On Shape Optimizing the Ratio of the First Two Eigenvalues of the Laplacian

J.-P. Haeberly

September 30, 1991

Abstract

We investigate numerically a 1956 conjecture of Payne, Polya, and Weinberger. The conjecture asserts that the ratio of the first two eigenvalues of the Laplacian on a bounded domain Ω of the plane with Dirichlet boundary conditions reaches its minimum value precisely when Ω is a disk. A crucial feature of this problem is the loss of smoothness of the objective function at the solution. The following results form the core of our numerical treatment. First, we construct finite dimensional families of deformations of a disk equipped with a uniform triangulation. This permits the formulation of a discrete model of the problem via finite element techniques. Second, we build on the work of M. Overton to derive optimality conditions in terms of Clarke's generalized gradients for nonsmooth functions. These ideas are then combined into an algorithm and implemented in Fortran.

Contents

1 Introduction

In 1956 a conjecture was formulated by Payne, Polya and Weinberger concerning the ratio of the two smallest eigenvalues of the Laplacian on bounded regions Ω of the plane [18]. More specifically, if $\lambda_1(\Omega) \leq \lambda_2(\Omega)$ denote these eigenvalues, then they conjectured that the ratio, λ_2/λ_1, attains its maximum precisely when Ω is a disk. Recently, while the present work was still in progress, Ashbaugh and Benguria have given a proof that the conjecture is indeed true [1,2].

In this thesis we investigate this problem from a different perspective, namely that of numerically minimizing this ratio for an appropriate discretized model. Optimization problems involving eigenvalues are particularly interesting and challenging because the optimization objective often forces some eigenvalues to coalesce at the solution point and this results in a loss of smoothness. This is precisely what occurs in the above problem since the second eigenvalue of the Laplacian on a disk has multiplicity two. Methods for the optimization of the largest eigenvalue of a symmetric matrix that can handle the lack of smoothness at the solution have recently been devised by Michael Overton [16,15] who implemented them in an algorithm which has been successfully applied to an extensive collection of problems [16,8]. We extend these techniques to apply in our context.

The main steps are as follows. First we use finite elements to describe a family $\Omega(x)$ of perturbations of a disk of radius R. Here x lies in a bounded open neighborhood U of $(1, \ldots, 1)$ in $\Re^m$, and $\Omega(1, \ldots, 1)$ is our approximation of the disk. Computing the eigenvalues of the Laplacian on $\Omega(x)$ is then reduced to computing the eigenvalues of the symmetric definite pencil $(A(x), B(x))$, where $A(x)$ is the stiffness matrix and $B(x)$ is the mass matrix corresponding to the given triangulation of $\Omega(x)$ [19]. Next we extend the work of Overton and Womersley [17] to sums $\sum_{i=1}^{k} \lambda_i$ of the first k eigenvalues of symmetric pencils (A, B). More precisely, we give a characterization of the generalized gradient of $\sum_{i=1}^{k} \lambda_i$ as a function of both A and B. We then apply this result to obtain a description of the generalized gradient $\sum_{i=1}^{k} \lambda_i(x)$, the sum of the first k eigenvalues of $A(x), B(x)$. This, combined with Clarke's calculus of generalized gradients [7], allows us to derive a characterization of the generalized gradient of the function $\rho(x) \equiv -(\lambda_1(x) + \lambda_2(x))/\lambda_1(x)$. This is the function that we wish to minimize. Indeed, minimizing $\rho(x)$ is equivalent to maximizing $\lambda_2(x)/\lambda_1(x)$. We are then able to derive optimality conditions for a minimizer of $\rho(x)$ and to formulate the appropriate variant of Overton's algorithm.

The thesis is organized as follows. We begin with a few preliminary remarks about eigenvalues of symmetric pencils in section 2. Then, in sections 3 and 4, we recall the results of Overton and Womersley [17] mentioned above. In section 5 we show how this leads to a characterization of the generalized gradient of the sum, $\sum_{i=1}^{k} \lambda_i$, of the first k eigenvalues of the symmetric pencil (A, B) by considering the symmetric matrix $G^{-1}AG^{-T}$, where G is the Choleski factor B. Then, composing the function $\sum_{i=1}^{k} \lambda_i$ with the smooth function associating $(A(x), B(x))$ to $x \in U$, we give a descritption of the generalized gradient of $\sum_{i=1}^{k} \lambda_i(x)$ as a function of x in section 6. Next we compute the generalized gradient of the function $\rho(x)$ in section 7, while in section 8 we digress briefly to derive the generalized gradient of the related function $\eta(x) \equiv \lambda_2(x) - \lambda_1(x)$, measuring the gap between λ_2 and λ_1. Optimality conditions for a minimizer of $\rho(x)$ are given in section 9, and the construction of our family of perturbations of a disk is given in section 10. Our modified version of Overton's algorithm is described in section 11. Finally a discussion of the numerical data generated by the algorithm is presented in section 12.

We wish to thank our advisor, Prof. M. Overton, for introducing us to this material and for his help and kindness (and patience!) during the preparation of this thesis.

Notations.
$\Re^{n \times n}$ denotes the vector space of $n \times n$ real matrices.
$\mathcal{S}\Re^n$ is the subspace of symmetric $n \times n$ matrices, $\mathcal{O}^n$ the group of orthogonal $n \times n$ matrices and $\mathcal{Q}^n$ the set of positive definite symmetric matrices.
$\mathcal{S}\Re^n$ is endowed with an inner product, $< \, , \, >$, defined by

$$< M, N >= \sum_{i,j} m_{ij} n_{ij} = Tr\, M^T N,$$

for $M = \{m_{ij}\}$, $N = \{n_{ij}\}$ in $\mathcal{S}\Re^n$.
If A_1 is an $n \times r$ matrix and A_2 is an $n \times s$ matrix, $A = [A_1 : A_2]$ denotes the $n \times (r + s)$ matrix obtained by juxtaposing the columns the A_1 and A_2.

2 Preliminaries

We collect in this section some well known facts for future reference.

First observe that $\mathcal{S}\Re^n \times \mathcal{Q}^n$ is a convex subset of $\mathcal{S}\Re^n \times \mathcal{S}\Re^n$. Indeed it is obvious that

$$M, N \in \mathcal{S}\Re^n \Rightarrow \lambda M + (1-\lambda)N \in \mathcal{S}\Re^n, \forall \lambda \in \Re$$

$$M, N \in \mathcal{Q}^n \Rightarrow \lambda M + (1-\lambda)N \in \mathcal{Q}^n, \forall \lambda \in [0,1].$$

Lemma 2.1 *A symmetric matrix B is positive definite if and only if all its eigenvalues are positive. It is positive semi-definite if and only if all its eigenvalues are nonnegative.*

Proof: Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ denote the eigenvalues of the symmetric matrix B. There exists an orthogonal matrix Q such that

$$Q^T B Q = diag\{\lambda_1, \ldots, \lambda_n\} \equiv \Lambda.$$

The Raleigh characterization of λ_n gives

$$\lambda_n = \min_{x \neq 0} \frac{x^T B x}{x^T x}$$

Then

$$x^T B x > 0, \, \forall x \neq 0 \Longleftrightarrow \lambda_n > 0$$

and

$$x^T B x \geq 0, \, \forall x \Longleftrightarrow \lambda_n \geq 0. \, \Box$$

Lemma 2.2 *Let $V = \{v_{ij}\}$ be a positive semi-definite symmetric matrix. Then we have*

$$(i) \; v_{ii} \geq 0, \, \forall i$$

$$(ii) \; v_{ij}^2 \leq v_{ii} v_{jj}, \, \forall i, j.$$

Proof: (i) This is trivial. Take x to be the vector $e_i = (0, \ldots, 1, \ldots, 0)$ with 1 in the i^{th} position. Then

$$v_{ii} = x^T V x \geq 0.$$

(ii) given $i < j$, consider the vector x with $x_i = \lambda$, $x_j = \mu$ and $x_k = 0$ for all $k \neq i$. Then

$$\begin{aligned} x^T V x &= \sum_{k,l} x_k x_l v_{kl} \\ &= x_i^2 v_{ii} + x_j^2 v_{jj} + 2 x_i x_j v_{ij} \\ &= \lambda^2 v_{ii} + \mu^2 v_{jj} + 2\lambda\mu v_{ij} \end{aligned}$$

Hence we must have

$$\lambda^2 v_{ii} + \mu^2 v_{jj} + 2\lambda\mu v_{ij} \geq 0, \forall\, \lambda,\, \mu \qquad (2.3)$$

Viewing (2.3) as a polynomial in λ with μ fixed, we see that the discriminant must be ≤ 0 . Thus

$$4\mu^2 v_{ij}^2 - 4\mu^2 v_{ii} v_{jj} \leq 0, \forall \mu$$

and the conclusion follows. □

Given a symmetric definite pencil (A, B), the spectrum $\Lambda(A, B)$ may be expressed in several equivalent ways. Let G denote the Choleski factor of B, that is G is a positive definite lower triangular matrix such that $B = GG^T$.

Lemma 2.4 *The following sets are equal.*

1. $\Lambda(A, B)$
2. $\Lambda(B^{-1}A)$
3. $\Lambda(G^{-1}AG^{-T})$

Proof: It is obvious that $\Lambda(A, B) = \Lambda(B^{-1}A)$, so we show that $\Lambda(A, B) = \Lambda(G^{-1}AG^{-T})$. Let $\lambda \in \Lambda(A, B)$ and let x be an eigenvector for λ. Write $y = G^T x$. Then

$$\begin{aligned} Ax = \lambda Bx &\iff Ax = \lambda GG^T x \\ &\iff AG^{-T} y = \lambda G y \\ &\iff G^{-1}AG^{-T} y = \lambda y \end{aligned}$$

Thus we have proved that $\lambda \in \Lambda(A, B)$ with eigenvector x if and only if $\lambda \in \Lambda(G^{-1}AG^{-T})$ with eigenvector $y = G^T x$. □

Remark 2.5 The advantage of viewing $\Lambda(A, B)$ as $\Lambda(G^{-1}AG^{-T})$ rather than as $\Lambda(B^{-1}A)$ is that $G^{-1}AG^{-T}$ is still a symmetric matrix. On the other hand, the dependence upon the data (A, B) is by far more explicit in the matrix $B^{-1}A$ than in $G^{-1}AG^{-T}$.

We can build upon the proof of lemma 2.4 a bit. Suppose $B \in \mathcal{Q}^n$ with Choleski factor G. Let $\mathcal{A}_1$ denote the set of orthonormal bases of $\Re^n$ (i.e. the set of orthonormal matrices), and let $\mathcal{A}_2$ denote the set of B-orthonormal bases of $\Re^n$, i.e. bases $\{v_1, \ldots, v_n\}$ such that

$$v_i^T B v_j = \delta_{ij}, \text{ for } 1 \leq i, j \leq n.$$

Lemma 2.6 *The matrix G induces a bijection of $\mathcal{A}_1$ onto $\mathcal{A}_2$ defined as*

$$\{v_1, \ldots, v_n\} \mapsto \{G^{-T} v_1, \ldots, G^{-T} v_n\}$$

Proof: Given $\{v_1, \ldots, v_n\} \in \mathcal{A}_1$. Then

$$\begin{aligned} (G^{-T} v_i)^T B (G^{-T} v_j) &= v_i^T (G^{-1} B G^{-T}) v_j \\ &= v_i^T v_j \\ &= \delta_{ij}. \ \square \end{aligned}$$

Now let $M \in \mathcal{S}\Re^n$. We recall a few facts about the eigenvalues λ_i of M. Write : $\lambda_1 \geq \ldots \geq \lambda_n$, and let $Q \in \mathcal{Q}^n$ be an orthogonal matrix whose columns $q_1, \ldots, q_n$ form an orthonormal basis of eigenvectors of M, ordered so that

$$Q^T M Q = \Lambda = diag\{\lambda_1, \ldots, \lambda_n\}.$$

For $k > 0$, write

$$H_k \equiv \{v \in \Re^n \,|\, v^T q_i = 0, 1 \leq i \leq k\}$$

and let $H_0 \equiv \Re^n$. Then we have the Raleigh quotient characterization of λ_k :

$$\lambda_k = \max_{v \in H_{k-1}, v \neq 0} \frac{v^T M v}{v^T v}.$$

Next let L_m denote the subspace of $\Re^n$ spanned by a set of m orthonormal vectors $\{v_1, \ldots, v_m\}$. We define the Raleigh trace of M on L_m by

$$Tr[L_m] \equiv \sum_{i=1}^{m} v_i^T M v_i.$$

Observe that $Tr[L_m]$ is independent of the choice of an orthonormal basis of the subspace L_m. Indeed, let $V = [v_1 \ldots v_m]$ so that $V^T V = I$ and

$$Tr[L_m] = Tr(V^T M V).$$

Any other orthonormal basis $\{w_1, \ldots, w_m\}$ of L_m with $W = [w_1 \ldots w_m]$ is such that

$$W = VA,$$

for some orthonormal $m \times m$ matrix A. It follows that

$$Tr(A^T V^T M V A) = Tr(V^T A V) = Tr[L_m].$$

Lemma 2.7 *([3]) We have the following characterization of the sum of the k eigenvalues $\lambda_{m+1}, \ldots, \lambda_{m+k}$:*

$$\sum_{i=m+1}^{m+k} \lambda_i = \max_{L_k \subset H_m} Tr[L_k].$$

Proof: Let $L_k \subset H_m$ be fixed. Then for any $v \in L_k$, we have $v^T q_i = 0$, for $1 \leq i \leq m$. Since $dim\, L_k = k$, there must exist a $v_{k+m} \in L_k$ such that

$$v_{k+m}^T q_i = 0, \; for \; 1 \leq i \leq k + m - 1.$$

Hence $v_{k+m} \in H_{k+m-1}$, and it follows that

$$\lambda_{k+m} \geq v_{k+m}^T M v_{k+m}.$$

Similarly, we construct $v_{m+j} \in L_k \cap H_{m+j-1}$ for $1 \leq j \leq k$ such that

$$v_{m+j}^T v_{m+j} = 1$$

and

$$v_{m+j}^T v_l = 0, \; for \; m + j + 1 \leq l \leq m + k.$$

Thus $\lambda_{m+j} \geq v_{m+j}^T M v_{m+j}$ and it follows that

$$\sum_{i=m+1}^{m+k} \lambda_i \geq Tr[L_k].$$

On the other hand, if we take L_k to be the span of $q_{m+1}, \ldots, q_{m+k}$, then we have

$$\sum_{i=m+1}^{m+k} \lambda_i = Tr[L_k]. \; \Box$$

3 Max characterization of $\sum_{i=1}^{k} \lambda_i$ for symmetric matrices

Let M be an arbitrary element of $\mathcal{S}\Re^n$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. Then lemma 2.7 implies that

$$\textstyle\sum_{i=1}^{k} \lambda_i = max \displaystyle\sum_{j=1}^{k} v_j^T M v_j$$

where the max is taken over all sets of k orthonormal vectors $\{v_1, \ldots, v_k\}$. Now

$$\begin{aligned} \sum_{j=1}^{k} v_j^T M v_j &= \sum_{j=1}^{k} < v_j v_j^T, M > \\ &= < \sum_{j=1}^{k} v_j v_j^T, M > . \end{aligned}$$

Let us write

$$\mathcal{S}_1^k \equiv \{V \in \mathcal{S}\Re^n \Bigg| V = \sum_{j=1}^{k} v_j v_j^T \}$$

where $\{v_1, \ldots, v_k\}$ runs over all orthonormal sets of k vectors.
Then

$$\textstyle\sum_{i=1}^{k} \lambda_i = \displaystyle\max_{V \in \mathcal{S}_1^k} < V, M > . \tag{3.1}$$

We now provide another characterization of $\sum_{i=1}^{k} \lambda_i$. We let

$$\mathcal{S}_2^k \equiv \{U \in \mathcal{S}\Re^n \,|\, 0 \leq U \leq I,\, TrU = k\}$$

and

$$\mathcal{D}_2^k \equiv \{D \in \mathcal{S}\Re^n \,|\, D \text{ is diagonal }, 0 \leq D \leq I, TrD = k\}.$$

Lemma 3.2 $\mathcal{S}_2^k = \{PDP^T \,|\, D \in \mathcal{D}_2^k,\ P \in \mathcal{O}^n\}$

Proof: It is obvious that $PDP^T \in \mathcal{S}_2^k$ for any $D \in \mathcal{D}_2^k$ and $P \in \mathcal{O}^n$. Indeed, the partial ordering $\leq$ on $\Re^{n \times n}$ is preserved by conjugation by orthogonal matrices, so that

$$0 \leq D \leq I \Longleftrightarrow 0 \leq PDP^T \leq I.$$

Also

$$Tr(PDP^T) = Tr(DP^TP) = TrD = k.$$

Conversely, given any $U \in \mathcal{S}_2^k$, a spectral decomposition of U yields

$$U = PDP^T, \text{ P orthogonal, D diagonal,}$$

i.e. the entries of D are the eigenvalues of U and the columns of P are the corresponding eigenvectors. Then, as before, we get

$$0 \leq U \leq I \Longrightarrow 0 \leq D \leq I$$

$$TrU = k \Longrightarrow TrD = k$$

and we conclude that $D \in \mathcal{D}_2^k$. □

Lemma 3.3 *$\mathcal{D}_2^k$ is a retract of $\mathcal{S}_2^k$, i.e. $\mathcal{D}_2^k \subset \mathcal{S}_2^k$ and there is a map*

$$r : \mathcal{S}_2^k \longrightarrow \mathcal{D}_2^k$$

such that $r(D) = D$ for all $D \in \mathcal{D}_2^k$.

