Z}^* - \widehat{Z}\}] + O(\|\widehat{Z} - \widehat{Z}^*\|^2). \tag{4.13}$$

By definition of Iteration 1, we also have

$$0 = \widehat{F}(\widehat{Z}) + [\widehat{F}_Z(\widehat{Z})\{\overline{Z} - \widehat{Z}\}], \tag{4.14}$$

noting that the Y_{11} component of $\overline{Z}$ is zero. The difference of these two equations gives

$$[\widehat{F}_Z(\widehat{Z})\{\overline{Z} - \widehat{Z}^*\}] = O(\|\widehat{Z} - \widehat{Z}^*\|^2). \tag{4.15}$$

Some comments here will be helpful. As usual, the proof of convergence of Newton's method involves three points: the current iterate, the new iterate, and the solution point. Here, these are respectively $\widehat{Z}$, $\overline{Z}$ and $\widehat{Z}^*$, the subtlety being that $\widehat{Z}^*$ is the solution to $\widehat{F}(Z) = 0$, an equation whose definition depends on $\widehat{Z}$. Equation (4.15) states that

$$[\widehat{F}_x\{\overline{x} - x^*\}] +$$

$$[\widehat{F}_{\{Y_{11}\}}\{-\widehat{Y}^*_{11}\}] + [\widehat{F}_{\{Y_{12}\}}\{\{\Delta Y\}_{12} - \widehat{Y}^*_{12}\}] + [\widehat{F}_{\{Y_{22}\}}\{\{\Delta Y\}_{22} - \widehat{Y}^*_{22}\}] +$$

$$[\widehat{F}_\omega\{\widehat{\lambda}_1 + \Delta\omega - \lambda^*_1\}] + [\widehat{F}_\Theta\{\widehat{\Lambda}_2 + \Delta\Theta - \Lambda^*_2\}] = O(\|\widehat{x} - x^*\|^2), \tag{4.16}$$

all of the derivatives being evaluated at $\widehat{Z}$, the appearance of $O(\|\widehat{x} - x^*\|^2)$ instead of $O(\|\widehat{Z} - \widehat{Z}^*\|^2)$ on the right-hand side being justified by (3.25). By Theorem 1, the $\widehat{F}_{\{Y_{11}\}}$ term on the left-hand side can be absorbed into the right-hand side, reducing (4.16) to a linear system of $\frac{n(n+1)}{2}$ equations in $\frac{n(n+1)}{2}$ variables. By precisely the argument which showed the equivalence of (4.1)–(4.2) with (4.5),(4.8),(4.10), this system can be reduced to $\frac{t(t+1)}{2}$ equations in $\frac{t(t+1)}{2}$ unknowns, namely

$$\widehat{K}\begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda^*_1 \\ \overline{x} - x^* \end{bmatrix} = O(\|\widehat{x} - x^*\|^2). \tag{4.17}$$

The proof is then complete if we can assert that the norm of the inverse of $\widehat{K}$ is bounded for $\widehat{x}$ in a neighborhood of x^*. Theorem 1 shows that there is an orthonormal basis of eigenvectors for $A(x^*)$, namely $Q^* = \widehat{Q}e^{\widehat{Y}^*}$, for which

$$\|\widehat{Q} - Q^*\| = \|\widehat{Q}^T(\widehat{Q} - Q^*)\| = \|I - e^{\widehat{Y}^*}\| = O(\|\widehat{Y}^*\|) = O(\|\widehat{x} - x^*\|). \tag{4.18}$$

Using this choice of Q^* in (4.11), we have

$$\|\widehat{K} - K^*\| = O(\|\widehat{x} - x^*\|). \tag{4.19}$$

Since K^* is nonsingular by assumption, and this nonsingularity is independent of the basis choice, the boundedness of the inverse of $\widehat{K}$ follows from the standard Banach lemma. ∎

Note that the use of the notation $O(\|\cdot\|^2)$ to denote neglected terms in the Taylor expansion is valid even though a family of functions $\widehat{F}$ is being considered, for a sequence of values $\widehat{Q}$ defining $\widehat{F}$. This is because the definition of $\widehat{F}$ in (3.9) shows that second and higher derivatives cannot blow up regardless of $\widehat{Q}$, given the corresponding smoothness assumptions on the matrix function $A(x)$, together with the orthogonality of $\widehat{Q}$.

5. The General Case. In this section we assume that $\frac{t(t+1)}{2} \leq m+1$. Since the codimension of $\Omega(t,\omega)$ is $\frac{t(t+1)}{2}$, and the dimension of the x,ω variable space is $m+1$, the opposite inequality can hold only nongenerically. Equality can be expected to hold only occasionally since relatively few of the integers have the form $\frac{t(t+1)}{2}$. In the general case, the constraints (3.12) are not enough to define x^* locally, so minimization of (3.11) must also be considered.

Define the Lagrangian function for (3.11)–(3.12) by

$$\widehat{L}(Z,U) = \omega - U \bullet \widehat{F}(Z) \tag{5.1}$$

where U is an $n \times n$ symmetric matrix of Lagrange multipliers corresponding to the $n \times n$ symmetric matrix constraint (3.12). The matrix U is called the *dual matrix* since its components are dual variables. The Frobenius inner product $A \bullet B$ was defined at the end of Section 2. Assuming a full rank condition to be discussed in detail later, the first-order necessary conditions for Z to minimize (3.11) subject to (3.12) are that, in addition to the satisfaction of (3.12) by Z, there exists U satisfying

$$\widehat{L}_Z(Z,U) = 0, \tag{5.2}$$

i.e.

$$U \bullet \widehat{F}_x(Z) = 0, \tag{5.3}$$

$$U \bullet \widehat{F}_Y(Z) = 0, \tag{5.4}$$

$$U \bullet \widehat{F}_\omega = 1, \tag{5.5}$$

and

$$U \bullet \widehat{F}_\Theta = 0. \tag{5.6}$$

Here (5.3), for example, is understood to mean $U \bullet [\widehat{F}_x(Z)\Delta x] = 0$ for all Δx, i.e. $U \bullet \frac{\partial \widehat{F}(Z)}{\partial x_k} = 0$, $1 \leq k \leq m$. A pair Z,U which satisfies conditions (5.3)–(5.6) is denoted $\widehat{Z}^*, \widehat{U}^*$.