Proof: Given $U \in \mathcal{S}_2^k$, define $r(U)$ to be the diagonal matrix consisting of the diagonal entries of U, i.e.

$$r(U)_{ij} = \begin{cases} u_{ii} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

It is obvious that $r(D) = D$ for $D \in \mathcal{D}_2^k$. We check that $r(U) \in \mathcal{D}_2^k$. $r(U)$ is diagonal by construction and clearly $Tr(r(U)) = Tr(U)$. Moreover, by lemma 2.2, we have

$$U \geq 0 \Longrightarrow u_{ii} \geq 0, \ \forall i$$

$$\begin{aligned} U \leq I &\Longleftrightarrow I - U \geq 0 \\ &\Longrightarrow 1 - u_{ii} \geq 0, \ \forall i \\ &\Longrightarrow u_{ii} \leq 1, \ \forall i \end{aligned}$$

Hence : $0 \leq r(U)_{ii} \leq 1$ for all $i = 1, \ldots, n$. But since $r(U)$ is diagonal, this implies that $0 \leq r(U) \leq I$. □

Now let $Q \in \mathcal{O}^n$ be such that $Q^TMQ = \Lambda$, where $\Lambda \equiv diag\{\lambda_1, \ldots, \lambda_n\}$.

Remark 3.4 Even with the order of the eigenvalues fixed the matrix Q is not uniquely determined, in general, due to possible eigenvalues with multiplicity greater than 1. Indeed, Q may be replaced by QP for any $P \in \mathcal{O}^n$ such that $P^T \Lambda P = \Lambda$. But

$$(\Lambda P)_{ij} = \sum_s \Lambda_{is} P_{sj} = \lambda_i P_{ij}$$

$$(P\Lambda)_{ij} = \sum_t P_{it} \Lambda_{tj} = \lambda_j P_{ij}$$

so that

$$\Lambda P = P\Lambda \Longrightarrow P_{ij} = 0 \text{ for all i,j with } \lambda_i \neq \lambda_j .$$

Hence P is block diagonal, $P = diag\,(P_1, \ldots, P_m)$, where m is the number of distinct eigenvalues of M. The size of P_l, for $1 \leq l \leq m$, is the multiplicity of the corresponding eigenvalue, and, of course, P_l is orthogonal.

It is clear that

$$\textstyle\sum_{i=1}^k \lambda_i = \max\limits_{D \in \mathcal{D}_2^k} < D, \Lambda > . \tag{3.5}$$

Since $r(\mathcal{S}_2^k) = \mathcal{D}_2^k$, we have

$$\max_{D \in \mathcal{D}_2^k} < D, \Lambda >= \max_{U \in \mathcal{S}_2^k} < r(U), \Lambda >$$

and, as Λ is diagonal, $< r(U), \Lambda >=< U, \Lambda >$. Thus

$$\begin{aligned} \textstyle\sum_{i=1}^k \lambda_i &= \max_{U \in \mathcal{S}_2^k} < U, \Lambda > \\ &= \max_{U \in \mathcal{S}_2^k} < U, Q^T M Q > \\ &= \max_{U \in \mathcal{S}_2^k} < Q U Q^T, M > . \end{aligned}$$

Furthermore, observe that the map

$$U \longmapsto Q U Q^T$$

induces a bijection of $\mathcal{S}_2^k$ onto itself, with inverse

$$V \longmapsto Q^T V Q$$

and it follows that

$$\textstyle\sum_{i=1}^{k} \lambda_i = \max_{V \in \mathcal{S}_2^k} < V, M > . \tag{3.6}$$

Next we determine which elements $U \in \mathcal{S}_2^k$ realize the maximum

$$\max_{V \in \mathcal{S}_2^k} < V, M > .$$

This requires more information about the eigenvalues of M. Let us assume that the k^{th}-eigenvalue has multiplicity t, so that

$$\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_k = \ldots = \lambda_{r+t} > \lambda_{r+t+1} \geq \ldots \geq \lambda_n \tag{3.7}$$

where $r \geq 0$. It is clear that those $D \in \mathcal{D}_2^k$ which realize

$$\max_{d \in \mathcal{D}_2^k} < D, \Lambda >$$

are of the form $D = diag(D_1, D_2, D_3)$ where the diagonal matrices D_1, D_2, D_3 satisfy

$$\begin{cases} D_1 = I_r, \text{ the identity } r \times r \text{ matrix} \\ D_3 = 0, \text{ the zero } (n-r-t) \times (n-r-t) \text{ matrix} \\ D_2 \in \mathcal{S}\Re^t, \text{ with } 0 \leq D_2 \leq I_t, \; TrD_2 = k - r. \end{cases}$$

Let us write $\mathcal{S}_2^p(t)$ for the set of all symmetric $t \times t$ matrices M of trace p with $0 \leq M \leq I_t$.

Lemma 3.8 *The matrices $U \in \mathcal{S}_2^k$ for which $r(U) = diag(D_1, D_2, D_3)$, D_1, D_2, D_3 as above, are precisely those matrices of the form*

$$U = \begin{bmatrix} I_r & 0 & 0 \\ 0 & U_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $U_2 \in \mathcal{S}_2^{k-r}(t)$.

Proof: Clearly any such $U \in \mathcal{S}_2^k$ satisfies

$$r(U) = diag(I_r, D_2, 0)$$

with $D_2 = r(U_2)$ of the appropriate type.

Now suppose that $U \in \mathcal{S}_2^k$ is such that $r(U) = diag(I_r, D_2, 0)$, and let us write

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{21} & U_{22} & U_{13} \\ U_{31} & U_{32} & U_{33} \end{bmatrix}$$

Note that if a matrix $V \in \Re^{n \times n}$ is ≥ 0, then so is any principal submatrix $\tilde{V}$ of V. Indeed, say $\tilde{V} \in \Re^{m \times m}$, then we have to show that $\tilde{x}^T \tilde{V} \tilde{x} \geq 0$ for all $\tilde{x} \in \Re^m$. But we can extend $\tilde{x}$ to a vector $x \in \Re^n$ by inserting 0's in the components corresponding to the rows of V not in $\tilde{V}$, and we get

$$\tilde{x}^T \tilde{V} \tilde{x} = x^T V x \geq 0.$$

Then, $0 \leq U \leq I$ immediately implies that $0 \leq U_{22} \leq I_t$. Obviously, $TrU_{22} = k - r$. It only remains to show that $U_{11} = I_r$ and all other blocks are 0. But this is an easy consequence of lemma 2.2. Indeed since $U \geq 0$ and $u_{jj} = 0$ if $j > r + t$, we conclude that

$$u_{ij} = 0 = u_{ji} \text{ for } j > r + t, \; i \leq j.$$

Thus $U_{31}, U_{32}, U_{33}, U_{23}, U_{13}$ are all 0. Moreover, since $V \equiv I - U \geq 0$,

$$v_{ii} = 0 \text{ for } 1 \leq i \leq r \text{ and } v_{ij} = u_{ij} \text{ for all i} \neq j,$$

we conclude that $u_{ji} = u_{ij} = v_{ij} = 0$ for $1 \leq i \leq r$, $i < j$. It follows that $U_{11} = I_r$ and $U_{12}, U_{13}, U_{21}, U_{31}$ are all 0. □

Now recall our choice of $Q \in \mathcal{O}^n$ such that $Q^T M Q = \Lambda$ and write

$$Q = [Q_1 : Q_2 : Q_3]$$

where $Q_1 \in \mathcal{O}^{n,r}$, $Q_2 \in \mathcal{O}^{n,t}$ and $Q_3 \in \mathcal{O}^{n,n-r-t}$. Thus

$$Q_1^T M Q_1 = diag\{\lambda_1, \ldots, \lambda_r\}$$

$$Q_2^T M Q_2 = \lambda_k I_t.$$

By lemma 3.8, the matrices $U \in \mathcal{S}_2^k$ realizing $\max_{V \in \mathcal{S}_2^k} < V, \Lambda >$ are of the form

$$U = \begin{bmatrix} I_r & 0 & 0 \\ 0 & U_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

with $U_2 \in \mathcal{S}\Re^t$, $0 \leq U_2 \leq I_t$ and $TrU_2 = k - r$. Then

$$QUQ^T = Q_1Q_1^T + Q_2U_2Q_2^T.$$

Since $< U, \Lambda >=< U, Q^TMQ >=< QUQ^T, M >$, we have proved the following result.

Theorem 3.9 *The matrices $U \in \mathcal{S}_2^k$ which realize*

$$\textstyle\sum_{i=1}^k \lambda_i = \max_{V \in \mathcal{S}_2^k} < V, M >$$

are of the form

$$U = Q_1Q_1^T + Q_2U_2Q_2^T$$

where $Q = [Q_1 : Q_2 : Q_3] \in \mathcal{O}^n$ satisfies $Q^TMQ = \Lambda$, and $U_2 \in \mathcal{S}\Re^t$ with $0 \leq U_2 \leq I_t$ and $TrU_2 = k - r$.

We have obtained three characterizations of $\sum_{i=1}^k \lambda_i$so far, namely equations (3.1), (3.5) and (3.6). We now introduce a set $\mathcal{D}_1^k$ whose relation to $\mathcal{S}_1^k$ is as that of $\mathcal{D}_2^k$ to $\mathcal{S}_2^k$. Define

$$\mathcal{D}_1^k \equiv \{D \in \mathcal{S}\Re^n \,|\, D \text{ is diagonal, } TrD = k \text{ and } d_{ii} = 0 \text{ or } 1, \text{ for all i } \}.$$

Lemma 3.10 *$\mathcal{D}_2^k$ is the convex hull of $\mathcal{D}_1^k$.*

Proof: Observe that the number of nonzero entries of a matrix $D \in \mathcal{D}_1^k$ is exactly k and that moreover they are all equal to 1. It is obvious that a convex combination of elements of $\mathcal{D}_1^k$ lies in $\mathcal{D}_2^k$. So we need to show that every $D \in \mathcal{D}_2^k$ is a convex combination of elements of $\mathcal{D}_1^k$.
We proceed by induction on the number $l(D)$ of nonzero entries of $D \in \mathcal{D}_2^k$. Clearly, the trace condition implies that $l(D) \geq k$, for all $D \in \mathcal{D}_2^k$, and that

$$l(D) = k \iff D \in \mathcal{D}_1^k \subset \mathcal{D}_2^k.$$

Obviously, $l(D) \leq n$ for all $D \in \mathcal{D}_2^k$.
Now suppose we have proved that any $D \in \mathcal{D}_2^k$ with $l(D) < m$, $(k < m \leq n)$, is a convex combination of elements of $\mathcal{D}_1^k$. We prove that the same holds for $D \in \mathcal{D}_2^k$ with $l(D) = m$. Let β denote the smallest nonzero entry of D, say $\beta = d_{ii}$ for some $1 \leq i \leq n$ (Note that i need not be uniquely determined). Define $\tilde{D} \in \mathcal{D}_1^k$ be such that

$$\begin{array}{ll} (i) & \tilde{d}_{ii} = 1 \\ (ii) & \tilde{d}_{jj} = 1 \implies d_{jj} > 0 \end{array}$$

(Again there may be many choices for such a $\tilde{D}$). Then let

$$D' \equiv D - \beta \tilde{D}.$$

Clearly D' is diagonal and $0 \leq D' \leq I$. Moreover,

$$Tr(D') = TrD - \beta Tr\tilde{D} = (1-\beta)k$$

and, since $(1-\beta) \neq 0$, we may define

$$\hat{D} \equiv \frac{1}{1-\beta} D'.$$

Observe that $\hat{D} \in \mathcal{D}_2^k$ and that $l(\hat{D}) \leq m-1 < m$, so that by induction, we have

$$\hat{D} = \sum_{j=1}^{s} \alpha_j D_j, \text{ with } 0 < \alpha_j \leq 1, \sum_{j=1}^{s} \alpha_j = 1 \text{ and } D_j \in \mathcal{D}_1^k.$$

Thus we get

$$\begin{aligned} D &= D' + \beta \tilde{D} \\ &= (1-\beta) \sum_{j=1}^{s} \alpha_j D_j + \beta \tilde{D} \\ &= \sum_{j=1}^{s} (1-\beta) \alpha_j D_j + \beta \tilde{D} \\ &= \sum_{r=1}^{s+1} \gamma_r \bar{D}_r \end{aligned}$$

where

$$\gamma_r = \begin{cases} (1-\beta)\alpha_r & 1 \leq r \leq s \\ \beta & r = s+1 \end{cases}$$

$$\bar{D}_r = \begin{cases} D_r & 1 \leq r \leq s \\ \tilde{D} & r = s+1 \end{cases}$$

Clearly, $\bar{D}_r \in \mathcal{D}_1^k$, for $1 \leq r \leq s+1$, and $0 \leq \gamma_r \leq 1$ with

$$\begin{aligned} \sum_{r=1}^{s+1} \gamma_r &= \sum_{r=1}^{s} (1-\beta)\alpha_r + \beta \\ &= (1-\beta) \sum_{r=1}^{s} \alpha_r + \beta \\ &= (1-\beta) + \beta = 1. \ \square \end{aligned}$$

Lemma 3.11 $\mathcal{S}_1^k = \{PDP^T \,|\, D \in \mathcal{D}_1^k,\ P \in \mathcal{O}^n\,\}$.

Proof: Let us write $\mathcal{C} = \{PDP^T \,|\, D \in \mathcal{D}_1^k,\ P \in \mathcal{O}^n\,\}$. Given $D \in \mathcal{D}_1^k$, let $i_1, \ldots, i_k$ denote the indices of the nonzero entries of D, i.e.

$$d_{tt} = 1 \iff t = i_j, \text{ for some } 1 \leq j \leq k.$$

Given $P \in \mathcal{O}^n$, write $P = [\,p_1, \ldots, p_n]$, where p_i denote the columns of P. Thus

$$p_i \in \Re^n,\ p_i^T p_j = \delta_{ij}, \text{ for } 1 \leq i, j \leq n.$$

Then $PD = [\,q_1, \ldots, q_n]$ where

$$q_j = \begin{cases} p_j & \text{if } j = i_s \text{ for some s} \\ 0 & \text{otherwise} \end{cases}$$

i.e. the columns p_j with $j \neq i_1, \ldots, i_k$ are replaced by 0 while the others are preserved. Hence $PDP^T = \sum_{t=1}^k p_{i_t} p_{i_t}^T$. We claim that

$$\mathcal{C} = \{QQ^T \,|\, Q \in \mathcal{O}^{n,k}\,\}.$$

Indeed, given PDP^T and $i_1, \ldots, i_k$ as above, let $Q = [\,p_{i_1}, \ldots, p_{i_k}] \in \mathcal{O}^{n,k}$. Then $PDP^T = QQ^T$. Conversely, given $Q = [q_1, \ldots, q_k] \in \mathcal{O}^{n,k}$, we may write

$$QQ^T = PDP^T$$

for many choices of P and D. For example, we may complete $\{q_1, \ldots, q_k\}$ to an orthonormal basis of $\Re^n$, say $\{q_1, \ldots, q_k, q_{k+1}, \ldots, q_n\}$ and let $D = diag\{\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0\}$ and $P = [q_1, \ldots, q_n]$. Thus we have

$$\begin{aligned} \mathcal{S}_1^k &= \left\{V \in \mathcal{S}\Re^n \,\middle|\, V = \sum_{i=1}^k v_i v_i^T,\ \{v_1, \ldots, v_k\} \text{ orthonormal set }\right\} \\ &= \{QQ^T \,|\, Q \in \mathcal{O}^{n,k}\,\} \\ &= \{V \in \mathcal{S}\Re^n \,|\, V = PDP^T,\ D \in \mathcal{D}_1^k,\ P \in \mathcal{O}^n\,\}. \ \Box \end{aligned}$$

Lemma 3.12 *$\mathcal{S}_2^k$ is the convex hull of $\mathcal{S}_1^k$.*

Proof: We have

$$\mathcal{S}_2^k = \{U \in \mathcal{S}\Re^n \,|\, U = PDP^T,\ D \in \mathcal{D}_2^k,\ P \in \mathcal{O}^n\,\}$$

$$\mathcal{S}_1^k = \{V \in \mathcal{S}\Re^n \,|\, V = PDP^T,\ D \in \mathcal{D}_1^k,\ P \in \mathcal{O}^n\,\}.$$

The result follows since $\mathcal{D}_2^k$ is the convex hull of $\mathcal{D}_1^k$ by lemma 3.10. $\Box$

Remark 3.13 Observe that the elements of $\mathcal{D}_1^k$ and $\mathcal{S}_1^k$ have rank exactly k while those of $\mathcal{D}_2^k$ and $\mathcal{S}_2^k$ have rank $\geq k$.

Remark 3.14 We proved earlier the existence of a retraction

$$r : \mathcal{S}_2^k \longrightarrow \mathcal{D}_2^k.$$

However, the restriction of r to $\mathcal{S}_1^k$ does not yield a retraction of $\mathcal{S}_1^k$ onto $\mathcal{D}_1^k$, but maps $\mathcal{S}_1^k$ into $\mathcal{D}_2^k$. We may summarize the relationships between the spaces $\mathcal{D}_1^k$, $\mathcal{D}_2^k$, $\mathcal{S}_1^k$ and $\mathcal{S}_2^k$ in the following diagram.

$$\begin{tikzcd}
\mathcal{D}_1^k \arrow[r, "\text{convex hull}"] \arrow[d] & \mathcal{D}_2^k \arrow[d, shift right] \\
\mathcal{S}_1^k \arrow[r, "\text{convex hull}"'] \arrow[ur, "r"] & \mathcal{S}_2^k \arrow[u, shift right, "r"']
\end{tikzcd}$$

Given a symmetric matrix M with spectral decomposition $Q^TMQ = \Lambda$, we trivially have the following characterization of $\sum_{i=1}^k \lambda_i$.

$$\textstyle\sum_{i=1}^k \lambda_i = \max_{\tilde{D} \in \mathcal{D}_1^k} < \tilde{D}, \Lambda > .$$

Taking the convex hull of $\mathcal{D}_1^k$ gives

$$\max_{\tilde{D} \in \mathcal{D}_1^k} < \tilde{D}, \Lambda >= \max_{D \in \mathcal{D}_2^k} < D, \Lambda > .$$

Then

$$\max_{D \in \mathcal{D}_2^k} < D, \Lambda >= \max_{U \in \mathcal{S}_2^k} < U, \Lambda > ,$$

since $\mathcal{D}_2^k$ is a retract of $\mathcal{S}_2^k$ and since, as Λ is diagonal,

$$< U, \Lambda >=< r(U), \Lambda >$$

for any $U \in \mathcal{S}_2^k$. Now, since $\mathcal{S}_2^k$ is the convex hull of $\mathcal{S}_1^k$, we conclude that

$$\max_{V \in \mathcal{S}_1^k} < V, \Lambda >= \max_{U \in \mathcal{S}_2^k} < U, \Lambda > .$$

Finally, since $\mathcal{S}_1^k$ and $\mathcal{S}_2^k$ are invariant with respect to conjugation by orthonormal matrices, we conclude that

$$\textstyle\sum_{i=1}^k \lambda_i = \max_{V \in \mathcal{S}_1^k} < V, M >= \max_{U \in \mathcal{S}_2^k} < U, M > .$$

4 The generalized gradient of $\sum_{i=1}^{k} \lambda_i : \mathcal{S}\Re^n \longrightarrow \Re$

We now consider $\sum_{i=1}^{k} \lambda_i$ as a function s_k defined on $\mathcal{S}\Re^n$,

$$s_k : \mathcal{S}\Re^n \longrightarrow \Re.$$

For $M \in \mathcal{S}\Re^n$ and $V \in \Re^{n\times n}$, let us write

$$\mathcal{L}_V(M) \equiv < V, M > .$$

Then $\mathcal{L}_V$ is a linear functional on $\mathcal{S}\Re^n$ and by (3.6) we have

$$s_k(M) = \max_{U\in\mathcal{S}_2^k} \mathcal{L}_U(M).$$

We obtain the following characterization of the generalized gradient ∂s_k of s_k.

Theorem 4.1 *For $M \in \mathcal{S}\Re^n$ we get*

$$\partial s_k(M) = \{U \in \mathcal{S}\Re^n \,|\, U = Q_1Q_1^T + Q_2U_2Q_2^T\,\}$$

where the eigenvalues of M satisfies (3.7),and $Q = [Q_1 : Q_2 : Q_3] \in \mathcal{O}^n$ and $U_2 \in \mathcal{S}\Re^t$ are as in theorem 3.9.

Proof: By [7, theorem 2.8.6, p. 92] and theorem 3.9, we know that $\partial s_k(M)$ is the convex hull of

$$\{U \in \mathcal{S}\Re^n \,|\, U = Q_1Q_1^T + Q_2U_2Q_2^T\,\}.$$

So, we only need to observe that this set is already convex. It is sufficient to prove that the set

$$\mathcal{A} = \{V \in \mathcal{S}\Re^t \,|\, 0 \le V \le I_t,\, TrV = p\,\},$$

for some fixed p, is convex. But for $V_1, V_2 \in \mathcal{A}$, $s \in [0,1]$, it is clear that

$$\begin{gathered} sV_1 + (1-s)V_2 \in \mathcal{S}\Re^t \\ 0 \le sV_1 + (1-s)V_2 \le I_t. \end{gathered}$$

Moreover,

$$Tr(sV_1 + (1-s)V_2) = sTrV_1 + (1-s)TrV_2 = p. \;\Box$$

Remark 4.2 Consider the special case $k = 1$. Then $r = 0$ in (3.7) and $s_1(M) = \lambda_1(M)$ for $M \in \mathcal{S}\Re^n$. We get

$$\partial\lambda_1(M) = \{V \in \mathcal{S}\Re^n \,\big|\, V = Q_1 U Q_1^T,\, U \in \mathcal{S}\Re^n,\, 0 \le U \le I_t, TrU = 1\,\}$$

where $Q_1 \in \mathcal{O}^{n,t}$, $Q_1^T M Q_1 = \lambda_1(M) I_t$. Observe that the condition $U \le I_t$ is redundant, i.e.

$$U \in \mathcal{S}\Re^t,\, U \ge 0,\, TrU = 1 \implies U \le I_t.$$

Indeed, let $\{\mu_1, \ldots, \mu_t\}$ denote the eigenvalues of U. We have $\mu_i \ge 0$ for all i since $U \ge 0$. Then

$$\begin{aligned} TrU = 1 \quad &\implies \quad \sum_{i=1}^{t} \mu_i = 1 \\ &\implies \quad \mu_i \le 1, \text{ for all i} \\ &\implies \quad U \le I_t. \end{aligned}$$

So we recover the characterization of $\partial\lambda_1(M)$ derived in [16].

5 The differential of $\Psi_V : \mathcal{S}\Re^n \times \mathcal{Q}^n \longrightarrow \Re$

We may now apply the results of section 3 to a symmetric definite pencil (A, B). So let G denote the Choleski factor of B and let $Q \in \mathcal{Q}^n$ be such that

$$Q^T(G^{-1}AG^{-T})Q = \Lambda$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ and $\lambda_1 \geq \ldots \geq \lambda_n$ are the eigenvalues of (A, B). Let $X \equiv G^{-T}Q$, so that

$$\begin{cases} X^TBX = I \\ X^TAX = \Lambda \end{cases}$$

Let us write g_k for $\sum_{i=1}^k \lambda_i$ viewed as a function on $\mathcal{S}\Re^n \times \mathcal{Q}^n$ and, for $V \in \Re^{n \times n}$, let

$$\Psi_V(A, B) \equiv < V, G^{-1}AG^{-T} > .$$

Since $\Lambda(A, B) = \Lambda(G^{-1}AG^{-T})$ and $G^{-1}AG^{-T}$ is symmetric, we get (see (3.6))

$$g_k(A, B) = \max_{V \in \mathcal{S}_2^k} \Psi_V(A, B) \tag{5.1}$$

and

$$g_k(A, B) = \max_{U \in X\mathcal{S}_2^kX^T} < U, A > . \tag{5.2}$$

Moreover, theorem 3.9 provides a description of those $V \in \mathcal{S}_2^k$ which realize the maximum. Equation (5.2) is not a very useful representation of $g_k(A, B)$ because the set over which the maximum is taken depends upon B. On the other hand, in order to use equation (5.1) for computing generalized gradients we need to be able to compute the derivative of the Choleski factor G as a function of B. Observe that the functions $\Psi_V : \mathcal{S}\Re^n \times \mathcal{Q}^n \longrightarrow \Re$, while linear in A, are not linear in B.

We now define maps α, β, and γ as follows.

$$\begin{array}{cccc} \alpha : & \mathcal{Q}^n & \longrightarrow & \mathcal{T}^n \\ & B & \longmapsto & G \end{array}$$

where $\mathcal{T}^n \subset \Re^{n \times n}$ denotes the subset of lower triangular matrices with positive diagonal entries. Let $\mathcal{V}^n$ denote the linear space of all lower triangular matrices. Then $\mathcal{V}^n \cong \Re^{n(n+1)/2}$ and we view $\mathcal{T}^n$ as an *open* subset of $\mathcal{V}^n$. Similarly $\mathcal{Q}^n$ is an *open* subset of $\mathcal{S}\Re^n \cong \Re^{n(n+1)/2}$.

$$\begin{array}{cccc} \beta : & \mathcal{T}^n & \longrightarrow & \mathcal{T}^n \\ & G & \longmapsto & G^{-1} \end{array}$$

Observe that the inverse of an element of $\mathcal{T}^n$ is indeed an element of $\mathcal{T}^n$.

$$\begin{array}{rccc} \gamma: & \mathcal{S}\Re^n \times \mathcal{T}^n & \longrightarrow & \mathcal{S}\Re^n \\ & (A, L) & \longmapsto & LAL^T \end{array}$$

Also, recall the map $\mathcal{L}_V$ defined in section 4,

$$\begin{array}{rccc} \mathcal{L}_V: & \mathcal{S}\Re^n & \longrightarrow & \Re^n \\ & M & \longmapsto & < V, M > \end{array}$$

where $V \in \mathcal{S}_2^k$. Then we have

$$\Psi_V(A, B) = \mathcal{L}_V(\gamma\,((1 \times \beta)\,((1 \times \alpha)\,(A, B)))). \qquad (5.3)$$

Note that $\mathcal{S}\Re^n \times \mathcal{Q}^n$ is an open subset of $\mathcal{S}\Re^n \times \mathcal{S}\Re^n$, so that the tangent space to $\mathcal{S}\Re^n \times \mathcal{Q}^n$ at a point (A, B) is

$$\mathcal{S}\Re^n \times \mathcal{S}\Re^n \cong \Re^{n(n+1)/2} \times \Re^{n(n+1)/2}.$$

Similarly, the tangent space to $\mathcal{T}^n$ at a point G is

$$\mathcal{V}^n \cong \Re^{n(n+1)/2}.$$

We do, however, write our matrices as square matrices; otherwise we should modify the definition of the inner product $< \, , >$ as follows : multiply the terms corresponding to the entries lying below the diagonal by a factor of 2. Now for $(S, T) \in \mathcal{S}\Re^n \times \mathcal{S}\Re^n$, the chain rule gives the following expression for the differential $D\Psi_V(A, B)$ evaluated at (S, T).

$$D\Psi_V(A, B)(S, T) \;=$$

$$D\mathcal{L}_V(\gamma\,((1 \times \beta)\,((1 \times \alpha)\,(A, B))))(D[\gamma \circ (1 \times \beta) \circ (1 \times \alpha)](A, B)(S, T).$$

But $\mathcal{L}_V$ is linear so that we get

$$D\Psi_V(A, B)(S, T) =< V, D[\gamma \circ (1 \times \beta) \circ (1 \times \alpha)](A, B)(S, T) > . \qquad (5.4)$$

We compute the differentials of the maps α, β and γ next.

The differential of β. Given $G \in \mathcal{T}^n$ and $L \in \mathcal{V}^n$, we must compute $D\beta(G)(L)$. Since $\beta(G)G = I$, for $1 \le l \le k \le n$ we have

$$\partial_{kl}\beta(G)G + \beta(G)E_{kl} = 0$$

where E_{kl} is the matrix with 1 in the (k,l)-entry and 0 everywhere else, so that

$$\partial_{kl}\beta(G) = -G^{-1}E_{kl}G^{-1}.$$

Hence, for $L = \{l_{ij}\} \in \mathcal{V}^n$,

$$\begin{aligned} D\beta(G)(L) &= \sum_{1\leq l\leq k\leq n} l_{kl}\partial_{kl}\beta(G) \\ &= \sum_{1\leq l\leq k\leq n} l_{kl}(-G^{-1}E_{kl}G^{-1}) \\ &= -G^{-1}(\sum_{1\leq l\leq k\leq n} l_{kl}E_{kl})G^{-1} \\ &= -G^{-1}LG^{-1} \in \mathcal{V}^n. \end{aligned} \tag{5.5}$$

The differential of α*.* It is clear that the map $\alpha : \mathcal{Q}^n \longrightarrow \mathcal{T}^n \subset \mathcal{V}^n$, $\alpha(B) = G$, is smooth; it involves only rational and square roots functions. Consider the map

$$\begin{array}{cccc} \hat{\alpha} : & \mathcal{V}^n & \longrightarrow & \mathcal{S}\Re^n \\ & L & \longmapsto & LL^T \end{array}$$

The restriction of $\hat{\alpha}$ to $\mathcal{T}^n$ maps $\mathcal{T}^n$ into $\mathcal{Q}^n$. Indeed, given $x \in \Re^n$, $L^Tx \neq 0$ when $x \neq 0$ since L^T is nonsingular for $L \in \mathcal{T}^n$. Thus,

$$x^TLL^Tx = (L^Tx)^T(L^Tx) > 0.$$

Moreover, $\hat{\alpha}$ is obviously smooth. Hence α is a diffeomorphism of $\mathcal{Q}^n$ onto $\mathcal{T}^n$ with inverse $\alpha^{-1} = \hat{\alpha}$, and

$$D\alpha(B) = [D\alpha^{-1}(\alpha(B))]^{-1}$$

We shall not compute $D\alpha(B)$ (while straightforward, it is complicated and unnecessary for our purposes); rather we will provide an algorithm for computing $D\alpha(B)(M)$ for $B \in \mathcal{Q}^n$ and $M \in \mathcal{S}\Re^n$, which is,after all, what we are interested in. Now

$$D\alpha(B)(M) = L \in \mathcal{V}^n$$

if and only if

$$D\alpha^{-1}(\alpha(B))(L) = M \tag{5.6}$$

Hence we may compute $D\alpha(B)(M)$ by solving equation 5.6 for L. For (k,l) with $1 \leq l \leq k \leq n$, we have

$$\partial_{kl}\alpha^{-1}(G) = \partial_{kl}(GG^T) = E_{kl}G^T + GE_{lk}.$$

Hence

$$\begin{aligned} D\alpha^{-1}(G)(L) &= \sum_{1\leq t\leq s\leq n} l_{st}\partial_{st}\alpha^{-1}(G) \\ &= \sum_{1\leq t\leq s\leq n} l_{st}[E_{st}G^T + GE_{ts}] \\ &= \left(\sum l_{st}E_{st}\right)G^T + G\left(\sum l_{st}E_{ts}\right) \\ &= LG^T + GL^T \in \mathcal{S}\Re^n. \end{aligned} \tag{5.7}$$

Thus we must solve the equation $LG^T + GL^T = M$ for the lower triangular matrix L. This is accomplished by the following algorithm.

Algorithm.

```
for $i = 1, n$
    for $j = 1, i-1$
        $l_{ij} = \frac{m_{ij} - \sum_{k=1}^{j-1} l_{ik}g_{jk} - \sum_{k=1}^{j} g_{ik}l_{jk}}{g_{jj}}$    (5.8)
    $l_{ii} = \frac{m_{ii} - 2\sum_{k=1}^{i-1} l_{ik}g_{ik}}{2g_{ii}}$    (5.9)
```

First observe that the algorithm is well defined.

1. $g_{ii} \neq 0$ for all $i = 1, n$ since $G \in \mathcal{T}^n$.

2. all the entries of L appearing on the right hand sides of (5.8) and (5.9) have been computed during prior loop iterations.

Next note that

$$(LG^T)_{ij} = \sum_{k=1}^{min(i,j)} l_{ik}g_{jk}.$$

To prove the algorithm correct, we order the entries l_{ij} of L using the lexicographic order on (i, j), where $1 \leq j \leq i \leq n$, and show by induction that, having computed $l_{i'j'}$ for all $(i', j') \leq (i, j)$, we indeed obtain l_{ij} by formulas (5.8) and (5.9).
$(i, j) = (1, 1)$: this is completely trivial; $[LG^T + GL^T]_{11} = l_{11}g_{11} + g_{11}l_{11}$, so that

$$l_{11} = \frac{m_{11}}{2g_{11}}$$

which is precisely (5.9) in this case.

general case : we have

$$[LG^T + GL^T]_{ij} = \sum_{k=1}^{min(i,j)} l_{ik}g_{jk} + \sum_{k=1}^{min(i,j)} g_{ik}l_{jk}. \qquad (\star)$$

If $i = j$, we get : $2\sum_{k=1}^{i} l_{ik}g_{ik} = m_{ii}$ and formula (5.9) follows. If $j < i$, then all l_{jk}, $1 \leq k \leq min(i,j) = j$, and all l_{ik}, $1 \leq k < j$ have been computed already according to the induction hypothesis, and formula (5.8) follows from $(\star)$.