In the following Newton iteration we shall, as in the previous section, impose the additional condition that $\{\Delta Y\}_{11} = 0$, and we shall therefore also *relax* the corresponding dual condition $U \bullet \widehat{F}_{\{Y_{11}\}}(Z) = 0$, replacing (5.4) by

$$U \bullet \widehat{F}_{\{Y_{12}\}}(Z) = 0, \qquad U \bullet \widehat{F}_{\{Y_{22}\}}(Z) = 0. \tag{5.7}$$

Each step of the iteration requires a dual matrix *estimate* $\widehat{U}$, which is necessary to define the Lagrangian function. It is important to note that a dual matrix estimate from the *previous* step of the iteration *cannot* be used, since the function $\widehat{F}$ changes from one iteration to the next, with the basis $\widehat{Q}$, which defines $\widehat{F}$, not converging in general.

ITERATION 3. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by (1.1)–(1.2), and* $\widehat{F}$ *by (3.9). Let* $\widehat{Z} = \{\widehat{x}, 0, \widehat{\lambda}_1, \widehat{\Lambda}_2\}$.

2. *Define* $\widehat{U}$ *to be any* $n \times n$ *symmetric matrix such that the norm of the residual of equations (5.3),(5.7),(5.5),(5.6), with* $Z = \widehat{Z}$, $U = \widehat{U}$, *is* $O(\|\widehat{Z} - \widehat{Z}^*\|)$.

3. *Solve the quadratic program*

$$\min_{\Delta Z}[\widehat{L}_Z(\widehat{Z},\widehat{U})\Delta Z] + \tfrac{1}{2}[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z\Delta Z] \tag{5.8}$$

$$\text{subject to} \qquad [\widehat{F}_Z(\widehat{Z})\Delta Z] = -\widehat{F}(\widehat{Z}) \tag{5.9}$$

with the restriction also that

$$\{\Delta Y\}_{11} = 0. \tag{5.10}$$

Set $\overline{Z} = \widehat{Z} + \Delta Z$.

4. *Replace* $\widehat{x}$ *by* $\overline{x}$*, the* x *component of* $\overline{Z}$*. Go to Step 1.*

Like Iteration 1, Iteration 3 can be substantially simplified using the structure of the problem. We begin with a closer look at the dual matrix. Suppose we choose

$$\widehat{U} = \begin{bmatrix} \widehat{U}_{11} & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.11}$$

and consider (5.3)–(5.6) with $Z = \widehat{Z}$, $U = \widehat{U}$. We see then that, for $U = \widehat{U}$, (3.19) implies (5.6) and (3.24) implies (5.4). In order to satisfy the condition in Step 2, then, we see from (3.18) and (3.23) that we need only ensure that

$$\text{tr } \widehat{U}_{11} = 1 + O(\|\widehat{x} - x^*\|) \tag{5.12}$$

and

$$\widehat{U}_{11} \bullet \widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{Q}_1 = O(\|\widehat{x} - x^*\|), \quad 1 \le k \le m. \tag{5.13}$$

This is a system of $m+1$ equations in $\frac{t(t+1)}{2}$ unknowns, which can also be written

$$\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} = e_1 + O(\|\widehat{x} - x^*\|). \tag{5.14}$$

As we shall see in Theorem 6 below, this can be achieved by solving the least squares problem

$$\min_{\widehat{U}_{11}} \|\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} - e_1\|. \tag{5.15}$$

The constraints (5.9)–(5.10) are identical to the condition in Step 2 of Iteration 1, the only difference being that the system of linear equations is underdetermined rather than square. The same argument given following Iteration 1 therefore shows that (5.9)–(5.10) is equivalent to the constraint (4.5) on Δx, $\Delta\omega$ together with (4.8),(4.10) defining $\{\Delta Y\}_{12}$, $\{\Delta Y\}_{22}$.

It is instructive to consider the special case $t = 1$ at this point: in this case the max eigenvalue function $\phi(x)$ is differentiable at x^*. Then $\widehat{Q}_1$ consists of a single column $\widehat{q}_1$, $\widehat{U}_{11}$ is a scalar which can be taken to be the number 1, (5.13) states that the gradient of ϕ at $x = \widehat{x}$ is $O(\|\widehat{x} - x^*\|)$, and the constraint (4.5) states that

$$\Delta\omega = [\phi_x \Delta x]. \tag{5.16}$$

Now let us consider the quadratic objective function (5.8). The linear term may be replaced by $\Delta\omega$, since the rest of this term is fixed by the constraint (5.9). To

evaluate the quadratic term in (5.8), we need to calculate the second derivatives of $\widehat{F}$. Clearly, all terms involving ω or Θ are zero. Differentiating (3.16)–(3.17) we obtain

$$\begin{aligned}
[\widehat{F}_{xx}(\widehat{Z})\Delta x\Delta x] &= -\widehat{Q}^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{Q};\\
[\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y] &= [\widehat{F}_{Yx}(\widehat{Z})\Delta Y\Delta x]\\
&= \{\Delta Y\}\widehat{Q}^T[A_x(\widehat{x})\Delta x]\widehat{Q} - \widehat{Q}^T[A_x(\widehat{x})\Delta x]\widehat{Q}\{\Delta Y\};\\
[\widehat{F}_{YY}(\widehat{Z})\Delta Y\Delta Y] &= \Delta Y\{\widehat{\Lambda}\{\Delta Y\} - \{\Delta Y\}\widehat{\Lambda}\} - \{\widehat{\Lambda}\{\Delta Y\} - \{\Delta Y\}\widehat{\Lambda}\}\Delta Y.
\end{aligned}$$

Since $\widehat{U}$ satisfies (5.11), we need only the 1,1 block of each of these terms. Using (5.10) and (3.20), we obtain

$$\begin{aligned}
[\widehat{F}_{xx}(\widehat{Z})\Delta x\Delta x]_{11} &= -\widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{Q}_1;\\
[\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y]_{11} &= \{\Delta Y\}_{12}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 + \widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2\{\Delta Y\}_{12}^T;\\
[\widehat{F}_{YY}(\widehat{Z})\Delta Y\Delta Y]_{11} &= \{\Delta Y\}_{12}\{-\widehat{\Lambda}_2\{\Delta Y\}_{12}^T + \{\Delta Y\}_{12}^T\widehat{\Lambda}_1\} +\\
&\quad \{\widehat{\Lambda}_1\{\Delta Y\}_{12} - \{\Delta Y\}_{12}\widehat{\Lambda}_2\}\{\Delta Y\}_{12}^T.
\end{aligned}$$

But since ΔY must satisfy the constraint (5.9), whose 1,2 block is (4.8), we see that

$$[\widehat{F}_{YY}(\widehat{Z})\Delta Y\Delta Y]_{11} = -[\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y]_{11}. \tag{5.17}$$

We therefore have

$$\begin{aligned}
[\widehat{F}_{ZZ}(\widehat{Z})\Delta Z\Delta Z]_{11} &= [\widehat{F}_{xx}(\widehat{Z})\Delta x\Delta x]_{11} + [\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y]_{11}\\
&\quad +[\widehat{F}_{Yx}(\widehat{Z})\Delta Y\Delta x]_{11} + [\widehat{F}_{YY}(\widehat{Z})\Delta x\Delta Y]_{11}\\
&= -\widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{Q}_1 + \{\Delta Y\}_{12}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1\\
&\quad +\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2\{\Delta Y\}_{12}^T.
\end{aligned}$$

Let us denote the right-hand side of this equation by $-\widehat{M}$; then we see that, under the constraints (5.9)–(5.10),

$$[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z\Delta Z] = \widehat{U}_{11}\bullet\widehat{M}.$$

Using (4.8) we see that the elements of the $t\times t$ matrix $\widehat{M}$ are given by

$$\widehat{M}_{ij} = \widehat{q}_i^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{q}_j + \sum_{s=t+1}^{n}\gamma_{ijs}\widehat{q}_i^T[A_x\Delta x]\widehat{q}_s\ \widehat{q}_j^T[A_x\Delta x]\widehat{q}_s \tag{5.18}$$

where $1\le i\le t$, $1\le j\le t$ and

$$\gamma_{ijs} = \frac{1}{\widehat{\lambda}_i-\widehat{\lambda}_s} + \frac{1}{\widehat{\lambda}_j-\widehat{\lambda}_s} = \frac{2}{\widehat{\lambda}_1-\widehat{\lambda}_s} + O(\|\widehat{x}-x^*\|). \tag{5.19}$$

Writing out the double sums in the square brackets explicitly we see that, under the constraints (5.9)–(5.10),

$$[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z\Delta Z] = \widehat{U}_{11}\bullet\widehat{M} = \{\Delta x\}^T\widehat{W}\{\Delta x\} \tag{5.20}$$

where $\widehat{W}$ is an m by m symmetric matrix whose k,l element satisfies

$$\widehat{W}_{kl} = \widehat{U}_{11}\bullet\widehat{G}^{kl} \tag{5.21}$$

with $\widehat{G}^{kl}$ defined to be the t by t symmetric matrix with elements

$$\{\widehat{G}^{kl}\}_{ij} = \widehat{q}_i^T \frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_l}\widehat{q}_j + \sum_{s=t+1}^{n} \gamma_{ij\,s}\widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{q}_s \;\widehat{q}_j^T \frac{\partial A(\widehat{x})}{\partial x_l}\widehat{q}_s\,. \tag{5.22}$$

Again, the case $t = 1$ is instructive: then, since $\widehat{U}_{11} = 1$, $\widehat{G}^{kl}$ is the scalar quantity (1.4) (with $i = 1$), i.e. the second partial derivative of ϕ at $x = \widehat{x}$, and $\widehat{W}$ is the Hessian matrix of ϕ at $x = \widehat{x}$.

Therefore, Iteration 3, with $\widehat{U}$ satisfying (5.11), reduces to:

ITERATION 4. *Given an initial value $\widehat{x}$:*

1. *Define $\widehat{\Lambda}$, $\widehat{Q}$ by (1.1)–(1.2).*
2. *Define $\widehat{U}_{11}$ by any t by t symmetric matrix such that (5.14) holds.*
3. *Define $\widehat{W}$ by (5.19)–(5.22). Solve the following quadratic program:*

$$\min_{\Delta\omega,\Delta x} \Delta\omega + \tfrac{1}{2}\{\Delta x\}^T \widehat{W}\{\Delta x\} \tag{5.23}$$

$$\text{subject to} \quad \widehat{K}\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix} = \widehat{b} \tag{5.24}$$

where the latter constraint is defined by (4.6)–(4.7). Set $\overline{x} = \widehat{x} + \Delta x$.