The differential of γ. The map γ is defined on the space $\mathcal{S}\Re^n \times \mathcal{T}^n$. We will use indices (k,l) to refer to the variables in the first component and indices (s,t) to refer to those in the second component. In all cases $1 \leq l \leq k \leq n$ and $1 \leq t \leq s \leq n$. Since $\gamma(A,L) = LAL^T$ for $A \in \mathcal{S}\Re^n$, $L \in \mathcal{T}^n$, we have

$$\begin{array}{rcl} \partial_{kl}\gamma(A,L) & = & LE_{kl}L^T \\ \partial_{st}\gamma(A,L) & = & E_{st}AL^T + LAE_{ts}. \end{array}$$

Thus for $(M,N) \in \mathcal{S}\Re^n \times \mathcal{V}^n$, we have

$$D\gamma(A,L)(M,N) = LML^T + NAL^T + LAN^T. \qquad (5.10)$$

Now, from (5.5), we readily obtain

$$D(1 \times \beta)(A,G)(M,N) = (M, -G^{-1}NG^{-1}) \qquad (5.11)$$

where $A \in \mathcal{S}\Re^n$, $G \in \mathcal{T}^n$, $M \in \mathcal{S}\Re^n$, $N \in \mathcal{V}^n$ and

$$1 \times \beta : \mathcal{S}\Re^n \times \mathcal{T}^n \longrightarrow \mathcal{S}\Re^n \times \mathcal{T}^n.$$

Similarly, from (5.7), we get

$$D(1 \times \alpha)(A,B)(M,N) = (M,L) \qquad (5.12)$$

where $A \in \mathcal{S}\Re^n$, $B \in \mathcal{Q}^n$, $M,N \in \mathcal{S}\Re^n$,

$$1 \times \alpha : \mathcal{S}\Re^n \times \mathcal{Q}^n \longrightarrow \mathcal{S}\Re^n \times \mathcal{T}^n$$

and $L \in \mathcal{V}^n$ solves $LG^T + GL^T = N$.
Finally, combining (5.4), (5.10), (5.11) and (5.12), we obtain the following expression for the differential of Ψ_V.

$$\begin{array}{rcl} D\Psi_V(A,B)(M,N) & = & < V, G^{-1}MG^T - G^{-1}LG^{-1}AG^{-T} - G^{-1}AG^{-T}L^TG^{-T} > \\ & = & < V, G^{-1}[M - LG^{-1}A - AG^{-T}L^T]G^{-T} > \qquad (5.13) \end{array}$$

where $L \in \mathcal{V}^n$ solves the equation $LG^T + GL^T = N$.

6 The generalized gradient of $\sum_{i=1}^{k} \lambda_i(\mathbf{x})$

We now assume that we are given a smooth function

$$\begin{array}{rccc} h: & \Omega & \longrightarrow & \mathcal{S}\Re^n \times \mathcal{Q}^n \\ & x & \longmapsto & (A(x), B(x)) \end{array}$$

where Ω is an open set in $\Re^m$, and we consider the composite

$$\begin{array}{rccc} f_k \equiv g_k \circ h: & \Omega & \longrightarrow & \Re \\ & x & \longmapsto & \sum_{i=1}^{k} \lambda_i(x). \end{array}$$

We want to compute the generalized gradient of f_k. We shall write $A_j(x)$, $B_j(x)$ for the partial derivatives of $A(x)$, respectively $B(x)$, with respect to the variable x_j. By equation 3.6, we know that

$$f_k(x) = \max_{V \in \mathcal{S}_2^k} \Psi_V(A(x), B(x)) \qquad (6.1)$$

where

$$\Psi_V(A(x), B(x)) = < V, G(x)^{-1} A(x) G(x)^{-T} >, \qquad (6.2)$$

with $G(x)$ the Choleski factor of $B(x)$.

Proposition 6.3 *The function $f_k : \Omega \longrightarrow \Re$ is regular and locally Lipschitz.*

Proof: Equation 6.1 gives $f_k(x)$ as a pointwise maximum over the compact set $\mathcal{S}_2^k$, so that the conclusion will follow from [7, theorem 2.8.2, p. 86] if we show that the functions $\Psi_V(A(x), B(x))$ are regular and locally Lipschitz of some rank l for all $V \in \mathcal{S}_2^k$. They certainly are regular since they are smooth functions of x. Moreover, the differentiability of $\Psi_V(A(x), B(x))$ implies that it is Lipschitz of rank l_V in a neighborhood of x. More precisely, by the Mean Value Theorem,

$$|\, \Psi_V(A(x'), B(x')) - \Psi_V(A(x''), B(x''))\, | \leq \|\, \nabla \Psi_V(A(x'''), B(x'''))\, \|\, \|\, x' - x''\, \| \; .$$

for some x''' on the line segment from x' to x''. Now let $\mathcal{N}$ be a convex compact neighborhood of x in Ω and let

$$l = max \, \|\, \nabla \Psi_V(A(y), B(y))\, \|$$

where the max is taken over all $y \in \mathcal{N}$ and $V \in \mathcal{S}_2^k$. Then, each $\Psi_V(A(x), B(x))$, $V \in \mathcal{S}_2^k$, is Lipschitz of rank l in $\mathcal{N}$. □

Recall that by theorem 3.9, the elements $V \in \mathcal{S}_2^k$ which achieve the maximum in 6.1 are precisely those elements of $U \in \mathcal{S}_2^k$ of the form

$$U = Q_1 Q_1^T + Q_2 U_2 Q_2^T$$

where $U_2 \in \mathcal{S}_2^{k-r}(t)$ and $Q_1 \in \mathcal{O}_{n,r}$, $Q_2 \in \mathcal{O}_{n,t}$ are such that

$$Q = [Q_1 : Q_2 : Q_3], \; Q_3 \in \mathcal{O}_{n,n-r-t}$$

diagonalizes $G(x)^{-1}A(x)G(x)^{-T}$, i.e.

$$Q^T G(x)^{-1} A(x) G(x)^{-T} Q = \Lambda$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ and the eigenvalues satisfy (3.7). Let us write $\mathcal{M}(x)$ for the set of such matrices U. Following the terminology of [16] we shall call the matrices $U \in \mathcal{M}(x)$, as well as the matrices $U_2 \in \mathcal{S}_2^{k-r}(t)$, dual matrices.

Theorem 6.4 *The generalized gradient of f_k at $x \in \Omega$ is given by*

$$\partial f_k(x) = \{v \in \Re^m \,\big|\, v_j =< U, G^{-1}[A_j(x) - L_j G^{-1} A(x) - A(x) G^{-T} L_j^T] G^{-T} >\} \tag{6.5}$$

where $G = G(x)$ is the Choleski factor of $B(x)$, $L_j = L_j(x)$ solves

$$L_j G^T + G L_j^T = B_j(x)$$

and $U \in \mathcal{M}(x)$.

Proof: Again this follows from [7, theorems 2.8.2, p.86 and 2.8.6, p. 92] once we show that the set on the right hand side of equation 6.5 is convex. Let $\mathcal{A}$ denote this set. Thus, given α_i, $1 \le i \le s$, with $0 \le \alpha_i \le 1$ and $\sum \alpha_i = 1$, and $v^i \in \mathcal{A}$, $1 \le i \le s$, we must show that

$$\sum_{i=1}^{s} \alpha_i v^i \in \mathcal{A}.$$

It is enough to check that $\sum \alpha_i v_j^i$ is of the required form for all $j = 1, \ldots, m$. Let

$$Z = G^{-1}[A_j(x) - L_j G^{-1} A(x) - A(x) G^{-T} L_j^T] G^{-T}.$$

Then, $v_j^i =< U_i, Z >$, $1 \le j \le m$, and

$$\sum_{i=1}^{s} \alpha_i v_j^i =< \sum_{i=1}^{s} \alpha_i U_i, Z > .$$

Now $U_i = Q_1Q_1^T + Q_2U_2^iQ_2^T$ for $U_2^i \in \mathcal{S}_2^{k-r}(t)$, and

$$\begin{aligned} \sum_{i=1}^{s} \alpha_i U_i &= \sum_{i=1}^{s} \alpha_i (Q_1Q_1^T + Q_2U_2^iQ_2^T) \\ &= Q_1Q_1^T + Q_2 \left(\sum_{i=1}^{s} \alpha_i U_2^i \right) Q_2^T. \end{aligned}$$

Thus, the result follows from the convexity of $\mathcal{S}_2^{k-r}(t)$. □

Now let $f_k^\circ(x;d)$ denote the generalized directional derivative of f_k at x in the direction d, [7, p. 25]. Then

$$f_k^\circ(x;d) = max\{< v, d > \mid v \in \partial f_k(x)\}$$

by [7, Proposition 2.1.2, p. 27]. Since f_k is regular at x, the usual one-sided directional derivative $f_k'(x;d)$ exists and is equal to $f_k^\circ(x;d)$, see [7, Definition 2.3.4, p. 39], so that

$$f_k'(x;d) = max\{< v, d > \mid v \in \partial f_k(x)\}.$$

Using our characterization of $v \in \partial f_k(x)$ given in equation(6.5), we get

$$\begin{aligned} < v, d > &= \sum_{j=1}^{m} v_j d_j \\ &= \sum_{j=1}^{m} d_j < U, G^{-1}[A_j(x) - L_jG^{-1}A(x) - A(x)G^{-T}L_j^T]G^{-T} > \\ &= < U, \sum_{j=1}^{m} d_jG^{-1}[A_k(x) - L_jG^{-1}A(x) - A(x)G^{-T}L_j^T]G^{-T} > \end{aligned}$$

with $U \in \mathcal{M}(x)$, i.e. $U = Q_1Q_1^T + Q_2U_2Q_2^T$. Thus

$$\begin{aligned} < v, d > &= < I_r, M_1(d) > + < U_2, M_2(d) > \\ &= TrM_1(d) + < U_2, M_2(d) >, \end{aligned}$$

where

$$M_1(d) \equiv \sum_{j=1}^{m} d_jQ_1^TG^{-1}[A_j(x) - L_jG^{-1}A(x) - A(x)G^{-T}L_j^T]G^{-T}Q_1 \in \mathcal{S}\Re^r$$

and

$$M_2(d) \equiv \sum_{j=1}^{m} d_j Q_2^T G^{-1}[A_j(x) - L_j G^{-1} A(x) - A(x) G^{-T} L_j^T] G^{-T} Q_2 \in \mathcal{S}\Re^t.$$

We obtain the following result.

Proposition 6.6 *The directional derivative $f_k'(x;d)$ is equal to*

$$TrM_1(d) + \textit{sum of the } (k-r) \textit{ largest eigenvalues of } M_2(d).$$

Proof: We have

$$\begin{aligned} f_k'(x;d) &= max\{< v, d > \mid v \in \partial f_k(x)\} \\ &= TrM_1(d) + \max_{U_2 \in \mathcal{S}_2^{k-r}(t)} < U_2, M_2(d) > \end{aligned}$$

But by the results of section 3,

$$\max_{U_2 \in \mathcal{S}_2^{k-r}(t)} < U_2, M_2(d) >$$

is precisely the sum of the $k-r$ largest eigenvalues of $M_2(d)$. □

7 The ratio of the first two eigenvalues of a symmetric definite pencil

We now consider the following problem.

Problem : Given a smooth function

$$h : \Omega \longrightarrow \mathcal{S}\Re^n \times \mathcal{Q}^n,\ \Omega \subset \Re^m,$$

we want to find

$$\min_{x \in \Omega} \frac{\lambda_1(x)}{\lambda_2(x)} \tag{7.1}$$

where $\lambda_1(x) \geq \lambda_2(x)$ are the two largest eigenvalues of the symmetric definite pencil $h(x) = (A(x), B(x))$. Observe that the above problem is most interesting when $\lambda_1(x) < 0$. We will come back to this point below, but first we recall Clarke's characterization of the generalized gradient of a quotient (see [7]). In the following discussion, the functions are assumed to be Lipschitz near the point x.

By [7, Proposition 2.3.1, p. 38], we know that

$$\partial(tf)(x) = t\partial f(x)$$

for any $t \in \Re$. In particular, $\partial(-f)(x) = -\partial f(x)$. We also have the chain rule for generalized gradients, [7, theorem 2.3.9, p. 42]. More precisely, given $g : \Re^n \longrightarrow \Re$ and $h : \Re^m \longrightarrow \Re^n$, with $h_i : \Re^m \longrightarrow \Re$, $1 \leq i \leq n$, denoting the components of h, let $f = g \circ h$ denote the composite of g and h. Then f is Lipschitz near x if g and h_i, $1 \leq i \leq n$, are Lipschitz near x, and

$$\partial f(x) \subset co \left\{ \sum_1^n \sigma_i \xi_i \,\middle|\, \xi_i \in \partial h_i(x), \sigma = (\sigma_1, \ldots, \sigma_n), \sigma \in \partial g(h(x)) \right\}.$$

Moreover, equality holds and f is regular at x if the following conditions hold.

1. g is regular at $h(x)$.

2. h_i is regular at x, $1 \leq i \leq n$.

3. for every $\sigma \in \partial g(h(x))$, $\sigma_i \geq 0$ for $1 \leq i \leq n$.

The quotient rule is now a consequence of the chain rule. Consider the map $g(u, v) = u/v$ defined on $\{(u, v) \in \Re^2 \,|\, v \neq 0\}$. This map is continuously differentiable, hence strictly differentiable, at any point. Thus g and $-g$ are both regular everywhere, and the generalized gradient of g at (u, v) is just the gradient of g, i.e.

$$\partial g(u, v) = (\frac{1}{v}, -\frac{u}{v^2}).$$

Given $\alpha, \beta : \Re^m \longrightarrow \Re$, we let $h(x)$ denote the maps

$$h(x) = (\pm\alpha(x), \pm\beta(x)),$$

and apply the chain rule to the composite $g \circ h$. We get

Lemma 7.2 *For $\alpha, \beta : \Re^m \longrightarrow \Re$ with $\beta(x) \neq 0$,*

$$\partial \left(\frac{\alpha}{\beta}\right)(x) \subset \frac{\beta(x)\partial\alpha(x) - \alpha(x)\partial\beta(x)}{\beta^2(x)}$$

and equality holds in any of the following cases.

1. $\alpha(x) \geq 0, \beta(x) > 0$, α, $-\beta$ *are regular at* x
2. $\alpha(x) \leq 0, \beta(x) > 0$, α, β *are regular at* x
3. $\alpha(x) \geq 0, \beta(x) < 0$, $-\alpha$, $-\beta$ *are regular at* x
4. $\alpha(x) \leq 0, \beta(x) < 0$, $-\alpha$, β *are regular at* x.

Proof: We prove the first case; the others are entirely similar. Now

$$\frac{\alpha(x)}{\beta(x)} = -g(\alpha(x), -\beta(x)).$$

Thus we conclude that

$$\partial \left(\frac{\alpha}{\beta}\right)(x) \subset co\left\{\frac{-\xi_1}{-\beta(x)} + \frac{\alpha(x)}{(-\beta(x))^2}\,\xi_2 \,\middle|\, \xi_1 \in \partial\alpha(x), \xi_2 \in \partial(-\beta)(x)\right\}$$

Since $\partial(-\beta)(x) = -\partial\beta(x)$, the set on the right hand side is equal to

$$\left\{\frac{\beta(x)\xi_1 - \alpha(x)\xi_2}{\beta^2(x)} \,\middle|\, \xi_1 \in \partial\alpha(x), \xi_2 \in \partial\beta(x)\right\}.$$

Next, we observe that this set is convex since $\partial\alpha(x)$ and $\partial\beta(x)$ are convex ([7, Proposition 2.1.2,p. 27]). Finally, we easily verify that the equality conditions are satisfied. We have

$$h(x) = (\alpha(x), -\beta(x)),$$

so that h_i is regular at x, for $i = 1, 2$, and

$$\partial(-g)(h(x)) = \left(\frac{1}{\beta(x)}, \frac{\alpha(x)}{\beta^2(x)}\right)$$

clearly has nonnegative components. □

We now return to problem (7.1). We make the following assumptions on the map h.

$$-A(x) \in \mathcal{Q}^n, \text{ for all } x \in \Omega \tag{7.3}$$

$$\lambda_1(x) \text{ is a simple eigenvalue for all } x \in \Omega. \tag{7.4}$$

Hypothesis (7.3) implies that $\lambda_1(x)$, and hence all the other eigenvalues, are strictly negative for all $x \in \Omega$. Hypothesis (7.4) implies that $\lambda_1(x)$ is a smooth function of x. Thus so is $-\lambda_1(x)$, and both are regular at x. We point out, however, that (7.4) is not a very restrictive assumption. Indeed, if x^* is a solution of (7.1) with $\lambda_1(x^*) < 0$ then $\lambda_2(x^*) < \lambda_1(x^*)$ unless λ_1 has multiplicity at least 2 near x^*.

Consider the following lemma, whose proof is completely obvious.