4. *Replace $\widehat{x}$ by $\overline{x}$ and go to Step 1.*

In the case $t = 1$, we see from (5.16) that (5.23)–(5.24) reduces to the ordinary Newton iteration

$$\min_{\Delta x}[\phi_x(\widehat{x})\Delta x] + \tfrac{1}{2}[\phi_{xx}(\widehat{x})\Delta x\Delta x].$$

Iteration 4 is the method given by [10], with two exceptions: (i) [10] addresses a slightly different problem, namely minimizing $\max(\lambda_1(x), -\lambda_n(x))$, with A assumed to be an affine matrix function; (ii) the method of [10] substitutes the quantities $2/\{\widehat{\lambda}_1 - \widehat{\lambda}_s\}$ for $\gamma_{ij\,s}$, dropping the last term on the right-hand side of (5.19). With this simplification, the corresponding formulas for (5.18), (5.22) can be written conveniently using matrix notation as

$$\widetilde{M} = \widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{Q}_1 + 2\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 D^{-1}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 \tag{5.25}$$

with $D = \widehat{\lambda}_1 I - \widehat{\Lambda}_2$,

$$\widetilde{G}^{kl} = \widehat{Q}_1^T \frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_l}\widehat{Q}_1 + 2\widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{Q}_2 D^{-1}\widehat{Q}_2^T \frac{\partial A(\widehat{x})}{\partial x_l}\widehat{Q}_1 \tag{5.26}$$

and

$$\widetilde{W}_{kl} = \widehat{U}_{11} \bullet \widetilde{G}^{kl}. \tag{5.27}$$

The use of $\widetilde{W}$ instead of $\widehat{W}$ does not affect the convergence rate of Iteration 4, but the advantage of the latter formula is that it leads to the following observation, due to M.K.H. Fan[2]:

THEOREM 4. *Suppose A is an affine function, i.e. $A_{xx} = 0$. Then if $\widehat{U}_{11}$ is positive semi-definite, $\widetilde{W}$ is also positive semi-definite, regardless of the magnitude of $\widehat{x} - x^*$.*

Proof. Since $A_{xx} = 0$, it is clear that, for any choice of Δx, $\widetilde{M}$ is positive semi-definite. Since $\widehat{U}_{11}$ is positive semi-definite, the inner product $\widehat{U}_{11} \bullet \widetilde{M}$ is nonnegative for all Δx, which is equivalent to the condition $\{\Delta x\}^T \widetilde{W} \{\Delta x\} \geq 0$ for all Δx. □

Clearly, the same result holds if $[A_{xx}(\widehat{x})\Delta x \Delta x]$ is positive semi-definite for all Δx. Furthermore, if $\widehat{x}$ is close enough to x^*, and $\widetilde{W}$ is positive definite, then $\widehat{W}$ is positive definite. However, even if A is affine, $\widehat{W}$ is not positive semi-definite in general. For example, suppose $n = 3$, $t = 2$, and $\widehat{Q} = I$. The condition that $\widehat{M}$ is positive semi-definite then reduces to the condition $\gamma_{113}\gamma_{223} \geq \gamma_{123}^2$, regardless of A_x. Choosing $\widehat{\Lambda} = \mathrm{Diag}(2, 1, 0)$ gives

$$\gamma_{113} = 1, \quad \gamma_{123} = \gamma_{213} = 1.5, \quad \gamma_{223} = 2.$$

so that $\widehat{M}$ is indefinite. Then $\widehat{U}_{11}$ can be chosen positive semi-definite such that (5.20) is negative. However, substituting $2/(\widehat{\lambda}_1 - \widehat{\lambda}_3)$ for γ_{ijs} results in the matrices $\widetilde{M}$ and $\widetilde{W}$, which are positive semi-definite.

The positive semi-definite condition on $\widehat{U}_{11}$ is a natural one, because, as indicated by the next two theorems, $\widehat{U}_{11}$ is an approximation to the matrix V^* given in (3.4). Specifically, note that equation (5.30) defining U_{11}^* in the following theorem is identical to equation (3.4) defining V^*. There is no condition on the definiteness of U_{11}^*, because in the formulation of the nonlinear program (3.13)–(3.14) we *assumed* that the optimal multiplicity t is known; consequently, indefiniteness of U_{11}^* indicates that t *was chosen incorrectly* and hence that x^* does *not* minimize ϕ.

THEOREM 5.

1. *Consider the r by $(m+1)$ matrix K^*, defined by (4.11), where $r = t(t+1)/2$. Then the rth singular value of K^* does not depend on the choice of basis Q_1^*.*

2. *Suppose that the rth singular value of K^* is nonzero, i.e. K^* has linearly independent rows. Consider the nonlinear program (3.13)–(3.15), noting that the latter constraint removes Y_{11} from the variable set. Let*

$$L^*(Z, U) = \omega - U \bullet F^*(Z).$$

A necessary condition for $Z^* = (x^*, 0, \lambda_1^*, \Lambda_2^*)$ to solve (3.13)–(3.15) is that there exists an n by n symmetric matrix U^*, satisfying

$$L_Z^*(Z^*, U^*) = 0. \tag{5.28}$$

Furthermore, U^* is unique, with

$$U^* = \begin{bmatrix} U_{11}^* & 0 \\ 0 & 0 \end{bmatrix}, \tag{5.29}$$

where the t by t block U_{11}^* satisfies

$$\{K^*\}^T \{\mathrm{vec}\ U_{11}^*\} = e_1. \tag{5.30}$$

3. *Define W^* to be the m by m symmetric matrix with elements*

$$W_{kl}^* = U_{11}^* \bullet G^{*kl}$$

where G^{*kl} is the t by t symmetric matrix with elements

$$G^{*kl} = Q_1^{*T} \frac{\partial^2 A(x^*)}{\partial x_k \partial x_l} Q_1^* + 2\{Q_1^*\}^T \frac{\partial A(x^*)}{\partial x_k} Q_2^* \{\lambda_1^* I - \Lambda_2^*\}^{-1} \{Q_2^*\}^T \frac{\partial A(x^*)}{\partial x_l} Q_1^*.$$

Then W^* is independent of the choice of basis Q_1^*.