Lemma 7.5 *Let $\alpha(x)$, $\beta(x)$, $x \in \Omega \subset \Re^m$ denote two real valued functions such that $\beta(x) \neq 0$ for all $x \in \Omega$ and*

$$\frac{\alpha(x)}{\beta(x)} > 0, \textit{ for all } x \in \Omega.$$

Then $x_0 \in \Omega$ minimizes $\frac{\alpha(x)}{\beta(x)}$ over Ω if and only if x_0 maximizes $\frac{\beta(x)}{\alpha(x)}$.

Thus, since $\frac{\lambda_1(x)}{\lambda_2(x)} > 0$, for all $x \in \Omega$, by (7.3), we get that (7.1) is equivalent to

$$\text{find} \qquad \min_{x\in\Omega} \frac{\lambda_2(x)}{-\lambda_1(x)}. \tag{7.6}$$

Finally, it is obvious that (7.6) is equivalent to

$$\text{find} \qquad \min_{x\in\Omega} \frac{\lambda_1(x) + \lambda_2(x)}{-\lambda_1(x)}.$$

So we have replaced problem 7.1 with the following one.

Problem : Given the smooth function

$$\begin{array}{rcll} h: & \Omega & \longrightarrow & \mathcal{S}\Re^n \times \mathcal{Q}^n, \\ & x & \longmapsto & (A(x), B(x)) \end{array}$$

let $f_k(x) = g_k(h(x))$, the sum of the k largest eigenvalues of $h(x)$, and let

$$\rho(x) = \frac{f_2(x)}{-f_1(x)}.$$

Then find

$$\min_{x \in \Omega} \rho(x). \tag{7.7}$$

Now, for $x \in \Omega$, $f_2(x) < 0$, $-f_1(x) > 0$, f_2 is regular at x by Proposition 6.3 and $-f_1$ is regular at x by 7.4. Hence we may apply lemma 7.2, case 2, to the function $\rho(x)$ and we have

$$\partial \rho(x) = \frac{f_2(x)\partial f_1(x) - f_1(x)\partial f_2(x)}{f_1^2(x)}. \tag{7.8}$$

We may apply the results of section 6 to the computation of the generalized gradient of $\rho(x)$. So let $G = G(x)$ be the Choleski factor of $B(x)$ and let Q be an orthonormal matrix which diagonalizes $G(x)^{-1}A(x)G(x)^{-T}$. Let t be the multiplicity of $\lambda_2(x)$, and let

$$Q = [\, Q_1 \,:\, Q_2 \,:\, Q_3 \,] \tag{7.9}$$

where $Q_1 = [q_1] \in \mathcal{O}^{n,1}$, $Q_2 \in \mathcal{O}^{n,t}$ and $Q_3 \in \mathcal{O}^{n,n-t-1}$. For $1 \le j \le m$, let $A_j(x)$ and $B_j(x)$ denote the partial derivatives of $A(x)$ and $B(x)$ respectively, and let $L_j = L_j(x)$ be the lower triangular matrix which solves the equation

$$L_j G^T + G L_j^T = B_j(x).$$

Write Z_j for the matrix

$$G^{-1}\left[A_j(x) - L_j G^{-1} A(x) - A(x) G^{-T} L_j^T\right] G^{-T}. \tag{7.10}$$

Then, theorem (6.4) gives

$$\partial f_1(x) = \left\{ w \in \Re^m \,\middle|\, w_j = < q_1 q_1^T, Z_j >,\ 1 \le j \le m \right\},$$

$$\partial f_2(x) = \left\{ v \in \Re^m \,\middle|\, v_j = < q_1 q_1^T, Z_j > + < Q_2 U_2 Q_2^T, Z_j >,\ 1 \le j \le m \right\},$$

where U_2 runs over all matrices in $\mathcal{S}_2^1(t)$, namely those symmetric $t \times t$ matrices U_2 with $0 \le U_2 \le I_t$ and $TrU_2 = 1$. The following theorem follows from equation (7.8).

Theorem 7.11 *The generalized gradient of ρ is given by*

$$\partial \rho(x) \;=\; \left\{ v \in \Re^m \,\middle|\, v_j = \left\langle \frac{\lambda_2(x) q_1 q_1^T - \lambda_1(x) Q_2 U_2 Q_2^T}{\lambda_1^2(x)}, Z_j \right\rangle \right\} \qquad (7.12)$$

$$\textit{for } 1 \leq j \leq m.$$

We also obtain the following characterization of the directional derivatives of $\rho(x)$. For $d \in \Re^m$, let

$$m(d) \;=\; \sum_{j=1}^{m} d_j q_1^T Z_j q_1 \in \Re, \qquad (7.13)$$

$$M(d) \;=\; \sum_{j=1}^{m} d_j Q_2^T Z_j Q_2 \in \mathcal{S}\Re^t. \qquad (7.14)$$

Proposition 7.15 *For $d \in \Re^m$, we have*

$$\rho'(x;d) = \frac{\lambda_2(x)}{\lambda_1^2(x)}\, m(d) - \frac{1}{\lambda_1(x)} \times (\text{the largest eigenvalue of } M(d)\ ).$$

Proof: As in Proposition 6.6, the directional derivative of ρ is given by

$$\rho'(x;d) = max\{ < v, d > \mid v \in \partial\rho(x) \}.$$

But

$$\begin{aligned} < v, d > \;&=\; \sum_{j=1}^{m} v_j d_j \\ &=\; \frac{\lambda_2(x)}{\lambda_1^2(x)}\, m(d) - \frac{1}{\lambda_1(x)} < U_2, M(d) > . \end{aligned}$$

Now, since $\frac{-1}{\lambda_1(x)} > 0$,

$$\rho'(x;d) = \frac{\lambda_2(x)}{\lambda_1^2(x)}\, m(d) - \frac{1}{\lambda_1(x)}\, max \left\{ < U_2, M(d) > \,\middle|\, U_2 \in \mathcal{S}_2^{k-r}(t) \right\},$$

and the result follows from section 3. □

8 The gap between the first two eigenvalues of a symmetric definite pencil

The discussion of the previous section on the ratio of the first two eigenvalues of the symmetric definite pencil $h(x) = (A(x), B(x))$, where

$$h : \Omega \longrightarrow \mathcal{S}\Re^n \times \mathcal{Q}^n,\ \Omega \subset \Re^m,$$

applies almost verbatim to the study of the gap between these eigenvalues. More precisely, if $\lambda_1(x) \geq \lambda_2(x)$ are the two largest eigenvalues of $h(x)$, we wish to maximize the gap between $\lambda_1(x)$ and $\lambda_2(x)$, or equivalently, to minimize the difference

$$f_2(x) - 2f_1(x) = \lambda_2(x) - \lambda_1(x).$$

We assume as before that hypotheses (7.3) and (7.4) hold so that $\lambda_1(x) < 0$ is simple, and hence smooth, for all x. We may then apply Clarke's result on the generalized gradient of a sum, ([7, Corollary 3, p. 40]), and our computations of the generalized gradient of $f_k(x)$ to obtain the following theorem.

Theorem 8.1 *Let $\eta(x) = f_2(x) - 2f_1(x)$. Then the generalized gradient of η is given by*

$$\partial\eta(x) \quad = \quad \left\{ v \in \Re^m \,\middle|\, v_j = \left\langle Q_2 U_2 Q_2^T - q_1 q_1^T, Z_j \right\rangle,\ \text{for } 1 \leq j \leq m \right\} \quad (8.2)$$

where U_2, Q_2 and Z_j, $1 \leq j \leq m$, are as in theorem (7.11).

Similarly, we obtain the characterization of the directional derivative of $\eta(x)$.

Proposition 8.3 *For $d \in \Re^m$, we have*

$$\eta'(x; d) = (\text{ the largest eigenvalue of } M(d)\ ) - m(d),$$

where

$$M(d) = \sum_{j=1}^{m} d_j Q_2^T Z_j Q_2.$$

$$m(d) = \sum_{j=1}^{m} d_j q_1^T Z_j q_1.$$

9 Optimality Conditions

We now describe optimality conditions for the constrained minimization of the functions $\rho(x)$ and $\eta(x)$ defined in the two previous sections.

Thus we let

$$\begin{array}{rcl} C=[c_1 \ldots c_m] & \in & \Re^{n_c \times m} \\ b & \in & \Re^{n_c} \\ l,u & \in & \Re^m \end{array}$$

and we consider the following two problems.

$$(P1) \qquad \begin{array}{l} \min\limits_x \ \rho(x) \\ subject \ \ to \\ \left\{ \begin{array}{ccccc} & Cx & = & b \\ l & \leq & x & \leq & u \end{array} \right. \end{array}$$

and

$$(P2) \qquad \begin{array}{l} \min\limits_x \ \eta(x) \\ subject \ \ to \\ \left\{ \begin{array}{ccccc} & Cx & = & b \\ l & \leq & x & \leq & u \end{array} \right. \end{array}$$

We assume that the hypotheses (7.3) and (7.4) hold.

We now apply the Lagrange multiplier rule, see [7, Theorem 6.1.1, p. 228], to derive optimality conditions for problems $(P1)$ and $(P2)$.

Theorem 9.1 *Let $x \in \Re^m$ be a feasible point for problems $(P1)$ and $(P2)$. Then*

1. *a necessary condition for x to solve $(P1)$ is that there exist a dual matrix $U_2 \in \mathcal{S}\Re^t$, where t is the multiplicity of λ_2 at x, and vectors of Lagrange multipliers $\alpha \in \Re^{n_c}$, and $\gamma \in \Re^m$ satisfying*

$$\frac{\lambda_2(x)}{\lambda_1^2(x)} q_1^T Z_j q_1 - \frac{1}{\lambda_1(x)} \left\langle U_2, Q_2^T Z_j Q_2 \right\rangle \ = \ < \alpha, c_j > + \gamma_j \qquad for \ \ 1 \leq j \leq m \quad (9.2)$$

$$Tr(U_2) \ = \ 1 \qquad (9.3)$$

$$0 \ \leq \ U_2 \ \leq \ I \qquad (9.4)$$

and

$$\left\{ \begin{array}{lll} \gamma_j = 0 & \mathit{if} & l_j < x_j < u_j \\ \gamma_j \geq 0 & \mathit{if} & x_j = l_j \\ \gamma_j \leq 0 & \mathit{if} & x_j = u_j \end{array} \right. \tag{9.5}$$

2. *similarly a necessary condition for x to be a solution of $(P2)$ is that there exist U_2, α, and γ as above satisfying all of the above equations but with (9.2) replaced by*

$$\left\langle U_2, Q_2^T Z_j Q_2 \right\rangle - < q_1 q_1^T, Z_j > \;=\; < \alpha, c_j > + \gamma_j, \qquad \mathit{for}\ 1 \leq j \leq m \tag{9.6}$$

Here the matrices U_2 and Z_j are as in theorems (7.11) and (8.1).

We devote the remainder of this section to showing how to derive a descent direction for the functions $\rho(x)$ and $\eta(x)$ in the case that all the optimality conditions are satisfied by U_2, α, and γ except possibly for condition (9.4).

Theorem 9.7 *Suppose that x, U_2, α, γ satisfy conditions (9.2), (9.3) and (9.5). Let θ be an eigenvalue of U_2 with normalized eigenvector v, and let $s \in \Re$. If $d \in \Re^m$, $\delta \in \Re$ solve the following system of equations*

$$\sum_{j=1}^{m} d_j Q_2^T U_2 Q_2 - \delta I = s v v^T \tag{9.8}$$

$$Cd = 0 \tag{9.9}$$

$$d_j = 0 \quad \mathit{if} \quad x_j = l_j \ \mathit{or}\ x_j = u_j \tag{9.10}$$

then d is a feasible direction for $(P1)$ and $(P2)$ with directional derivatives

1.

$$\rho'(x;d) = \left\{ \begin{array}{ll} \frac{s(\theta-1)}{\lambda_1(x)} & \mathit{if} \quad s > 0 \\ \\ \frac{s\theta}{\lambda_1(x)} & \mathit{if} \quad s \leq 0 \end{array} \right.$$

2.

$$\eta'(x;d) = \left\{ \begin{array}{ll} -s(\theta - 1) & \mathit{if} \quad s > 0 \\ -s\theta & \mathit{if} \quad s \leq 0 \end{array} \right.$$

Proof: Recall the definitions of the matrix $M(d) = \sum_{j=1}^{m} d_j Q_2^T Z_j Q_2$ and the number $m(d) = \sum_{j=1}^{m} d_j q_1^T Z_j q_1$ (see (7.14) and (7.13)). Equation (9.8) gives

$$M(d) = \delta I + svv^T,$$

so that all the eigenvalues of $M(d)$ are equal to δ except for one which equals $\delta + s$. Let $l = max\{\delta, \delta + s\}$. Now, taking the inner product of equation (9.8) with the matrix U_2 yields

$$\sum_{j=1}^{m} d_j \left\langle U_2, Q_2^T Z_j Q_2 \right\rangle - \delta \left\langle U_2, I \right\rangle = \left\langle U_2, svv^T \right\rangle,$$

or, equivalently,

$$\sum_{j=1}^{m} d_j \left\langle U_2, Q_2^T Z_j Q_2 \right\rangle - \delta Tr(U_2) = s\theta. \qquad (9.11)$$

Recall from Proposition 7.15 that

$$\rho'(x; d) = \frac{\lambda_2(x)}{\lambda_1^2(x)} m(d) - \frac{l}{\lambda_1(x)}.$$

If the optimality conditions (9.2) and (9.3) are satisfied, equation (9.11) becomes

$$\sum_{j=1}^{m} d_j \left(\frac{\lambda_2(x)}{\lambda_1(x)} q_1^T Z_j q_1 - \lambda_1(x) (< \alpha, c_j > + \gamma_j) \right) - \delta = s\theta,$$

i.e.

$$\frac{\lambda_2(x)}{\lambda_1(x)} m(d) - \delta - \lambda_1(x) \left(\alpha^T C d + \gamma^T d \right) = s\theta.$$

But $Cd = 0$ by assumption, and moreover, if optimality condition (9.5) is satisfied, then $\gamma^T d = 0$ as well. Hence

$$\frac{\lambda_2(x)}{\lambda_1(x)} m(d) - \delta = s\theta,$$

and we conclude that

$$\begin{aligned} \rho'(x; d) &= \frac{\lambda_2(x)}{\lambda_1^2(x)} m(d) - \frac{l}{\lambda_1(x)} \\ &= \begin{cases} \frac{s(\theta - 1)}{\lambda_1(x)} & \text{if} \quad s > 0 \\ \frac{s\theta}{\lambda_1(x)} & \text{if} \quad s \leq 0. \end{cases} \end{aligned}$$

We consider the function $\eta(x)$ next. Recall from Proposition 8.3 that

$$\eta'(x; d) = l - m(d).$$

Equation (9.11) becomes

$$\sum_{j=1}^{m} d_j [q_1^T Z_j q_1 + < \alpha, c_j > + \gamma_j] - \delta = s\theta,$$

or equivalently

$$m(d) = \alpha^T C d + \gamma^T d - \delta = s\theta,$$

and we get

$$m(d) - \delta = s\theta.$$

We conclude that

$$\eta'(x; d) = \begin{cases} -s(\theta - 1) & \text{if} \quad s > 0 \\ -s\theta & \text{if} \quad s \leq 0. \end{cases}$$

□

It is now easy to generate a descent direction for $\rho(x)$ or $\eta(x)$ in the case the matrix U_2 does not satisfy

$$0 \leq U_2 \leq I.$$

Indeed, U_2 must have an eigenvalue θ falling outside the interval $[0, 1]$. If $\theta < 0$, we simply choose s to be negative, say $s = -1$, while if $\theta > 1$, we choose s to be positive, e.g. $s = 1$.

10 Perturbations of a disk

We now describe our families of perturbations of a disk. Consider a disk D centered at the origin (for simplicity), and of fixed area. We write D as the union of a disk D_0 of radius R_0 and an annulus D_1 of width R_1. We divide D_1 into sectors of angle $\theta = 2\pi/M$ and we write D_1 as a union of isoparametric quadrilateral elements with 8 nodes as in the following picture.

where one rectangular element is shaded and the nodes are displayed. The disk D_0 is triangulated with isoparametric triangle of type 2 [5]. Now the nodes of the rectangular elements are all of the form

$$(R_0 + \frac{k}{n} R_1)(\cos m\theta, \sin m\theta)$$

where n is an even integer equal to twice the number of "layers" of rectangles in D_1 ($n = 4$, $M = 16$ in the above picture), k is an integer between 0 and n specifying a node along a single ray, θ is the angle of a single sector and m is an integer between 0 and $M - 1$ specifying a single sector.

We perturb the disk D by perturbing the annulus D_1 while leaving the disk D_0 fixed as follows.