4. *The null space of K^* is independent of the choice of basis Q_1^*. Consequently, if N^* is a matrix with orthonormal columns spanning the null space of K^*, the eigenvalues of the reduced Hessian matrix*

$$\{N^*\}^T \begin{bmatrix} 0 & 0 \\ 0 & W^* \end{bmatrix} N^* \tag{5.31}$$

are independent of the choice of bases Q_1^*, N^*. (The matrix in the center of this expression has dimension $m+1$ by $m+1$.)

Proof.

1. The rth singular value of K^* can be written

$$\min_{\|S\|_F=1} \|\{K^*\}^T \{\text{vec } S\}\|,$$

where S is a t by t symmetric matrix. (The quantity (4.12) is zero in the general case that K^* has more columns than rows.) The quantity being minimized is

$$(\{\text{tr } S\}^2 + \sum_{k=1}^{m} \{S \bullet \{Q_1^*\}^T \frac{\partial A(x^*)}{\partial x_k} Q_1^*\}^2 \ )^{\frac{1}{2}}.$$

This minimum value is independent of the choice of basis Q_1^*, since any rotation of the basis can be absorbed into S.

2. Let

$$U^* = \begin{bmatrix} U_{11}^* & U_{12}^* \\ \{U_{12}^*\}^T & U_{22}^* \end{bmatrix}.$$

We claim that (5.28) is equivalent to the two conditions (5.29)–(5.30). To see that (5.28) implies (5.29)–(5.30), observe, by analogy with (5.3)–(5.7) and (3.18)–(3.24), that $U^* \bullet F_\Theta^* = 0$ implies the diagonal elements of U_{22}^* are zero, while $U^* \bullet F_{Y_{22}}^*(Z^*) = 0$ and $U^* \bullet F_{Y_{12}}^*(Z^*) = 0$, together with (3.1), imply respectively that the off-diagonal elements of U_{22}^* and all elements of U_{12}^* are zero. The conditions $U^* \bullet F_\omega^* = 1$ and $U^* \bullet F_x^*(Z^*) = 0$ then reduce to (5.30). Conversely, if (5.29)–(5.30) hold, it is easily verified that (5.28) holds. The linear independence of the columns of $\{K^*\}^T$, equivalently the columns of the coefficient matrix of the linear system (5.28), provides a constraint qualification guaranteeing the existence and uniqueness of U^*.

3. Let M^* be defined by (5.25) with $\widehat{x}, \widehat{\Lambda}, \widehat{Q}$ replaced respectively by x^*, Λ^*, Q^*. (This is equivalent to (5.18) in this case since $\lambda_1^* = \cdots = \lambda_t^*$.) When Q_1^* is postmultiplied by a t by t orthogonal matrix P, it has the following effect: the first column of K^* is unchanged and the others are replaced by vec $P^T Q_1^* \frac{\partial A(x^*)}{\partial x_k} Q_1^* P$; the matrix M^* is replaced by $P^T M^* P$; the matrix U_{11}^* is replaced by $P^T U_{11}^* P$. By analogy with (5.20), $\{\Delta x\}^T W^* \{\Delta x\} = U_{11}^* \bullet M^*$ for all $\{\Delta x\}$, so it follows that W^* is independent of the choice of basis Q_1^*.

4. The null space of K^* is

$$\{v : \quad K^* v = 0\}$$

i.e.

$$\{v = (v_0 \ v_1 \dots v_m)^T : \quad v_0 I + \sum_{k=1}^{m} v_k \{Q_1^*\}^T \frac{\partial A(x^*)}{\partial x_k} Q_1^* = 0\},$$

which is unchanged if Q_1^* is postmultiplied by an orthogonal matrix.

□

The previous theorem was concerned only with quantities involving x^* and F^*. In order to prove convergence of Iterations 3 and 4, however, we need to quantify the relationship between $\widehat{U}$ and $\widehat{U}^*$, the latter quantity being the dual matrix associated with the solution of (3.11)–(3.12).

THEOREM 6. *Suppose K^* has linearly independent rows and that $\widehat{x}$ is sufficiently close to x^*. Consider the nonlinear program (3.11)–(3.12), which has no constraint that $Y_{11} = 0$. A necessary condition for $\widehat{Z}^* = (x^*, \widehat{Y}^*, \lambda_1^*, \Lambda_2^*)$ to solve (3.11)–(3.12) is that there exists an n by n symmetric matrix $\widehat{U}^*$ satisfying*

$$\widehat{L}_Z(\widehat{Z}^*, \widehat{U}^*) = 0, \tag{5.32}$$

i.e. (5.3)–(5.6) hold for $Z = \widehat{Z}^*$, $U = \widehat{U}^*$. Furthermore, $\widehat{U}^*$ is unique. Now assume that the discrepancy in (5.3)–(5.6), with $Z = \widehat{Z}$, $U = \widehat{U}$ is $O(\|\widehat{Z} - \widehat{Z}^*\|)$, as required by Iteration 3. Then

$$\|\widehat{U} - \widehat{U}^*\| = O(\|\widehat{Z} - \widehat{Z}^*\|). \tag{5.33}$$

Furthermore, such a matrix $\widehat{U}$ is obtained by using the block structure (5.11) and solving the least squares problem (5.15).

Proof. From Theorem 5, the independence of the rows of K^* and the independence of the columns of the coefficient matrix defining the linear system (5.28) are equivalent. Using (4.18)–(4.19), it follows that if $\|\widehat{x} - x^*\|$ is sufficiently small, the columns of the linear system (5.32) are also independent. (The fact that the columns of the latter system have more rows than the columns of the former, because of the presence of the additional variables Y_{11}, does not affect the linear independence.) This rank condition provides a constraint qualification guaranteeing the existence and uniqueness of $\widehat{U}^*$, satisfying (5.32), i.e.

$$\widehat{U}^* \bullet \widehat{F}_Z(\widehat{Z}^*) = v, \tag{5.34}$$

where v is a vector with one nonzero element, namely 1, in the position corresponding to the variable ω. By definition, $\widehat{U}$ satisfies

$$\widehat{U} \bullet \widehat{F}_Z(\widehat{Z}) = v + O(\|\widehat{Z} - \widehat{Z}^*\|),$$

which has no equations corresponding to Y_{11}. Subtracting this equation from the corresponding equations in (5.34), ignoring the Y_{11} equations in (5.34), and noting that $\widehat{F}_Z$ is Lipschitz, gives

$$\{\widehat{U} - \widehat{U}^*\} \bullet \widehat{F}_Z(\widehat{Z}) = O(\|\widehat{Z} - \widehat{Z}^*\|).$$

The independence of the columns of the coefficient matrix defining this system then gives (5.33).

The proof of the final statement of the theorem is as follows. From (5.30),

$$K^*\{K^*\}^T\{\text{vec } U_{11}^*\} = K^* e_1$$

and, from (5.15),

$$\widehat{K}\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} = \widehat{K}^T e_1.$$

It follows as a consequence, using (4.19) and the fact that K^* is full rank, that

$$\|\widehat{U}_{11} - U_{11}^*\| = O(\|\widehat{x} - x^*\|).$$

Combining this equation with (4.19) and (5.30) gives

$$\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} = e_1 + O(\|\widehat{x} - x^*\|)$$

from which the result follows. □

We are now ready to prove the main convergence theorem.

THEOREM 7. *Suppose that K^* has independent rows and that the reduced Hessian (5.31) is positive definite. Then there exist constants ϵ and C such that, if $\|\widehat{x}-x^*\| \leq \epsilon$, then*

$$\|\overline{x} - x^*\| \leq C\|\widehat{x} - x^*\|^2$$

for both Iterations 3 and 4. Consequently, both iterations generate points $\widehat{x}$ which converge quadratically to the solution x^*.

Proof. From Theorem 6, assuming that $\widehat{x}$ is sufficiently close to x^*, a necessary condition for a pair $\widehat{Z}^*, \widehat{U}^*$ to solve the nonlinear program (3.11)–(3.12) (*without* the condition $Y_{11} = 0$ imposed), is that, in addition to (3.12), the equation (5.32) holds. Theorem 1 shows that we can take the $\widehat{Y}^*$ component of $\widehat{Z}^*$ to satisfy $\|\widehat{Y}^*\| = O(\|\widehat{x} - x^*\|)$ and $\|\widehat{Y}_{11}^*\| = O(\|\widehat{x} - x^*\|^2)$. Furthermore, we can expand $\widehat{F}$ in a Taylor series just as in the proof of Theorem 3, obtaining all of equations (4.13)–(4.16) exactly as before, the only difference being that these equations are not square systems. Specifically, (4.16), with its Y_{11} terms absorbed into the right-hand side, gives

$$[\widehat{F}_Z(\widehat{Z})\{\overline{Z} - \widehat{Z}^*\}] = O(\|\widehat{x} - x^*\|^2). \tag{5.35}$$

Now let us expand (5.32) in a Taylor series. We have

$$0 = \widehat{L}_Z(\widehat{Z}^*, \widehat{U}^*) = \widehat{L}_Z(\widehat{Z}, \widehat{U}) + [\widehat{L}_{ZZ}(\widehat{Z}, \widehat{U})\{\widehat{Z}^* - \widehat{Z}\}] + [\widehat{L}_{ZU}(\widehat{Z}, \widehat{U})\{\widehat{U}^* - \widehat{U}\}]$$

$$+O(\|\widehat{Z} - \widehat{Z}^*\|^2 + \|\widehat{Z} - \widehat{Z}^*\|\|\widehat{U} - \widehat{U}^*\|)$$

using the linearity of $\widehat{L}(Z, U)$ in U. Note that the terms in square brackets, although involving second-order differentiation, are summed over only one argument and are therefore vectors of length $n(n+1)/2 + m + 1 - t$, the number of variables in Z. This system of equations has a row and a column corresponding to each element of $Z = (x, Y, \omega, \Theta)$. Let us *discard* the *rows* corresponding to Y_{11}, and *absorb* the *columns* corresponding to Y_{11} into the O term, which is permissible since $\widehat{Y} = 0$, $\widehat{Y}_{11}^* = O(\|\widehat{x} - x^*\|^2)$. Using the fact that $\widehat{L}_{ZU} = -\widehat{F}_Z$, this gives

$$0 = \widehat{L}_Z(\widehat{Z}, \widehat{U}) + [\widehat{L}_{ZZ}(\widehat{Z}, \widehat{U})\{\widehat{Z}^* - \widehat{Z}\}] - \{\widehat{U}^* - \widehat{U}\} \bullet \widehat{F}_Z(\widehat{Z}) + O(\|\widehat{x} - x^*\|^2) \tag{5.36}$$

with the understanding that all Y_{11} terms are omitted. The $O(\|\widehat{x} - x^*\|^2)$ term on the right-hand side is justified by (3.25) and (5.33).