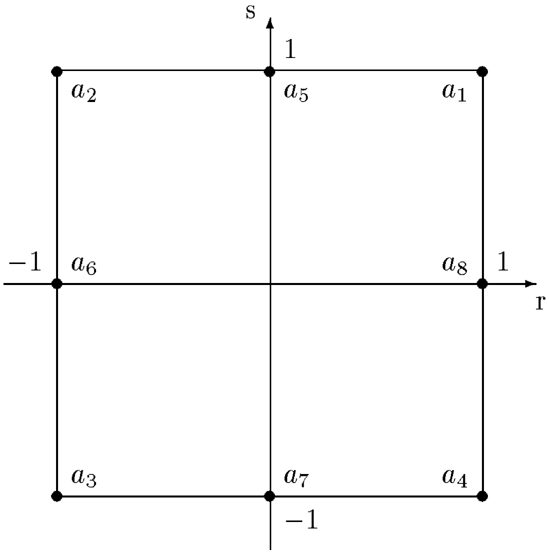
for each ray we introduce a parameter x_m and we perturb the nodes by

$$(R_0 + x_m \frac{k}{n} R_1)(\cos m\theta, \sin m\theta);$$

the parameter x_m is taken to lie in a small neighborhood of 1 (not containing 0!).

Thus, writing $x = (x_0, \ldots, x_{M-1})$, we obtain a family $\Omega(x)$ of deformations of our original disk D, parametrized by x lying in a neighborhood of $(1, \ldots, 1)$ in $\Re^m$. We need to derive sufficient conditions on the values of R_0, R_1, the angle θ, and the parameters x_i, to insure the nondegeneracy of the rectangular elements. To this end, we describe in some details the construction of these elements.

Consider the standard restricted biquadratic element $\hat{K}$ (see [5])



equipped with the subspace of Q_2 spanned by the following basis dual to the nodes $a_1, \ldots, a_8$. (Recall that Q_2 is the space of polynomials in r, s spanned by the monomials $r^n s^m$ with $0 \leq n, m \leq 2$ [5]).

$$\begin{aligned}
a_1 = (1,1) \ : \ \lambda_1 &= \frac{1}{4}[(1+r)(1+s) - (1-r^2)(1+s) - (1+r)(1-s^2)] \\
a_2 = (-1,1) \ : \ \lambda_2 &= \frac{1}{4}[(1-r)(1+s) - (1-r^2)(1+s) - (1-r)(1-s^2)] \\
a_3 = (-1,-1) \ : \ \lambda_3 &= \frac{1}{4}[(1-r)(1-s) - (1-r^2)(1-s) - (1-r)(1-s^2)]
\end{aligned}$$

$$\begin{aligned}
a_4 = (1,-1) \,:\, \lambda_4 &= \frac{1}{4}[(1+r)(1-s) - (1-r^2)(1-s) - (1+r)(1-s^2)] \\
a_5 = (0,1) \,:\, \lambda_5 &= \frac{1}{2}(1-r^2)(1+s) \\
a_6 = (-1,0) \,:\, \lambda_6 &= \frac{1}{2}(1-r)(1-s^2) \\
a_7 = (0,-1) \,:\, \lambda_7 &= \frac{1}{2}(1-r^2)(1-s) \\
a_8 = (1,0) \,:\, \lambda_8 &= \frac{1}{2}(1+r)(1-s^2)
\end{aligned}$$

If we let a_9 denote $(0,0)$, then $\{\lambda_1, \ldots, \lambda_8\}$ spans the subspace Q_2' of Q_2 consisting of those $p(r,s) \in Q_2$ such that

$$\chi(p) = 0$$

where

$$\chi(p) \equiv 4p(a_9) + \sum_{i=1}^{4} p(a_i) - 2\sum_{i=5}^{8} p(a_i).$$

Then the corresponding isoparametric elements K are constructed as follows. Given 8 points $b^1, \ldots, b^8$ in $\Re^2$, we define the map

$$\begin{aligned}
F : \quad \hat{K} \quad &\longrightarrow \quad \Re^2 \\
(r,s) \quad &\longmapsto \quad \sum_{1}^{8} \lambda_j(r,s) b^j
\end{aligned}$$

and we set $K = F(\hat{K})$. The element K is nondegenerate if the map F is invertible. Let us write $F = (F_1, F_2)$ where

$$F_i(r,s) = \sum_{j=1}^{8} \lambda_j(r,s) b_i^j, \; i = 1,2.$$

Then the rows of the jacobian matrix JF of F are the gradients of F_1, F_2,

$$\nabla F_i = \sum_{j=1}^{8} \nabla \lambda_j b_i^j, \; i = 1,2.$$

We compute the gradients $\nabla \lambda_j$ for $1 \le j \le 8$.

$$\nabla \lambda_1 = \left(\frac{1}{4}(1+s)(2r+s), \frac{1}{4}(1+r)(2s+r)\right)$$

$$
\begin{aligned}
\nabla\lambda_2 &= \left(\frac{1}{4}(1+s)(2r-s), \frac{1}{4}(1-r)(2s-r)\right)\\
\nabla\lambda_3 &= \left(\frac{1}{4}(1-s)(2r+s), \frac{1}{4}(1-r)(2s+r)\right)\\
\nabla\lambda_4 &= \left(\frac{1}{4}(1-s)(2r-s), \frac{1}{4}(1+r)(2s-r)\right)\\
\nabla\lambda_5 &= \left(-r(1+s), \frac{1}{2}(1-r^2)\right)\\
\nabla\lambda_6 &= \left(-\frac{1}{2}(1-s^2), -s(1-r)\right)\\
\nabla\lambda_7 &= \left(-r(1-s), -\frac{1}{2}(1-r^2)\right)\\
\nabla\lambda_8 &= \left(\frac{1}{2}(1-s^2), -s(1+r)\right)
\end{aligned}
$$

Now let K denote one of the elements of the ring D_1. The nodes b^j, $1 \leq j \leq 8$, of the perturbed K are of the form

$$
\begin{aligned}
b^1 &= (R_0 + \alpha\left(\frac{k+1}{n}\right) R_1)(\cos\theta_1, \sin\theta_1)\\
b^8 &= (R_0 + \alpha\left(\frac{k}{n}\right) R_1)(\cos\theta_1, \sin\theta_1)\\
b^4 &= (R_0 + \alpha\left(\frac{k-1}{n}\right) R_1)(\cos\theta_1, \sin\theta_1)\\
\\
b^2 &= (R_0 + \beta\left(\frac{k+1}{n}\right) R_1)(\cos\theta_2, \sin\theta_2)\\
b^6 &= (R_0 + \beta\left(\frac{k}{n}\right) R_1)(\cos\theta_2, \sin\theta_2)\\
b^3 &= (R_0 + \beta\left(\frac{k-1}{n}\right) R_1)(\cos\theta_2, \sin\theta_2)\\
\\
b^5 &= (R_0 + \gamma\left(\frac{k+1}{n}\right) R_1)(\cos\left(\frac{\theta_1+\theta_2}{2}\right), \sin\left(\frac{\theta_1+\theta_2}{2}\right))\\
b^7 &= (R_0 + \gamma\left(\frac{k-1}{n}\right) R_1)(\cos\left(\frac{\theta_1+\theta_2}{2}\right), \sin\left(\frac{\theta_1+\theta_2}{2}\right))
\end{aligned}
$$

where α, β, γ are the deformation parameters, k is an odd integer, $0 \leq k \leq n$.

The jacobian determinant *det JF* of the map F is given by

$$\left(\sum_{j=1}^{8}(\lambda_j)_r b_1^j\right)\left(\sum_{j=1}^{8}(\lambda_j)_s b_2^j\right) - \left(\sum_{j=1}^{8}(\lambda_j)_r b_2^j\right)\left(\sum_{j=1}^{8}(\lambda_j)_s b_1^j\right)$$

where $(\lambda_j)_r$, $(\lambda_j)_s$ denote the partial derivatives of λ_j with respect to r, s respectively. Let $\phi = (\theta_2 - \theta_1)/2$. We have the following theorem whose proof is given in the appendix.

Theorem 10.1 *The jacobian determinant det JF of the map F is nonzero for all $R_0 > 0$, $R_1 > 0$ and $\alpha, \beta, \gamma, \phi$ such that*

$$\left\{\begin{array}{rcl} \mid \alpha - 1 \mid & \leq & \epsilon \\ \mid \beta - 1 \mid & \leq & \epsilon \\ \mid \gamma - 1 \mid & \leq & \epsilon \end{array}\right.$$

where

$$\epsilon = \frac{1}{2}\frac{1-\cos\phi}{1+\cos\phi}$$

and $3/5 < \cos\phi < 1$.

Remark 10.2

1. *The condition on the angle ϕ gives*

$$0 < \phi < \text{ (approximately) } 53^\circ.$$

2. *The corresponding range of values for ϵ are :*

$$\begin{array}{rll} \epsilon = & 0.2 & \textit{for } \cos\phi = 3/5 \\ \epsilon = & 0 & \textit{for } \cos\phi = 1. \end{array}$$

The above theorem shows that all the quadrilateral elements of the ring D_1 are nondegenerate if M is chosen larger than 8, and if the parameters x_i, $0 \leq M - 1$, satisfy $\mid x_i - 1 \mid \leq \epsilon$, where

$$\epsilon = \frac{1}{2}\frac{1-\cos\frac{2\pi}{M}}{1+\cos\frac{2\pi}{M}}.$$

11 Description of the Algorithm

We present the algorithm we used to minimize the function the ratio λ_1/λ_2 of the two smallest eigenvalues of the Laplacian as a function of the shape of the domain Ω. This is a suitably modified version of an algorithm developed by M. Overton [16]. We briefly describe the changes that need to be made and refer the reader to [16] for a detailed account.

First of all, we consider the family $\Omega(x)$, $x \in \Re^m$, of deformations of a disk constructed in section 10. The regions $\Omega(x)$ are equipped with a triangulation made of isoparametric rectangles and triangles of type 2 which is uniform in x. By this we mean that the number of elements as well as their relative position with respect to one another is independent of the choice of x. Then, standard finite element techniques allow us to transform the problem of computing the eigenvalues of the Laplacian on $\Omega(x)$ into the following finite dimensional generalized eigenvalue problem [5,6,19]. Compute the eigenvalues of the symmetric definite pencil $(A(x), B(x))$, where $A(x)$ and $B(x)$ are the stiffness and mass matrix , respectively, of the Laplacian associated to the given triangulation. $(A(x), B(x))$ is a smooth function of x, thanks to the independence of the triangulation upon the parameters.

Thus, following our discussion in section 7, we attempt to minimize the function

$$\rho(x) = -\frac{\lambda_1(x) + \lambda_2(x)}{\lambda_1(x)}$$

where $\lambda_2(x) \leq \lambda_1(x) < 0$ are the two largest eigenvalues of the pencil $(-A(x), B(x))$. The basic strategy is to generate a sequence of iterates x^ν converging to a (local) minimizer of $\rho(x)$. Hence, starting with an initial guess x^0 near $(1, \ldots, 1)$, so that the corresponding region $\Omega(x)$ is near the disk, the algorithm ought to produce a sequence x^ν converging to $(1, \ldots, 1)$. Now the point $x^{\nu+1}$ is obtained as $x^\nu + d$ and the primary task of the algorithm is to generate the step d. The scheme implemented by Overton is to get d by partially solving a linear program whose constraints are derived from the optimality conditions for a minimizer of the function to be optimized (see [16] for an explanation of partial solution of the linear program).

In our particular situation, the linear program that needs to be solved has the following form. It depends upon the multiplicity t^* of λ_2 at the solution. We compute an estimate t of t^* based on our knowledge of the current iterate x^ν and in terms of a tolerance τ as follows.

$$\lambda_2(x^\nu) - \lambda_{t+1}(x^\nu) \quad \leq \quad \tau\, max\{1, |\lambda_2(x^\nu)|\}$$

$$\lambda_2(x^\nu) - \lambda_{t+2}(x^\nu) \quad > \quad \tau\, max\{1, |\lambda_2(x^\nu)|\}.$$

We also normalize the problem of finding

$$\min_x \rho(x)$$

by requiring that the area of the region $\Omega(x)$ remains constant. Thus we introduce a linear constraint on d of the form $c^T d = 0$, where the vector c is the gradient of the area of $\Omega(x^\nu)$ with respect to x. Also let $\hat{q}_i = q_i(x^\nu)$, $1 \leq i \leq n$, form an orthonormal basis of eigenvectors for $(A(x^\nu), B(x^\nu))$, and let $\hat{Z}_j = Z_j(x^\nu)$ be as in (7.10). We write $\hat{Q}_2 = [\hat{q}_2 \ldots \hat{q}_{1+t}]$, $\hat{\lambda}_1 = \lambda_1(\hat{x})$, $\hat{\lambda}_2 = \lambda_2(\hat{x})$, and $\hat{A}_j = A_j(x^\nu)$ for the j^{th}-partial derivative of $A(x)$ at x^ν. The linear program is then given by

$$(LP^\nu) \qquad \min_{d,\delta} \frac{\hat{\lambda}_2}{\hat{\lambda}_1^2} \sum_{j=1}^{m} d_j \hat{q}_1^T \hat{Z}_j \hat{q}_1 - \frac{\delta}{\hat{\lambda}_1}$$

subject to

$$\delta I - \sum_{j=1}^{m} d_j \hat{Q}_2^T \hat{Z}_j \hat{Q}_2 = diag(0, \hat{\lambda}_3 - \hat{\lambda}_2, \ldots, \hat{\lambda}_{1+t} - \hat{\lambda}_2) \qquad (11.1)$$

$$\delta - \sum_{j=1}^{m} d_j \hat{q}_1^T \hat{Z}_j \hat{q}_1 \leq \hat{\lambda}_1 - \hat{\lambda}_2 \qquad (11.2)$$

$$\delta - \sum_{j=1}^{m} d_j \hat{q}_k^T \hat{Z}_j \hat{q}_k \geq \hat{\lambda}_k - \hat{\lambda}_2,\ 2 + t \leq k \leq n \qquad (11.3)$$

$$c^T d = 0 \qquad (11.4)$$

$$l \leq d \leq u \qquad (11.5)$$

$$||d||_\infty \leq r \qquad (11.6)$$

where r is a trust region radius updated by the algorithm, and l, u are determined by the neighborhood of $(1, \ldots, 1)$ in $\Re^m$ over which $\rho(x)$ is to be optimized. The set of constraints (11.1) arise from a linearization of a suitable system of nonlinear equations characterizing the conditions

$$\lambda_2(x) = \ldots = \lambda_{1+t}(x) = \omega;$$

constraints (11.2) and (11.3) ensure that, up to first order, the values of λ_1 and $\lambda_{2+t}, \ldots, \lambda_n$ at the new iterate are still greater, respectively less, than that of λ_2. Note that the constraints (11.1) form a system of $t(t+1)/2$ linear equations; each equation has a Lagrange multiplier associated to it. These Lagrange multipliers are assembled into a symmetric matrix U with diagonal entries corresponding to the multipliers of the diagonal equations of (11.1) and the off-diagonal elements of U corresponding to half the multipliers of the corresponding off-diagonal equations of (11.1). (The factor 1/2 comes from our definition of the inner product $<,>$). The symmetric matrix U provides an estimate for the dual matrix U_2 described in section 9. More precisely, we have

$$U_2 = -\lambda_1(x^\nu)U.$$

12 Numerical Results

We implemented the algorithm described above in Fortran 77. The program includes a subroutine written by Overton to partially solve the linear program presented in the previous section. The first few eigenvalues and eigenvectors of the pencil $(A(x), B(x))$ are computed via subspace iteration as described in [4], and subroutines from the Linpack [10] and Eispack [12] libraries are used to solve the resulting linear systems and reduced generalized eigenvalue problems. The bulk of the code that was written specifically for this problem is devoted to implementing the parametrization of the shape Ω as a function of a vector x of parameters as described in section 10 and the computation of the matrices $A(x)$ and $B(x)$ as well as their partial derivatives with respect to the parameters x_i.

In order to test the quality of our discrete model of the geometry we used our code to verify numerically the celebrated Faber-Krahn inequality, namely that among all regions of the plane of a given area the disk is the shape with the smallest first eigenvalue for the Laplacian. Of course, this problem is much simpler than the Payne, Polya, Weinberger conjecture since the function to be minimized, namely $\lambda_1(x)$, is smooth at the solution. We set the values of the radius of the disk D_0 and the width of the annulus D_1 both equal to 0.6. The annulus was divided into 32 sectors, so that the region Ω depends upon 32 parameters. There were four layers of rectangles in D_1 for a total of 64 rectangles, and there were 64 triangles in D_0. The total number of vertices was 337, including the boundary vertices, resulting in banded matrices $A(x)$ and $B(x)$ of size 305 by 305 with a bandwidth equal to 59. Finally the initial value of the trust radius was set to 0.1. The algorithm performed very well, converging to a disk in a few steps. For example, starting with an ellipse as the initial shape, all 32 parameters agreed to four significant digits after 10 iterations. (Recall that in our model disks are represented by setting all parameters equal to a common value). On the other hand, the next 15 iterations failed to produce agreement on the next digit. These results show that our discretization of the geometry is adequate, but at the same time reveal the limited 'resolution' of our model, namely only the first four or five digits of the parameters are geometrically significant for the triangulation of Ω given above.