The necessary condition for a pair $\Delta Z, \Delta U$ to solve the quadratic program defining a step of Iteration 3 is, in addition to the constraints (5.9)–(5.10), that there exists a dual matrix ΔU such that

$$\Delta U \bullet \widehat{F}_Z(\widehat{Z}) = \widehat{L}_Z(\widehat{Z}, \widehat{U}) + [\widehat{L}_{ZZ}(\widehat{Z}, \widehat{U})\Delta Z]. \tag{5.37}$$

where rows and columns of the coefficient matrix corresponding to Y_{11} have been omitted because of (5.10). Noting that $\Delta Z = \overline{Z} - \widehat{Z}$ and subtracting (5.36) from (5.37) gives

$$[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\{\overline{Z}-\widehat{Z}^*\}] - \{\overline{U}-\widehat{U}^*\} \bullet \widehat{F}_Z(\widehat{Z}) = O(\|\widehat{x}-x^*\|^2), \tag{5.38}$$

where $\overline{U} = \widehat{U} + \Delta U$.

Equations (5.35),(5.38) state the first-order optimality conditions for the quadratic program

$$\min_{\overline{Z}-\widehat{Z}^*} \quad h \cdot \{\overline{Z}-\widehat{Z}^*\} + \tfrac{1}{2}[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\{\overline{Z}-\widehat{Z}^*\}\{\overline{Z}-\widehat{Z}^*\}] \tag{5.39}$$

$$\text{subject to} \qquad [\widehat{F}_Z(\widehat{Z})\{\overline{Z}-\widehat{Z}^*\}] = O(\|\widehat{x}-x^*\|^2) \tag{5.40}$$

where the first term in (5.39) is an inner product, with h (which has the same structure as Z) satisfying $h = O(\|\widehat{x}-x^*\|^2)$. It is understood that there are no Y_{11} terms in $\overline{Z}$, $\widehat{Z}^*$. Note that the Hessian and constraint coefficients of this quadratic program are identical to those of (5.8)–(5.9). We shall now simplify this quadratic program, using an argument similar to that which reduced (5.8)–(5.9) to (5.23)–(5.24). First consider the linear term in (5.39). We have

$$h \cdot \{\overline{Z}-\widehat{Z}^*\} = \widetilde{h} \cdot \begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} + \psi \tag{5.41}$$

where $\widetilde{h} \in \Re^{m+1}$ and $\psi \in \Re$ satisfy $\widetilde{h} = O(\|\widehat{x}-x^*\|^2)$ and $\psi = O(\|\widehat{x}-x^*\|^4)$. This equation holds because of the constraint (5.40), which defines the Y and Θ elements of $\overline{Z}-\widehat{Z}^*$ in terms of the x and ω components, by analogy with (4.8)–(4.10). Now consider the quadratic term in (5.39). The argument that showed that the quadratic form in (5.8) reduces to that in (5.23) uses (5.17), which follows from the 1,2 block of (5.9), namely (4.8). We now use a similar argument to simplify the quadratic term in (5.39). Instead of (4.8), we have, from the 1,2 block of (5.40),

$$-\widehat{Q}_1^T[A_x(\widehat{x})\{\overline{x}-x^*\}]\widehat{Q}_2 - \widehat{\Lambda}_1\{\Delta Y - \widehat{Y}^*\}_{12} + \{\Delta Y - \widehat{Y}^*\}_{12}\widehat{\Lambda}_2 = O(\|\widehat{x}-x^*\|^2).$$

Instead of (5.17), we conclude that

$$[\widehat{F}_{YY}(\widehat{Z})\{\Delta Y - \widehat{Y}^*\}\{\Delta Y - \widehat{Y}^*\}]_{11} + [\widehat{F}_{xY}(\widehat{Z})\{\overline{x}-x^*\}\{\Delta Y - \widehat{Y}^*\}]_{11}$$

$$= O(\|\widehat{x}-x^*\|^2\|\Delta Y - \widehat{Y}^*\|).$$

Again using (5.40) to define $\Delta Y - \widehat{Y}^*$ in terms of the x and ω components of $\overline{Z}-\widehat{Z}^*$, we see that the right-hand side consists of two terms, of which one can be absorbed into the first term of (5.41), and the other into the second. We therefore see that, just as the quadratic form in (5.8) reduces to that in (5.23), the quadratic form in (5.39) reduces to

$$\psi + \widetilde{h} \cdot \begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} + \tfrac{1}{2}\{\overline{x}-x^*\}^T\widehat{W}\{\overline{x}-x^*\} \tag{5.42}$$

where $\widetilde{h} = O(\|\widehat{x}-x^*\|^2)$. The constraint (5.40) reduces to (4.17), i.e.

$$\widehat{K}\begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} = O(\|\widehat{x}-x^*\|^2). \tag{5.43}$$

The optimality conditions for the quadratic program defined by (5.42)–(5.43) are

$$\left[\begin{array}{cc} \left[\begin{array}{cc} 0 & 0 \\ 0 & \widehat{W} \end{array}\right] & -\widehat{K}^T \\ \widehat{K} & 0 \end{array}\right] \left[\begin{array}{c} \left[\begin{array}{c} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{array}\right] \\ \text{vec } \{\overline{U}_{11} - \widehat{U}_{11}^*\} \end{array}\right] = O(\|\widehat{x} - x^*\|^2), \tag{5.44}$$

By assumption, K^* has full rank and (5.31) is positive definite, so

$$\left[\begin{array}{cc} \left[\begin{array}{cc} 0 & 0 \\ 0 & W^* \end{array}\right] & \{K^*\}^T \\ K^* & 0 \end{array}\right]$$

is nonsingular. Therefore, using (4.18)–(4.19) and noting that $\|\widehat{W} - W^*\| = O(\|\widehat{x} - x^*\|)$, we see that the inverse of the coefficient matrix of (5.44) is bounded for $\widehat{x}$ near x^*. The desired quadratic contraction is therefore proved. □