The behavior of the algorithm on the main problems, namely that of minimizing the ratio $\frac{\lambda_1(x)}{\lambda_2(x)}$ or maximizing the gap between $\lambda_1(x)$ and $\lambda_2(x)$ was much less satisfying. First, recall that the parameters with an even index, x_{2i}, correspond to rays that bisect the rectangles in the annulus D_1, while the odd indexed parameters x_{2i+1} correspond to rays that coincide with the sides of the rectangles (see section 10). We worked with the same triangulation of $\Omega(x)$ as that used for the Faber-Krahn inequality and we used various initial shapes and several values for the multiplicity threshold τ (see section 11). In all cases, we observed the following pattern as the algorithm proceeds. At startup the multiplicity of λ_2 is 1 and the algorithm makes moderate progress, that is the values of the parameters converge to a common value but very slowly, the objective function, either $\rho(x)$ or $\eta(x)$, is reduced by smaller and smaller steps, and the trust radius is steadily decreased. At the same time the gap between λ_2 and λ_3 narrows until the threshold τ forces the multiplicity of λ_2 to jump to 2. At this point the algorithm goes through a few iterations without succeeding in producing a decrease in the objective function. For each such failure the trust radius is halved, and the algorithm eventually produces a descent direction. Then the following happens for the next few iterations. The objective function is being steadily reduced, but the parameters no longer converge to a common value. Instead, the even indexed parameters are being reduced, while the odd ones are being increased. Thus $\Omega(x)$ resembles a polygon whose sides are caving in more and more toward the interior of the polygon. At the same time the gap between λ_2 and λ_3 widens and the trust radius is steadily increasing. (At each iteration the trust radius is either doubled, halved, or left unchanged, depending upon the ratio of the actual increase in the objective function to the first order estimate of this increase; when the multiplicity is 2 this ratio is usually close to 1 and the trust radius is doubled, whereas when the multiplicity is 1 this ratio is small and the trust radius is halved). Eventually, the threshold τ forces the multiplicity back to 1, and the whole cycle starts again. The algorithm oscillates between these two opposite behaviors for a while, the actual number of cycles depending upon the value of the convergence tolerance (the algorithm stops when the norm of the step produced falls below a convergence threshold, and when the optimality conditions are satisfied, of course).

At present we still do not understand what is happening. The first guess would be that there remains a bug in the program, but extensive checking and testing of the code has failed to produce the source of the problem so far. It is also possible that our geometric model, while adequate for the optimization of $\lambda_1(x)$, is not sensitive enough for the ratio $\rho(x)$ or the gap $\eta(x)$. Finally, but least likely, the behavior of the algorithm may actually reflect wild and unpleasant properties of $\rho(x)$ and $\eta(x)$ near their minimum.

A Appendix

In this appendix we give the proof of theorem 10.1. So let K denote the isoparametric rectangular element of the ring D_1 with nodes

$$\begin{aligned}
b^1 &= (R_0 + \alpha\left(\frac{k+1}{n}\right)R_1)(1,0)\\
b^8 &= (R_0 + \alpha\left(\frac{k}{n}\right)R_1)(1,0)\\
b^4 &= (R_0 + \alpha\left(\frac{k-1}{n}\right)R_1)(1,0)\\
\\
b^2 &= (R_0 + \alpha\left(\frac{k+1}{n}\right)R_1)(\cos\theta, \sin\theta)\\
b^6 &= (R_0 + \alpha\left(\frac{k}{n}\right)R_1)(\cos\theta, \sin\theta)\\
b^3 &= (R_0 + \alpha\left(\frac{k-1}{n}\right)R_1)(\cos\theta, \sin\theta)\\
\\
b^5 &= (R_0 + \alpha\left(\frac{k+1}{n}\right)R_1)(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})\\
b^7 &= (R_0 + \alpha\left(\frac{k-1}{n}\right)R_1)(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})
\end{aligned}$$

where n is a positive even integer and k is an odd integer with $0 < k < n$ and let

$$F(r,s) = \sum_1^8 \lambda_j(r,s)b^j$$

be the map sending the standard rectangle $\hat{K}$ onto K (cf section 10). We are going to determine bounds on the values of θ and α, β, γ which will ensure the nonvanishing of the jacobian determinant, $det\, JF$, of F. These conditions will apply also to the other elements of D_1 since nondegeneracy is clearly rotation invariant. So far we have $0 < \theta < 2\pi$, and we assume that

$$\mid \alpha - 1 \mid, \mid \beta - 1 \mid, \mid \gamma - 1 \mid \leq \epsilon$$

for some ϵ to be determined.

The following abbreviations will be used below.

$$A \;=\; R_0 + \alpha\frac{k}{n}R_1 + \alpha\frac{R_1}{n}s$$

$$
\begin{aligned}
B &= R_0 + \beta\frac{k}{n}R_1 + \beta\frac{R_1}{n}s \\
C &= R_0 + \gamma\frac{k}{n}R_1 + \gamma\frac{R_1}{n}s
\end{aligned}
$$

We need to compute the following expressions.

$$
\begin{aligned}
(\lambda_1)_r + (\lambda_8)_r + (\lambda_4)_r &= \frac{1}{2}(2r+1) \\
(\lambda_1)_r - (\lambda_4)_r &= \frac{1}{2}s(2r+1) \\
(\lambda_2)_r + (\lambda_6)_r + (\lambda_3)_r &= \frac{1}{2}(2r-1) \\
(\lambda_2)_r - (\lambda_3)_r &= \frac{1}{2}s(2r-1) \\
(\lambda_5)_r + (\lambda_7)_r &= -2r \\
(\lambda_5)_r - (\lambda_7)_r &= -2rs \\
(\lambda_1)_s + (\lambda_8)_s + (\lambda_4)_s &= 0 \\
(\lambda_1)_s - (\lambda_4)_s &= \frac{1}{2}r(1+r) \\
(\lambda_2)_s + (\lambda_6)_s + (\lambda_3)_s &= 0 \\
(\lambda_2)_s - (\lambda_3)_s &= -\frac{1}{2}r(1-r) \\
(\lambda_5)_s + (\lambda_7)_s &= 0 \\
(\lambda_5)_s - (\lambda_7)_s &= 1 - r^2
\end{aligned}
$$

We now compute the four entries of the jacobian matrix JF.

(1,1)-entry :

$$
\sum_{j=1}^{8}(\lambda_j)_r b_1^j = \frac{1}{2}A(2r+1) + \frac{1}{2}B(2r-1)\cos\theta - 2Cr\cos\frac{\theta}{2}
$$

(2,2)-entry :

$$
\sum_{j=1}^{8}(\lambda_j)_s b_2^j = -\frac{1}{2}\beta\frac{R_1}{n}r(1-r)\sin\theta + \gamma\frac{R_1}{n}(1-r^2)\sin\frac{\theta}{2}
$$

(2,1)-entry :

$$
\sum_{j=1}^{8}(\lambda_j)_r b_2^j = \frac{1}{2}B(2r-1)\sin\theta - 2Cr\sin\frac{\theta}{2}
$$

(1,2)-entry :

$$\sum_{j=1}^{8}(\lambda_j)_s b_1^j = \frac{1}{2}\alpha\frac{R_1}{n}r(1+r) - \frac{1}{2}\beta\frac{R_1}{n}r(1-r)\cos\theta + \gamma\frac{R_1}{n}(1-r^2)\cos\frac{\theta}{2}$$

Now :

$$[\frac{1}{2}A(2r+1) + \frac{1}{2}B(2r-1)\cos\theta - 2Cr\cos\frac{\theta}{2}] \times$$
$$[-\frac{1}{2}\beta\frac{R_1}{n}r(1-r)\sin\theta + \gamma\frac{R_1}{n}(1-r^2)\sin\frac{\theta}{2}] =$$

$$\begin{aligned} = \; & -\frac{1}{4}A\beta\frac{R_1}{n}r(1-r)(2r+1)\sin\theta + \frac{1}{2}A\gamma\frac{R_1}{n}(2r+1)(1-r^2)\sin\frac{\theta}{2} \\ & -\frac{1}{4}B\beta\frac{R_1}{n}r(1-r)(2r-1)\cos\theta\sin\theta \\ & +\frac{1}{2}B\gamma\frac{R_1}{n}(2r-1)(1-r^2)\cos\theta\sin\frac{\theta}{2} \\ & +C\beta\frac{R_1}{n}r^2(1-r)\cos\frac{\theta}{2}\sin\theta - 2C\gamma\frac{R_1}{n}r(1-r^2)\cos\frac{\theta}{2}\sin\frac{\theta}{2}. \end{aligned}$$

and

$$[\frac{1}{2}\alpha\frac{R_1}{n}r(1+r) - \frac{1}{2}\beta\frac{R_1}{n}r(1-r)\cos\theta + \gamma\frac{R_1}{n}(1-r^2)\cos\frac{\theta}{2}] \times$$
$$[\frac{1}{2}B(2r-1)\sin\theta - 2Cr\sin\frac{\theta}{2}] =$$

$$\begin{aligned} = \; & \frac{1}{4}B\alpha\frac{R_1}{n}r(1+r)(2r-1)\sin\theta - C\alpha\frac{R_1}{n}r^2(1+r)\sin\frac{\theta}{2} \\ & -\frac{1}{4}B\beta\frac{R_1}{n}r(1-r)(2r-1)\cos\theta\sin\theta \\ & +C\beta\frac{R_1}{n}r^2(1-r)\cos\theta\sin\frac{\theta}{2} + \frac{1}{2}B\gamma\frac{R_1}{n}(2r-1)(1-r^2)\cos\frac{\theta}{2}\sin\theta \\ & -2C\gamma\frac{R_1}{n}r(1-r^2)\cos\frac{\theta}{2}\sin\frac{\theta}{2}. \end{aligned}$$

Thus we have

$$\begin{aligned} det\, JF \; = \; & -\frac{1}{4}A\beta\frac{R_1}{n}r(1-r)(2r+1)\sin\theta + \frac{1}{2}A\gamma\frac{R_1}{n}(2r+1)(1-r^2)\sin\frac{\theta}{2} \\ & +\frac{1}{2}B\gamma\frac{R_1}{n}(2r-1)(1-r^2)\cos\theta\sin\frac{\theta}{2} \end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}B\gamma\frac{R_1}{n}(2r-1)(1-r^2)\cos\frac{\theta}{2}\sin\theta \\
&+C\beta\frac{R_1}{n}r^2(1-r)\cos\frac{\theta}{2}\sin\theta - C\beta\frac{R_1}{n}r^2(1-r)\cos\theta\sin\frac{\theta}{2} \\
&-\frac{1}{4}B\alpha\frac{R_1}{n}r(1+r)(2r-1)\sin\theta + C\alpha\frac{R_1}{n}r^2(1+r)\sin\frac{\theta}{2}.
\end{aligned}
$$

But

$$
\begin{gathered}
\cos\theta\sin\frac{\theta}{2} - \cos\frac{\theta}{2}\sin\theta = -\sin\frac{\theta}{2} \\
\cos\frac{\theta}{2}\sin\theta - \cos\theta\sin\frac{\theta}{2} = \sin\frac{\theta}{2} \\
\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}
\end{gathered}
$$

so that

$$
\begin{aligned}
det\, JF \;=\;& -\frac{1}{4}A\beta\frac{R_1}{n}\, r(1-r)(2r+1)\sin\theta - \frac{1}{4}B\alpha\frac{R_1}{n}\, r(1+r)(2r-1)\sin\theta \\
&+\frac{1}{2}A\gamma\frac{R_1}{n}(2r+1)(1-r^2)\sin\frac{\theta}{2} + C\alpha\frac{R_1}{n}\, r^2(1+r)\sin\frac{\theta}{2} \\
&-\frac{1}{2}B\gamma\frac{R_1}{n}(2r-1)(1-r^2)\sin\frac{\theta}{2} + C\beta\frac{R_1}{n}\, r^2(1-r)\sin\frac{\theta}{2} \\
\\
=\;& \left[\frac{1}{2}\frac{R_1}{n}\sin\frac{\theta}{2}\right]\{-A\beta r(1-r)(2r+1)\cos\frac{\theta}{2} \\
&-B\alpha r(1+r)(2r-1)\cos\frac{\theta}{2} + A\gamma(2r+1)(1-r^2) \\
&+2C\alpha r^2(1+r) - B\gamma(2r-1)(1-r^2) + 2C\beta r^2(1-r)\}.
\end{aligned}
$$

Now, $\sin\frac{\theta}{2} \neq 0$ for $0 < \theta < 2\pi$, so that $det\, JF \neq 0$ if and only if $Q \neq 0$, where

$$
\begin{aligned}
Q \;=\;& A\gamma(2r+1)(1-r^2) - B\gamma(2r-1)(1-r^2) + 2C\alpha r^2(1+r) \\
&+2C\beta r^2(1-r) - \cos\frac{\theta}{2}\left[A\beta r(1-r)(2r+1) + B\alpha r(1+r)(2r-1)\right].
\end{aligned}
$$

We compute

$$
A\gamma(2r+1)(1-r^2) = \left(\gamma R_0 + \alpha\gamma\frac{k}{n}R_1 + \alpha\gamma\frac{R_1}{n}s\right)\left[-2r^3 - r^2 + 2r + 1\right]
$$

$$
-B\gamma(2r-1)(1-r^2) = \left(-\gamma R_0 - \beta\gamma\frac{k}{n}R_1 - \beta\gamma\frac{R_1}{n}s\right)\left[-2r^3 + r^2 + 2r - 1\right]
$$

$$\begin{aligned}
2C\alpha r^2(1+r) &= \left(2\alpha R_0 + 2\alpha\gamma\frac{k}{n}R_1 + 2\alpha\gamma\frac{R_1}{n}s\right)[r^3+r^2] \\
2C\beta r^2(1-r) &= \left(2\beta R_0 + 2\beta\gamma\frac{k}{n}R_1 + 2\beta\gamma\frac{R_1}{n}s\right)[-r^3+r^2] \\
A\beta r(1-r)(2r+1) &= \left(\beta R_0 + \alpha\beta\frac{k}{n}R_1 + \alpha\beta\frac{R_1}{n}s\right)[-2r^3+r^2+r] \\
B\alpha r(1+r)(2r-1) &= \left(\alpha R_0 + \alpha\beta\frac{k}{n}R_1 + \alpha\beta\frac{R_1}{n}s\right)[2r^3+r^2-r].
\end{aligned}$$

Thus we get the following expressions for the various terms of Q

$$\begin{aligned}
R_0\textbf{-term :}\quad & 2[r^3(\alpha-\beta)+r^2(\alpha+\beta-\gamma)+\gamma]. \\
\tfrac{k}{n}R_1\textbf{-term :}\quad & \gamma[\alpha(1+r)^2+\beta(1-r)^2]. \\
\tfrac{R_1}{n}s\textbf{-term :}\quad & \gamma[\alpha(1+r)^2+\beta(1-r)^2]. \\
\cos\tfrac{\theta}{2}\textbf{-term :}\quad & R_0[2r^3(\alpha-\beta)+r^2(\alpha+\beta)-r(\alpha-\beta)] \\
& +2\frac{k}{n}R_1\alpha\beta r^2 + 2\frac{R_1}{n}\alpha\beta s r^2.
\end{aligned}$$

Hence we may rewrite Q as

$$\begin{aligned}
Q &= 2R_0[r^3(\alpha-\beta)+r^2(\alpha+\beta-\gamma)+\gamma] \\
&\quad +\frac{k}{n}R_1\gamma[\alpha(1+r)^2+\beta(1-r)^2] + \frac{R_1}{n}s\gamma[\alpha(1+r)^2+\beta(1-r)^2] \\
&\quad -\cos\frac{\theta}{2}\left\{R_0[2r^3(\alpha-\beta)+r^2(\alpha+\beta)-r(\alpha-\beta)]\right. \\
&\quad \left. +2\frac{k}{n}R_1\alpha\beta r^2 + 2\frac{R_1}{n}\alpha\beta s r^2\right\} \\
&= m(r)s+b(r)
\end{aligned}$$

where

$$m(r) = \frac{R_1}{n}\left[\left(\gamma\alpha+\gamma\beta-2\cos\frac{\theta}{2}\alpha\beta\right)r^2 + 2\gamma(\alpha-\beta)r + \gamma(\alpha+\beta)\right]$$

and

$$\begin{aligned}
b(r) &= \left[2R_0(\alpha-\beta)(1-\cos\frac{\theta}{2})\right]r^3 \\
&\quad + \left[2R_0(\alpha+\beta-\gamma) + \frac{k}{n}R_1\gamma(\alpha+\beta) - \cos\frac{\theta}{2}R_0(\alpha+\beta) - 2\cos\frac{\theta}{2}\frac{k}{n}R_1\alpha\beta\right]r^2 \\
&\quad + \left[2\frac{k}{n}R_1\gamma(\alpha-\beta) + \cos\frac{\theta}{2}R_0(\alpha-\beta)\right]r + \left[2R_0\gamma + \frac{k}{n}R_1\gamma(\alpha+\beta)\right],
\end{aligned}$$

so that

$$b(r) = e(r) + km(r),$$

where

$$\begin{aligned} e(r) &= R_0 \left\{ 2(\alpha - \beta)(1 - \cos\frac{\theta}{2})r^3 + [2(\alpha + \beta - \gamma) - (\alpha + \beta)\cos\frac{\theta}{2}]r^2 \right. \\ &\quad \left. + [(\alpha - \beta)\cos\frac{\theta}{2}]r + 2\gamma \right\}. \end{aligned}$$

So we have obtained the following expression for Q

$$Q = m(r)s + e(r) + km(r).$$

Now $Q(r, s)$ vanishes for some $r, s \in [-1, 1]$ precisely when either of the following two conditions are met.