6. Concluding Remarks. The convergence proof just given is complicated, because of the disparity in the number of free parameters in the equations $\widehat{F} = 0$ and $F^* = 0$, even as $\widehat{x} \to x^*$. An alternative analysis of the same method has been given recently by Shapiro and Fan [15], in contemporary, independent work. Our results and those of [15] complement each other nicely. The analysis in [15] is shorter than ours but rests on several nontrivial results. The principal idea is that although eigenvectors are not smooth, eigenprojections are differentiable, and indeed derivative formulas are known (Kato [7]). Shapiro and Fan show how to construct a smoothly varying orthonormal basis for the eigenprojection which agrees with a given orthonormal basis of eigenvectors at a point, though not in a neighborhood of the point. Neither the results from Kato nor the construction of the eigenprojection basis could be said to be elementary, though both are powerful. By contrast, our convergence proof is completely self-contained. The Hessian formulas arise simply from differentiating the function $\widehat{F}$ and do not require any machinery from Kato. The only outside result which is needed is Theorem 1, whose proof is elementary [4].

Appendix. The following shows that any real orthogonal matrix P with $\det P = 1$ may be written in the form $P = e^Y$, where $Y = -Y^T$. This derivation was suggested by J.-P. Haeberly. It is undoubtedly well known, though we lack a standard reference.

An orthogonal matrix has eigenvalues of the form ± 1 and $\cos\theta \pm i \sin\theta$, with a corresponding orthogonal set of eigenvectors. Thus, there exists an orthogonal matrix V such that

$$V^T P V = \text{Diag}(D_1, \ldots, D_k)$$

where each D_j is either the number ± 1, or a 2×2 matrix of the form

$$\left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right].$$

Since $\det P = 1$, the number of -1's that occur must be even, so we may assume that the D_j's are either the number $+1$ or a 2×2 matrix as above. But $1 = e^0$, and

$$\left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right] = \exp \left[\begin{array}{cc} 0 & -\theta \\ \theta & 0 \end{array}\right].$$

Hence, $\mathrm{Diag}(D_1, \ldots, D_k) = e^X$ for some block diagonal matrix X with nonzero diagonal blocks of the form

$$\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}.$$

Note that $X = -X^T$. Defining $Y = VXV^T$, we have

$$P = V\mathrm{Diag}(D_1, \ldots, D_k)V^T = Ve^XV^T = e^Y.$$

It remains to show that Y is skew-symmetric:

$$(VXV^T)^T = VX^TV^T = V(-X)V^T = -VXV^T.$$

The matrix Y is not unique, since incrementing θ by multiples of 2π does not change e^Y, but the solution set consists of isolated points in matrix space. In our local convergence analysis, we are concerned only with $P = e^Y$ in a neighborhood of the identity matrix and the corresponding Y in a neighborhood of the zero matrix (see Theorem 3.1).

Acknowledgements. We thank Jean-Pierre Haeberly and the anonymous referees for helping us to improve the discussion in Section 3.

REFERENCES

[1] J. Cullum, W. Donath, and P. Wolfe, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Mathematical Programming Study, 3 (1975), pp. 35–55.

[2] M. Fan, 1988. Private communication.

[3] R. Fletcher, *Semi-definite constraints in optimization*, SIAM Journal on Control and Optimization, 23 (1985), pp. 493–513.

[4] S. Friedland, J. Nocedal, and M. Overton, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM Journal on Numerical Analysis, 24 (1987), pp. 634–667.

[5] J. Goodman, *Newton's method for constrained optimization*, Mathematical Programming, 33 (1985), pp. 162–171.

[6] J.-P. A. Haeberly and M. Overton, *A hybrid algorithm for optimizing eigenvalues of symmetric definite pencils*, SIAM Journal on Matrix Analysis and Applications, 15 (1994), pp. 1141–1156.

[7] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.

[8] P. Lancaster, *On eigenvalues of matrices dependent on a parameter*, Numerische Mathematik, 6 (1964), pp. 377–387.

[9] J. Nocedal and M. Overton, *Numerical methods for solving inverse eigenvalue problems*, in Lecture Notes in Mathematics 1005, V. Pereyra and A. Reinoza, eds., New York, 1983, Springer-Verlag.

[10] M. Overton, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM Journal on Matrix Analysis and Applications, 9 (1988), pp. 256–268.

[11] ———, *Large-scale optimization of eigenvalues*, SIAM Journal on Optimization, 2 (1992), pp. 88–120.

[12] M. Overton and R. Womersley, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Mathematical Programming, 62 (1993), pp. 321–357.

[13] M. Overton and X.-J. Ye, *Towards second-order methods for structured nonsmooth optimization*, in Advances in Optimization and Numerical Analysis, S. Gomez and J.-P. Hennart, eds., The Netherlands, 1994, Kluwer, pp. 97–110.

[14] E. Polak and Y. Wardi, *A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints*, Journal of Structural Mechanics, 11 (1983), pp. 561–577.

[15] A. Shapiro and M. Fan, *On eigenvalue optimization*, SIAM Journal on Optimization. To appear.

[16] R. Tapia, *A stable approach to Newton's method for general mathematical programming problems in R^n*, Journal of Optimization Theory and Applications, 14 (1974), pp. 453–476.

[17] J. von Neumann and E. Wigner, *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen*, Physik. Zeitschr., 30 (1929), pp. 467–470.

[18] G. Watson, *Algorithms for minimum trace factor analysis*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 1039–1053.

[19] ———, *Computing the structured singular value*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 1054–1066.