1. $\exists r \in [-1, 1]$ such that $m(r) = b(r) = 0$
2. $m(r) \neq 0$ and $s = -\frac{b(r)}{m(r)} \in [-1, 1]$.

Remark 1.1 We consider the special case $\alpha = \beta = \gamma = 1$, i.e. the case of the unperturbed element. Then

$$\begin{aligned} m(r) &= 2\frac{R_1}{n}[(1 - \cos\frac{\theta}{2})r^2 + 1] \\ e(r) &= 2R_0[(1 - \cos\frac{\theta}{2})r^2 + 1] \\ b(r) &= 2[(1 - \cos\frac{\theta}{2})r^2 + 1]\left(R_0 + k\frac{R_1}{n}\right). \end{aligned}$$

Clearly $m(r) > 0$ for all $r \in \Re$ and

$$\begin{aligned} s &= -\frac{2[(1 - \cos\frac{\theta}{2})r^2 + 1]\left(R_0 + \frac{k}{n}R_1\right)}{2\frac{R_1}{n}[(1 - \cos\frac{\theta}{2})r^2 + 1]} \\ &= -(n\frac{R_0}{R_1} + k). \end{aligned}$$

Since $0 < k < n$, we have

$$k + n\frac{R_0}{R_1} > 1, \text{ if } R_0 > 0,$$

so that Q does not vanish on $[-1, 1] \times [-1, 1]$.

Let us return to the general case. We proceed in two steps. First, we show how to ensure that $m(r) \neq 0$ for $r \in [-1,1]$. Second, we show how to force $-b(r)/m(r)$ to lie outside of $[-1,1]$ for $r \in [-1,1]$.

(1) First we force the coefficients of r^2 in $m(r)$ to be ≥ 0. (Observe that this is the case when $\alpha = \beta = \gamma = 1$). Thus we want

$$\gamma\alpha + \gamma\beta - 2\cos\frac{\theta}{2}\alpha\beta \geq 0$$

or equivalently

$$\gamma\frac{(\alpha+\beta)}{\alpha\beta} \geq 2\cos\frac{\theta}{2}.$$

Let $g(\alpha,\beta,\gamma) \equiv \gamma\frac{(\alpha+\beta)}{\alpha\beta}$. We have

$$\begin{aligned} g_\alpha &= -\frac{\gamma}{\alpha^2} \\ g_\beta &= -\frac{\gamma}{\beta^2} \text{ (by symmetry)} \\ g_\gamma &= \frac{\alpha+\beta}{\alpha\beta}. \end{aligned}$$

Let $\mathcal{C}$ denote the cube in $\Re^3$ given by

$$\mathcal{C} = \{(\alpha,\beta,\gamma) | \mid \alpha - 1 \mid, \mid \beta - 1 \mid, \mid \gamma - 1 \mid \leq \epsilon\}$$

for some $\epsilon > 0$ to be determined. Clearly, g does not have any critical points in $\mathcal{C}$, so to find the minimum of g on $\mathcal{C}$ it is enough to consider the boundary $\partial\mathcal{C}$. But actually, examining the restriction of g to the six faces of $\partial\mathcal{C}$, we see that it suffices to consider the vertices of $\partial\mathcal{C}$. Moreover, as $g(\alpha,\beta,\gamma)$ is obviously symmetric in α and β, and

$$g(\alpha,\beta,1+\epsilon) \geq g(\alpha,\beta,1-\epsilon),$$

we only have three vertices to consider.

1. $\alpha = 1-\epsilon, \beta = 1-\epsilon, \gamma = 1-\epsilon$.

$$g(\alpha,\beta,\gamma) = 2 > 2\cos\frac{\theta}{2} \text{ for any } 0 < \theta < 2\pi$$

2. $\alpha = 1+\epsilon, , \beta = 1-\epsilon, , \gamma = 1-\epsilon.$

$$g(\alpha,\beta,\gamma) = \frac{2}{1+\epsilon}$$

Thus we get, provided that $\cos\frac{\theta}{2} \neq 0$,

$$\epsilon < sec\,\frac{\theta}{2} - 1.$$

3. $\alpha = 1+\epsilon\,, \beta = 1+\epsilon\,, \gamma = 1-\epsilon.$

$$g(\alpha,\beta,\gamma) = 2\frac{1-\epsilon}{1+\epsilon}$$

Thus we get

$$\epsilon < \frac{1-\cos\frac{\theta}{2}}{1+\cos\frac{\theta}{2}}.$$

Observe that

$$\frac{1-\cos\frac{\theta}{2}}{1+\cos\frac{\theta}{2}} \leq \frac{1-\cos\frac{\theta}{2}}{\cos\frac{\theta}{2}} = sec\,\frac{\theta}{2} - 1 \text{ if } \cos\frac{\theta}{2} \neq 0,$$

and thus a sufficient condition for the coefficient of r^2 in $m(r)$ to be positive is

$$\mid \alpha - 1 \mid, \mid \beta - 1 \mid, \mid \gamma - 1 \mid \leq \epsilon, \text{ where } \epsilon < \frac{1-\cos\frac{\theta}{2}}{1+\cos\frac{\theta}{2}}. \qquad (1.2)$$

Remark 1.3 We want $\epsilon < 1$! This will follow from (1.2) if we restrict θ to lie in $(0,\pi)$.

Now that the coefficient of r^2 is positive, the critical point r_0 of $m(r)$ yields a global minimum of $m(r)$ on $\Re$. We have

$$\begin{aligned} r_0 &= \frac{\gamma(\alpha-\beta)}{\alpha\gamma+\beta\gamma-2\alpha\beta\cos\frac{\theta}{2}} \\ m(r_0) &= \frac{R_1}{n}\left[\gamma(\alpha+\beta) - \frac{\gamma^2(\alpha-\beta)^2}{\alpha\gamma+\beta\gamma-2\alpha\beta\cos\frac{\theta}{2}}\right] \end{aligned}$$

Thus

$$\begin{aligned} m(r_0) > 0 &\iff \gamma(\alpha+\beta) - \frac{\gamma^2(\alpha-\beta)^2}{\alpha\gamma+\beta\gamma-2\alpha\beta\cos\frac{\theta}{2}} > 0 \\ &\iff \frac{2\gamma}{\alpha+\beta} > \cos\frac{\theta}{2} \end{aligned}$$

Clearly for $(\alpha, \beta, \gamma) \in \mathcal{C}$, we have

$$\frac{2\gamma}{\alpha+\beta} \geq \frac{1-\epsilon}{1+\epsilon}$$

and we know that

$$\frac{1-\epsilon}{1+\epsilon} > \cos\frac{\theta}{2},$$

so that we may conclude that $m(r) > 0$ for all $r \in \Re$, when $(\alpha, \beta, \gamma) \in \mathcal{C}$ with ϵ satisfying equation (1.2).

(2) Let us write $x = \cos\frac{\theta}{2}$. We now choose

$$\epsilon = \frac{1}{2}\frac{1-x}{1+x}\,. \tag{1.4}$$

We want to find a condition on x ensuring that

$$-\frac{b(r)}{m(r)} \notin [-1,1], \text{ for all } r \in [-1,1].$$

Recall that $b(r) = e(r) + km(r)$, so that

$$-\frac{b(r)}{m(r)} = -k - \frac{e(r)}{m(r)}$$

and $-k \leq -1$ since k is an odd integer with $0 < k < n$. Thus it suffices to show that

$$e(r) > 0, \text{ for all } r \in [-1,1].$$

Now $e(r) = R_0 p(r)$ where

$$p(r) = 2(\alpha-\beta)(1-x)r^3 + [2(\alpha+\beta-\gamma) - (\alpha+\beta)x]r^2 + (\alpha-\beta)xr + 2\gamma.$$

Let

$$\begin{aligned} \rho &= 2(\alpha+\beta-\gamma) - (\alpha+\beta)x \\ \mu &= 2(\alpha-\beta)(1-x) \\ \delta &= (\alpha-\beta)x \end{aligned}$$

and observe that, writing $p(r) = p(r, \alpha, \beta, \gamma)$, we have

$$p(r, \beta, \alpha, \gamma) = p(-r, \alpha, \beta, \gamma),$$

so that, to show $p(r) > 0$ for all $r \in [-1,1]$ and for all $(\alpha,\beta,\gamma) \in \mathcal{C}$, it is enough to consider $r \in [-1,1]$ and

$$(\alpha,\beta,\gamma) \in \hat{\mathcal{C}} \equiv \{(\alpha,\beta,\gamma) \,|\, \alpha \geq \beta\} \cap \mathcal{C}.$$

When $\alpha = \beta$ we have

$$p(r) = 2[\alpha(2-x) - \gamma]r^2 + 2\gamma \geq 2\gamma > 0$$

for all $r \in \Re$, provided $\alpha(2-x) - \gamma \geq 0$. But, $\gamma \leq 1 + \epsilon$, $\alpha \geq 1 - \epsilon$ and $2 - x > 0$, so it suffices to show that

$$1 + \epsilon \leq (1-\epsilon)(2-x),$$

i.e.

$$\epsilon \leq \frac{1-x}{3-x}\,.$$

By our choice of ϵ, (1.4), we have

$$\frac{1}{2}\frac{1-x}{1+x} \leq \frac{1-x}{3-x} \iff x \geq 1/3$$

and we get

$$\cos\frac{\theta}{2} \geq 1/3. \tag{1.5}$$

We have shown that, when $\alpha = \beta$,

$$p(r) \geq 2\gamma \geq 2(1-\epsilon), \text{ for all } r \in \Re.$$

We now show that

$$\rho \geq \frac{-3x^2 + 4x - 1}{1+x}, \text{ for}(\alpha,\beta,\gamma) \in \hat{\mathcal{C}}. \tag{1.6}$$

We let $g(\alpha,\beta,\gamma) = 2(\alpha+\beta-\gamma) - (\alpha+\beta)x$, and we have

$$\left.\begin{array}{rcrcl} g_\alpha & = & 2-x & > & 0 \\ g_\beta & = & 2-x & > & 0 \\ g_\gamma & = & -2 & < & 0 \end{array}\right\} \text{ for values of } \alpha,\beta,\gamma \in \Re.$$

It follows that the minimum of g on $\mathcal{C}$ occurs at the vertices of $\mathcal{C}$. However g is symmetric in α,β and

$$g(\alpha,\beta,\gamma) \geq g(\alpha,\beta,1+\epsilon),$$

so there are only three vertices to consider.

1. $\alpha = 1 + \epsilon \, , \beta = 1 - \epsilon \, , \gamma = 1 + \epsilon$.

$$\rho = 2(1-x) - 2\epsilon \tag{1.7}$$

2. $\alpha = 1 + \epsilon \, , \beta = 1 + \epsilon \, , \gamma = 1 + \epsilon$.

$$\rho = 2(1-x) - 2\epsilon(1-x) \tag{1.8}$$

3. $\alpha = 1 - \epsilon \, , \beta = 1 - \epsilon \, , \gamma = 1 + \epsilon$.

$$\rho = 2(1-x) - 2\epsilon(3-x). \tag{1.9}$$

Now for $0 < \theta < \pi$, we have

$$1 - x < 1 < 3 - x$$

so that

$$(1.9) < (1.7) < (1.8).$$

By (1.4) again, we have

$$(1.9) = \frac{-3x^2 + 4x - 1}{1 + x},$$

and inequality (1.6) follows.
The roots of $-3x^2 + 4x - 1$ are 1 and 1/3, so that

$$\frac{-3x^2 + 4x - 1}{1 + x} > 0, \text{ for all } 1/3 < x < 1.$$

Thus we have shown that for $(\alpha, \beta, \gamma) \in \hat{\mathcal{C}}$,

$$\rho \geq \frac{-3x^2 + 4x - 1}{1 + x} > 0, \text{ for } 1/3 < \cos\frac{\theta}{2} < 1.$$

Moreover, if $\alpha > \beta$, then

$$0 < \mu \leq 4\epsilon(1-x) \text{ and } 0 < \delta \leq 2\epsilon x.$$

But

$$4\epsilon(1-x) = 2\frac{(1-x)^2}{1+x} \text{ and } 2\epsilon x = \frac{x(1-x)}{1+x}.$$

Thus, for $\alpha > \beta$, we have

$$p(r) = \mu r^3 + \rho r^2 + \delta r + 2\gamma > 0, \text{ for all } r \in [0,1],$$

so it suffices to consider $r \in [-1,0]$. Then $r^3 \leq 0$, $r \leq 0$, so that

$$p(r) \geq \frac{2(1-x)^2}{1+x} r^3 + \frac{-3x^2+4x-1}{1+x} r^2 + \frac{x(1-x)}{1+x} r + 2(1-\epsilon),$$

and

$$\begin{aligned} p(r) > 0 \iff & \ 2(1-x)^2 r^3 + (-3x^2+4x-1)r^2 \\ & \ x(1-x)r + 2(1-\epsilon)(1+x) > 0. \end{aligned}$$

Now

$$1 - \epsilon = 1 - \frac{1}{2}\frac{1-x}{1+x} = \frac{1+3x}{2(1+x)},$$

so that

$$2(1+x)(1-\epsilon) = 1 + 3x\,,$$

and we must show that

$$2(1-x)^2 r^3 + (-3x^2+4x-1)r^2 + x(1-x)r + 1 + 3x > 0, \ \forall\, r \in [-1,0]$$

i.e.

$$(-3x^2+4x-1)r^2 + 1 + 3x > -r[2(1-x)^2 r^2 + x(1-x)].$$

Since $-r \leq 1$ on $[-1,0]$, it is enough to show that

$$2(1-x)^2 r^2 + x(1-x) < (-3x^2+4x-1)r^2 + 1 + 3x, \ \forall\, r \in [-1,0].$$

This will certainly be the case if

(a) $x(1-x) < 1 + 3x$

(b) $2(1-x)^2 < -3x^2 + 4x - 1$

But (a) holds if and only if $(1+x)^2 > 0$, which holds for all $x \neq -1$.
On the other hand, (b) holds if and only if $-5x^2 + 8x - 3 > 0$. The roots of this polynomial are 1 and 3/5. Thus the inequality (b) holds for all

$$3/5 < x < 1.$$

To summarize, we have proved theore 10.1, namely that the jacobian $det\,JF$ does not vanish for all α, β, γ and θ such that

$$\mid \alpha - 1 \mid, \mid \beta - 1 \mid, \mid \gamma - 1 \mid \leq \frac{1}{2} \frac{1 - \cos\frac{\theta}{2}}{1 + \cos\frac{\theta}{2}}$$

and

$$3/5 < \cos\frac{\theta}{2} < 1.$$

References

[1] M.S. Ashbaugh and R.D. Benguria, *Proof of the Payne-Polya-Weinberger Conjecture*, Bull. Amer. Math. Soc. 25 (1991), 19–29.

[2] M.S. Ashbaugh and R.D. Benguria, *Sharp Bound for the Ratio of the First Two Eigenvalues of Dirichlet Laplacians and Extensions*, to appear in Annals of Math.

[3] C. Bandle, *Isoperimetric Inequalities and Applications*, Monographs and Studies in Mathematics 7, Pitman, London, 1980.

[4] K. Bathe and E. Wilson, *Numerical Methods in Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[5] P. Ciarlet, *Numerical Analysis of the Finite Element Method*, Les Presses de l'Université de Montréal, Montréal, 1976.

[6] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Hooland, New York, 1978.

[7] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Les publications **CRM**, Université de Montréal, Montréal, 1989.

[8] S. Cox and M. Overton, *On the optimal design of columns against buckling*, to appear in SIAM J. Math. Analysis.

[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Wiley, New York, 1989.

[10] J.J. Dongarra , et al., *LINPACK Users Guide*, SIAM, Philadelphia, 1978.

[11] R. Fletcher, *Practical Methods of Optimization*, Wiley, New York, 1987.

[12] B.S. Garbow , et al., *Matrix Eigensystem Routines: EISPACK Guide Extension*, Lecture Notes in Computer Science 51, Springer-Verlag, New York, 1977.

[13] P. Gill, W. Murray and M. Wright, *Practical Optimization*, Academic Press, London, 1981.

[14] J. Kuttler and V. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev. 26 (1984), 163–193.

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

[16] M. Overton, *Large-scale Optimization of Eigenvalues*, To appear in SIAM Journal on Optimization.

[17] M. Overton and R. Womersley, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, submitted to Math. Programming.

[18] L.E. Payne, G. Polya, and H.F. Weinberger, *On the ratio of consecutive eigenvalues*, J. Math. and Phys. 35 (1956), 289-298.

[19] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, 1973.